

Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Michał Borkowski

Nr albumu: 370727
Marian Dziubiak

Nr albumu: 370784

Jakub Bujak

Nr albumu: 370737
Marek Puzyna

Nr albumu: 371359

Kompilacja NianioLanga do efektywnych konstrukcji języka C

Praca licencjacka
na kierunku INFORMATYKA

Praca wykonana pod kierunkiem
mgr. Radosława Bartosiaka
Instytut Informatyki

Maj 2018

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data

Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data

Podpisy autorów pracy

Streszczenie

W pracy opisujemy...

Słowa kluczowe

kompilacja, języki programowania, analiza semantyczna, NianioLang, system typów

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.3 Informatyka

Klasyfikacja tematyczna

D. Software

D.3. Programming languages

D.3.3. Language constructs and features

Spis treści

Wprowadzenie	5
1. Wstęp	7
1.1. Wprowadzenie do NianioLanga	7
1.2. Typy w NianioLangu	7
1.3. Cele projektu	8
1.4. Podstawowe pojęcia	8
2. Metodyka pracy	9
2.1. Korzystanie z systemu kontroli wersji	9
2.2. Zgłaszanie zmian i code review	10
2.3. Techniki komunikacji w zespole	10
3. Kompilator NianioLanga	11
3.1. Budowanie drzewa AST	11
3.2. Analiza semantyczna	12
3.3. Architektura nlasma	13
3.4. Translacja drzewa AST do nlasma	14
3.5. Generowanie kodu C na podstawie nlasma	14
3.6. Implementacja typów NianioLanga w C	14
4. Zmiana systemu typów	17
4.1. Rozdzielenie typu <code>ptd::sim</code>	17
4.2. Typy <code>own</code>	17
5. Rozszerzenie nlasma	19
5.1. Przekazywanie informacji o typach z drzewa AST	19
5.2. Statyczne sprawdzanie poprawności	19
6. Nowe implementacje typów	21
6.1. Typy proste	21
6.1.1. Liczby całkowite	21
6.1.2. Łańcuchów znaków	21
6.1.3. Wartości logiczne	21
6.1.4. Ograniczenia kompilacji typów prostych	22
6.2. Typy złożone	22
6.2.1. Tablice	22
6.2.2. Rekordy	22
6.2.3. Typy wariantowe	23

6.2.4. Tablice asocjacyjne	23
6.3. Konwersja $own \rightarrow im$	23
7. Efekty optymalizacji i wnioski	27
7.1. Porównanie czasu wykonania programów	27
8. Wkład poszczególnych członków zespołu	29
Bibliografia	31

Wprowadzenie

NianioLang jest językiem programowania ogólnego przeznaczenia. Jego twórcą jest założyciel firmy Atinea, na zlecenie której realizowana jest poniższa praca – Andrzej Gąsienica-Samek. Istniejący kompilator umożliwia translację NianioLanga do kilku języków, między innymi do Javy, JavaScriptu i C. Ze względu na chęć uproszczenia kompilacji NianioLanga utworzono środowisko uruchomieniowe dostarczające odpowiednie abstrakcje i pozwalające na korzystanie w C z dynamicznych struktur odpowiadających typom, jakie są dostępne do użycia w NianioLangu. Takie rozwiązanie nie jest niestety optymalne, szczególnie w przypadku niskopoziomowego języka jakim jest C. W tej pracy opisujemy wprowadzenie nowych typów danych i ich wsparcia w kompilatorze, co umożliwi generowanie natywnego kodu w C i znacznie zwiększy wydajność kompilowanych aplikacji przy zachowaniu podstawowych założeń języka.

Rozdział 1

Wstęp

1.1. Wprowadzenie do NianioLanga

NianioLang jest proceduralnym, imperatywnym językiem, którego celem jest uproszczenie pisania rozproszonych aplikacji stosując wzorzec projektowy Nianio[1]. Celem twórców języka jest dostarczenie narzędzia umożliwiającego operowanie na niemutowalnych strukturach o semantyce zbliżonej do języków funkcyjnych, ale prostszego w użyciu. Konstrukcje takie jak wskaźniki zostały usunięte, a w ich miejsce, w przypadku kompilacji do C, zaimplementowano system zarządzania obiektami w pamięci przez zliczanie referencji, aby pozbyć się jednego z najczęstszych źródeł błędów, występujących przy programowaniu niskopoziomym. Takie zabiegi zmniejszają nieco wydajność języka, ale jego twórcy doszli do wniosku, że zysk z wysokopoziomowego podejścia do pisania aplikacji jest wystarczająco duży, aby tworzenie aplikacji w NianioLangu było opłacalne.

Kompilator NianioLanga jest rozwijany przez firmę Atinea, która używa NianioLanga w swoich projektach (m.in. InstaDB.com). Kod źródłowy kompilatora jest dostępny na platformie GitHub[2] na licencji MIT.

1.2. Typy w NianioLangu

W NianioLangu mamy doczynienia z kilkoma wbudowanymi typami. Wszelkie typy tworzone przez użytkownika, to właściwie aliasy na typy wbudowane, co pomaga w zarządzaniu abstrakcją w programie. Dostępnych jest pięć wbudowanych typów:

- `ptd::sim` – liczby całkowite, zmiennoprzecinkowe oraz ciągi znaków
- `ptd::rec` – rekordy, czyli odpowiedniki struktur znanych np. z C
- `ptd::hash` – słowniki o kluczach będących ciągami znaków i wartościach danego typu
- `ptd::var` – typ wariantowy, który reprezentuje obiekty, mogące być w dokładnie jednym z określonego zbioru stanów i dodatkowo zawierać dane o typie właściwym dla danego stanu
- `ptd::arr` – tablice danych tego samego typu

Typy te mogą być ze sobą łączone w bardziej skomplikowane konstrukcje, na przykład `ptd::arr(ptd::sim())` definiuje typ tablicy wartości prostych, a `ptd::rec({a => ptd::sim(), b => ptd::sim()})` definiuje typ rekordu o dwóch polach będących wartościami prostymi.

Podstawowymi założeniami systemu typów w NianioLangu są: niemutowalność struktur i opcjonalne typowanie.

- Niemutowalność struktur w NianioLangu jest pojęciem słabszym, niż w językach funkcyjnych. Oznacza ona, że język daje gwarancję, iż między dwoma kolejnymi dostęпами do zmiennej jej wartość nie ulegnie zmianie. W klasycznych językach imperatywnych, posiadających wskaźniki lub referencje nie jest to prawdą – jeśli istnieją dwa wskaźniki do jednej zmiennej, wartość odczytana z jednego z nich może się zmieniać nawet bez jego jawnej modyfikacji.
- Opcjonalne typowanie oznacza gwarancję, że dodanie lub usunięcie typów z programu nie zmieni jego semantyki. W skrajnym przypadku można usunąć z programu całą informację o typach i będzie on działał bez zmian (oczywiście może to być ze szkodą dla łatwości utrzymania lub wydajności). Jest to podejście przeciwne do stosowanego w takich językach jak C++, czy Haskell, których złożone systemy typów mają istotny wpływ na semantykę programu.

W wielu językach istnieje znacznie więcej wbudowanych typów, jednak powyższe są wystarczające, aby zbudować skomplikowane aplikacje, a jednocześnie dość proste, żeby rozpoczęcie programowania w NianioLangu nie było dla programisty wyzwaniem. W rozdziale Kompilator NianioLanga opisana została implementacja powyższych typów w języku C.

1.3. Cele projektu

Celem projektu była modyfikacja kompilatora NianioLanga, w taki sposób, aby kod wynikowy w C zawierał mniejszą liczbę wywołań funkcji oraz skomplikowanych struktur dla prostych typów istniejących już w języku C. Dzięki temu zmniejszony został czas wykonania aplikacji, pozwalając kompilatorowi GCC na zastosowanie dodatkowych optymalizacji. W tym celu wprowadzone zostały nowe typy z przestrzeni nazw `own`, które będą bardziej niskopoziomowymi odpowiednikami typów z przestrzeni nazw `ptd` (rekordy, tablice i warianty). Dzięki nałożonym na nie ograniczeniom możliwe jest znaczne zwiększenie wydajności programów pisanych w NianioLangu przy zachowaniu niezmienionej semantyki języka. Jednocześnie zmienione zostały pewne podstawowe typy, mianowicie `ptd::sim` został rozbity na `ptd::int` oraz `ptd::string`, oraz wprowadzony został typ `ptd::bool`. Liczby całkowite i wartości boole'owskie mają w języku C natywną reprezentację i używając ich bezpośrednio uzyskujemy prostszy kod wynikowy w C, który jest łatwiejszy do optymalizacji przez kompilator GCC.

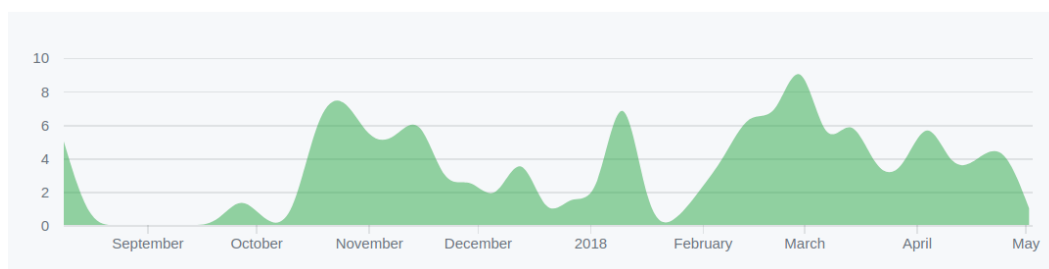
Skutkiem zmian jakie musiały zostać wprowadzone by osiągnąć cel projektu było utracenie możliwości kompilacji NianioLanga do języków innych niż C do czasu implementacji nowych typów w pozostałych językach. To zadanie pozostaje jednak poza zakresem niniejszej pracy.

1.4. Podstawowe pojęcia

Rozdział 2

Metodyka pracy

Prace nad kodem prowadzone były regularnie od połowy października do maja, z wyłączeniem sesji na przełomie lutego i stycznia. Na rysunku 2.1 przedstawiony został wykres zmian, które wpływały na główną gałąź repozytorium kodu. Co tydzień zespół prezentował swoje poczynania opiekunowi, co dostarczało stałej motywacji do implementacji kolejnych elementów projektu.



Rysunek 2.1: Liczba commitów na głównej gałęzi repozytorium w czasie trwania projektu.

2.1. Korzystanie z systemu kontroli wersji

Podczas pracy zespołowej niezmiennie istotna jest wymiana i zarządzanie fragmentami kodu, jakie piszą poszczególni członkowie zespołu. Ponieważ każdy w zespole pracował już wcześniej z systemem kontroli wersji Git, to został on użyty w tym projekcie. Dodatkowym atutem tego systemu kontroli wersji jest szeroka dostępność darmowych publicznych repozytoriów kodu online. Kod projektu umieszczony został na portalu GitHub, który jest największym serwisem pozwalającym na hostowanie kodu online.

Początkowo każdy w zespole miał swoją gałąź w repozytorium, oznaczoną jego imieniem, na której wprowadzał swoje zmiany. Jednak w trakcie pracy za praktyczniejsze podejście zostało uznane tworzenie osobnych gałęzi kodu dla każdego zadania. Umożliwiło to pracę nad kilkoma niezależnymi zadaniami jednocześnie i ułatwiło przeglądanie kodu poprzez tematyczny podział zmian.

GitHub oprócz repozytorium kodu, udostępnia również narzędzia do zarządzania zadaniami w projekcie, co zostało wykorzystane. Każde duże i średnie zadanie było zapisywane w systemie zgłoszeń. Po ustaleniu osoby wykonującej zadanie było ono przypisywane na tę osobę, a po wgraniu zmian na główną gałąź zadanie było zamykane. Dzięki temu w dowolnej chwili możliwe było określenie nad czym pracuje dany członek zespołu i jakie otwarte zadania

pozostały do przydzielenia.

2.2. Zgłaszanie zmian i code review

Po napisaniu kodu rozwiązującego dane zadanie, zgłaszany był tzw. Pull Request, czyli żądanie zatwierdzenia zmian. Aby zmiany zostały zatwierdzone i umieszczone w głównej gałęzi projektu, musiały być zatwierdzone przez conajmniej jednego członka zespołu, innego niż ten który zgłosił PR. GitHub udostępnia widok naniesionych zmian wraz z możliwością dodawania komentarzy do konkretnych fragmentów kodu. Możliwy więc był dialog między recenzentem kodu, a jego twórcą, tak aby ostatecznie zatwierdzone zmiany spełniały oczekiwania obu stron.

2.3. Techniki komunikacji w zespole

Większość komunikacji w zespole odbywała się online przez portal społecznościowy Facebook. W początkowej fazie projektu, głównym tematem dyskusji była specyfika języka NianioLang, z którym większość członków zespołu miała po raz pierwszy styczność. Na dalszym etapie projektu, na konwersacji grupowej na tym portalu, zespół ustalał większość detali implementacyjnych przed rozpoczęciem ich wdrażania. Ten środek komunikacji był najchętniej wykorzystywany ze względu na najkrótszy czas reakcji członków zespołu.

Drugim kanałem komunikacji były zadania i PRy na GitHubie. Zadania w prosty sposób opisywały co i kto ma zrobić, a opisy PRów określały co zostało przez daną osobę zrobione. Ta forma komunikacji w projekcie była najmniej używana.

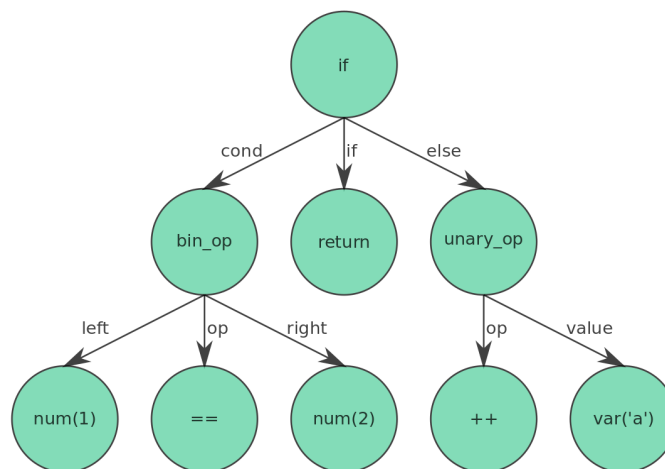
Ostatnim sposobem komunikacji były cotygodniowe spotkania podczas zajęć z Zespołowego Projektu Programistycznego, po których zespół spędzał 10-20 minut na dogadywaniu się, kto wybiera które zadanie, a w przypadku pytań lub problemów, na ich rozwiązaniu.

Rozdział 3

Kompilator NianioLanga

3.1. Budowanie drzewa AST

Pierwszym etapem kompilacji jest parsowanie programu wejściowego do drzewa AST. Takie drzewo jest strukturą przechowującą informacje o składniowej roli wyrażeń, które przetwarzać można na dalszych etapach kompilacji. Celem jest przetworzenie tekstu programu do postaci przypominającej derywację gramatyki języka. Chociaż książkowa definicja drzewa AST mówi wyłącznie o składni, kompilator NianioLanga już na tym etapie dokonuje pewnej podstawowej analizy semantycznej, gdyż wydzielenie tych czynności spowodowałoby dużą duplikację kodu związanego z czynnością obchodzenia drzewa.



Rysunek 3.1: Drzewo AST dla programu `if (1 == 2) {return;} else {a++;}`

Najpierw parser próbuje przeczytać listę importowanych modułów, gdyż zgodnie ze składnią języka musi ona zostać umieszczona na początku programu. Po tym etapie następuje parsowanie listy funkcji. Każdy etap parsowania odbywa się poprzez analizę możliwie najdłuższej ilości tekstu, dopisanie błędów, ustawienie wskaźnika następnej pozycji oraz innych pomocniczych wartości, które są przechowywane w zmiennej reprezentującej stan parsera, a na końcu zwrócenie sparsowanej wartości. Nie ma znaczenia, co zwróci dana funkcja w razie błędu, gdyż wtedy kompilacja zostanie zatrzymana, a użytkownik poinformowany o błędzie. W przypadku, gdy nie jest jasne, co należy sparsować (na przykład nie wiadomo, czy oczekujemy kolejnej deklaracji importu modułu, czy definicji funkcji), próbuje się przeczytać różne

wartości, aż któraś zostanie poprawnie sparsowana. Ważne jest, by różne możliwości przetwarzania od największej do najmniejszej, jeżeli reprezentacja jedna może być początkiem drugiej, np. najpierw parsować należy operatory, a dopiero potem wyrażenia, chyba że kolejność nie ma znaczenia. Przykładowo w wyrażeniu $1 + 2 - 3$ nie ma znaczenia, czy spróbujemy sparsować najpierw operator dodawania, czy odejmowania, ale operator musi być sparsowany zanim sparsowana zostanie liczba. Innym przykładem jest `if`, po którym może nastąpić `else` i najpierw należy spróbować sparsować blok `else` zanim przejdzie się do parsowania następnych komend.

Parsowanie funkcji odbywa się poprzez sparsowanie nagłówka, a następnie sparsowanie jej wnętrza jako jednej komendy. Przyjęte jest, że komenda może być także listą komend, co parsuje się w sposób rekurencyjny. Ułatwia to parsowanie takich komend jak `for` czy `if`. Podejście to umożliwia także tworzenie rekurencyjnych funkcji, z których każda odpowiedzialna jest za jeden element gramatyki języka. Funkcje te sposób działania opierają na opisanym w poprzednim paragrafie schemacie.

Projekt nie wymagał znaczących zmian na etapie parsowania, dlatego nie zostaną one tutaj omówione. Jedynym wyjątkiem jest dodanie sprawdzania po sparsowaniu funkcji, czy nie definiuje ona typu, a następnie wpisania tego typu do drzewa. Zostało to jednak wykonane przy użyciu dostępnych już funkcji. Jest to wykonywane, by na dalszych etapach można było zapytać o dowolny definiowany typ, co ułatwia proces kompilacji. Sytuacja błędna nie jest jednak obsługiwana, stanie się to dopiero podczas sprawdzania poprawności typów.

3.2. Analiza semantyczna

Drugim etapem kompilacji jest sprawdzenie poprawności typów. Wbrew nazwie etap ten nie ma na celu wyłącznie sprawdzenia, czy program jest poprawny typowo – do drzewa programu wpisywane są informacje o uzyskanych typach. Jest to konieczne, by móc dopasować typy `NiaoLanga` do typów języka wynikowego. Ponadto zachodzi tutaj analiza opisana w rozdziale `Konwersja own → im`.

Na tym etapie sprawdza się, czy zaimportowane moduły faktycznie istnieją, czy definicje funkcji nie dublują się, oraz w końcu czy używane typy zgodne są z oczekiwanymi. Trzeci element rozumiany jest dwojako. Po pierwsze, jeżeli dana zmienna ma zadeklarowany przez użytkownika typ, sprawdza się, czy używana jest ona wyłącznie przez funkcje, które takiego typu oczekują. Ze względu na nietrywialny system typów wydzielone są specjalne funkcje sprawdzające czy i w jakim zakresie dwa typy są zgodne. Przykładowo jeżeli `im` używany jest w kontekście `hash`, to zgodnym typem jest `hash`. Jeżeli zaś funkcja oczekuje `int` a podawany jest `array`, to zgłaszany jest błąd. W tym kontekście o typach myśleć można jako o zbiorach możliwych wartości, na których wykonywane są operacje, przede wszystkim przecięcia.

Drugie rozumienie, które stosuje type checker, jest podobne do pierwszego, lecz w kontekście wnioskowania typów na podstawie wykonywanych na nich operacji. Jeżeli dana zmienna inicjalizowana jest jako `int`, a następnie używana jako `hash`, to zgłaszany jest błąd. Warto zaznaczyć, że operacje na typach wywnioskowanych są subtelniejsze, gdyż wiele typów jest konwertowalnych na `im` oraz w drugą stronę, przez co nie jest możliwa całkowita, ścisła kontrola typów, ta ma miejsce dopiero w runtime. Ponownie, wydzielone są funkcje operujące na typach w analogii do zbiorów.

Samą budowę type checker ma podobną w pewnym sensie do parsera. W rekurencyjny sposób sprawdzamy, jakie typy mają poszczególne elementy drzewa oraz czy są one poprawne. Tak jak w parserze zapisane jest, że operator `+` zawiera słowo kluczowe `+`, tak odpowiednia funkcja w type checkerze wie, że argumenty tego operatora muszą mieć typ zwracany `int`.

Ponownie pozwala to na podział funkcji w taki sposób, by każda dotyczyła jednego, jasno wyszczególnionego elementu.

3.3. Architektura nlasma

nlasma to język stosowany w kompilatorze jako pomost pomiędzy NianioLangiem a językiem wyjściowym, gdyż bezpośrednia kompilacja NianioLanga do C, lub też innego języka, może okazać się trudna, ponieważ używane struktury językowe znacząco od siebie odbiegają. Umożliwia to w łatwy sposób tworzenie modułów kompilujących NianioLanga do różnych języków. Taka technika jest szeroko stosowana w świecie kompilatorów, jednym z najszerzej znanych przykładów jest LLVM. Obecność autorskiego nlasma podyktowana jest przede wszystkim potrzebą języka pośredniego, który z jednej strony będzie możliwie najprostszy, z drugiej w pewnym sensie zachowa strukturę programu wyjściowego.

Funkcje w nlasmie nie operują na zmiennych, lecz na rejestrach. Każdy rejestr ma swój unikalny w obrębie funkcji numer oraz typ. Także argumenty funkcji są rejestrami. Wprowadzenie typów do nlasma było jednym z elementów projektu, gdyż o ile poprzednio wszystkie zmienne konwertowane były do typu `im`, o tyle teraz przechować należy strukturę typów `own` czy `int`. Typy w nlasmie mają charakter wyłącznie informacyjny, to znaczy nie mają żadnego znaczenia semantycznego.

Najważniejszą cechą nlasma jest nie drzewiasta, lecz liniowa struktura programu. Oznacza to, że argumentami wywołań mogą być wyłącznie rejestry, a nie zmienne czy inne wywołania. W znacznym stopniu ułatwia to dalszą pracę kompilatora, przede wszystkim z powodu mniejszej liczby niuansów dotyczących na przykład wartości zwracanych, przez co kod generatora języka docelowego może być znacznie prostszy. Ponadto generowanie kodu może odbywać się w sposób liniowy, to znaczy każda komenda nlasma może mieć jednoznaczne, proste odzwierciedlenie w języku docelowym. Dzięki temu generator może dokonywać analizy wyłącznie w zakresie charakterystycznym dla danego języka, na przykład do generatora C zostało wprowadzone sortowanie topologiczne deklaracji typów.

Inną ważną cechą nlasma, wynikającą częściowo z poprzedniej, jest brak pętli. Translator NianioLanga do nlasma musi rozbić każdą pętlę na etykietę oraz komendę `goto`. Takie podejście ułatwia tłumaczenie kodu na język docelowy, gdyż w różnych językach pętle działają w odrobinę inny sposób, zaś `goto` tworzy nieskomplikowany, wspólny interfejs.

Ponadto nlasma zawiera polecenia służące do zarządzania pamięcią. Zakłada się, że język docelowy będzie korzystał ze zliczania referencji w celu zwalniania pamięci. Aby nie pisać tej samej funkcjonalności oddzielnie dla wielu generatorów, nlasma zawiera polecenia mówiące o tym, że licznik referencji danego rejestru (o ile jest on wskaźnikiem) należy zmniejszyć lub zwiększyć, co można w prosty sposób przełożyć na wywołania innych języków. Oczywiście, jeżeli generator wybierze inną metodę zarządzania pamięcią, na przykład generując kod Javy, może te polecenia po prostu zignorować. W przypadku generowania języka z manualnym zarządzaniem pamięcią, na przykład C, polecenia te są bardzo pomocne.

Przykładowe komendy nlasma:

- `if_goto` – sprawdza wartość w rejestrze. Jeżeli prawda, wykonuje skok pod etykietę
- `clear` – sygnalizuje, że wartość licznika referencji powinna spaść
- `call` – wywołuje funkcję o danej nazwie z danymi rejestrami jako argumentami, zapisuje wynik w danym rejestrze
- `move` – kopiuje wartość między dwoma rejestrami

3.4. Translacja drzewa AST do nlasma

3.5. Generowanie kodu C na podstawie nlasma

Jak zostało to omówione w poprzednim rozdziale, `nlas` został zaprojektowany z myślą o ułatwieniu dalszej kompilacji. Wobec tego generowanie kodu C jest bardzo proste.

Pierwszym etapem jest wygenerowanie kodu dla pliku nagłówkowego modułu. Zawiera on między innymi komentarz z informacją o tym, że dany plik został wygenerowany automatycznie, instrukcje importujące nagłówki C biblioteki standardowej NianioLanga, czy też pewne instrukcje inicjujące, na przykład ustawiające wartości stałych. Warto wspomnieć, że pliki `.c` i `.h` generowane są równolegle, gdyż pozwala to na przetwarzanie zgodnie z kolejnością semantyczną, to znaczy na przykład generowanie wszelkich informacji o funkcji można oddelegować do raz wywoływanej procedury.

Drugim etapem jest wygenerowanie kodu odpowiedzialnego za obsługę typów, czyli ich deklaracje, definicje oraz potrzebne do obsługi funkcje. Przed zmianami wprowadzonymi w ramach tego projektu etap ten nie występował, gdyż kod C używał ogólnego typu `ImmT`, który obsługiwać można biblioteką standardową NianioLanga. Obecnie, ponieważ dla każdego typu `own` tworzony jest oddzielny typ na poziomie C, a funkcje go obsługujące generowane są na poziomie NianioLanga, to podejście zostało zmienione. Ze względu na fakt, że definicje w NianioLangu nie muszą występować w określonej kolejności, lecz w C już tak, konieczne jest sortowanie topologiczne, z uwzględnieniem możliwości wyodrębnionej deklaracji typu. Jest to potrzebne do obsługi sytuacji, w której typ A potrzebuje w swojej definicji posługiwać się w pełni typem B, który z kolei posługuje się wskaźnikiem na typ A.

Trzecim i ostatnim etapem jest wygenerowanie kodu odpowiedzialnego za same funkcje NianioLanga. Komendy `nlas` w bardzo prosty sposób tłumaczą się na język C, wobec czego generator na tym etapie wykonuje niewiele logiki, lecz jedynie pracę związaną z właściwie słownikowym przełożeniem instrukcji mając jedynie na uwadze pewne szczegóły techniczne. Do tych szczegółów należy, pośród innych pomniejszych zadań, utworzenie wrapperów na funkcje tak, by można było je wywoływać także mając daną tablicę argumentów, obsługa wskaźników i zarządzanie pamięcią oraz tworzenie komentarzy będących informacją, której linii oryginalnego programu odpowiada dana seria instrukcji.

Po wykonaniu wszystkich trzech etapów pliki `.c` i `.h` danego modułu gotowe są do zapisania na dysku, a kompilator przechodzi do tłumaczenia kolejnego modułu.

3.6. Implementacja typów NianioLanga w C

Implementacja typów `ptd` opiera się na generycznym typie `ImmT`. Typ ten zdefiniowany jest jako `void*`, a w czasie wykonania programu wskazuje na konkretne struktury danego obiektu. Pierwsze bajty każdego obiektu zawierają informacje o tym, jaki jest jego typ, oraz jaką wartość ma licznik referencji. Jest to powszechny schemat symulowania programowania obiektowego w językach, które takowego nie wspierają, a takim językiem jest C.

Operacje na danych definiowane są przez szereg funkcji dostępnych w bibliotece standardowej NianioLanga, dokompilowywanej do każdego programu, to znaczy każdy moduł NianioLanga importuje jej nagłówki. Zawierają one podstawową funkcjonalność obsługi arytmetyki, struktur danych, operacji I/O.

Cała obsługa jest dość prosta i ma charakter możliwie najbardziej generyczny, wobec czego jest niezwykle niewydajna, w szczególności w kontekście operacji na liczbach lub wartościach logicznych. Celem projektu była zmiana tej sytuacji i umożliwienie kompilatorowi C

zastosowanie większej liczby optymalizacji, co opisane jest w kolejnych rozdziałach.

Rozdział 4

Zmiana systemu typów

4.1. Rozdzielenie typu `ptd::sim`

4.2. Typy `own`

Jaki jest cel tych typów, ich semantyka, jakie ograniczenia na ich użycie nakładamy (w stosunku do typów `ptd`).

Rozdział 5

Rozszerzenie nlasma

5.1. Przekazywanie informacji o typach z drzewa AST

5.2. Statyczne sprawdzanie poprawności

Rozdział 6

Nowe implementacje typów

6.1. Typy proste

Pierwszą zmianą, jaka została wprowadzona w generatorze kodu, było uproszczenie, tam gdzie to możliwe, kodu generowanego dla operacji na typach prostych. Rozdzielenie typu `ptd::sim` na osobne typy dla liczb naturalnych, łańcuchów znaków i wartości logicznych umożliwiło bardziej uważny dobór instrukcji języka C, co przełożyło się na wydajniejszy kod wynikowy.

6.1.1. Liczby całkowite

Wartości typu `ptd::int`, czyli liczby całkowite, w wyniku wprowadzonych zmian są kompilowane do zmiennych języka C o typie `int`. Zapewnia to dobrą wydajność operacji arytmetycznych, które na obecnych procesorach wykonują się w niewielkiej liczbie cykli. Użycie typu języka C dedykowanego dla liczb całkowitych umożliwia też kompilatorowi `gcc` lepszą optymalizację na dalszym etapie kompilacji (na przykład umożliwia ewaluację stałych w czasie kompilacji).

6.1.2. Łańcuchów znaków

Implementacja łańcuchów znaków nie została zmieniona w stosunku do istniejącej wcześniej. Decyzja ta była spowodowana efektywnością już istniejącej implementacji – łańcuchy znaków są kompilowane to tablice typu `char []`, opakowanych w struktury zawierające rozmiar i pojemność tablicy. Dostęp takich zmiennych odbywa się poprzez specjalne funkcje, zapewniające utrzymanie poprawnego stanu całej struktury. Jest to minimalny narzut potrzebny żeby zachować bezpieczeństwo języka i uchronić programistę przed błędami, takimi jak przepełnienie bufora. Z tego względu została podjęta decyzja o niepowielaniu istniejącej implementacji i wykorzystaniu tej już istniejącej.

6.1.3. Wartości logiczne

Jak zostało to opisane w rozdziale Implementacja typów NianioLanga w C, dotychczasowa implementacja wartości logicznych opierała się na dwuwartościowych języka NianioLang. Z uwagi na to, że warianty są dość wysokopoziomowym mechanizmem, bez bezpośredniego odzwierciedlenia w języku C, kod wynikowy generowany dla wartości logicznych nie był efektywny.

W wyniku przeprowadzonych zmian wartości logiczne są kompilowane do zmiennych typu `bool`, zdefiniowanego w pliku `stdbool.h` biblioteki standardowej języka C. Podobnie, jak w

przypadku liczb całkowitych zapewnia to dobrą wydajność operacji logicznych ze względu na bezpośrednie wsparcie procesora i kompilatora `gcc`.

6.1.4. Ograniczenia kompilacji typów prostych

Z uwagi na leżące u podstaw języka założenie o opcjonalności typów nie było możliwe zagwarantowanie efektywnej kompilacji typów prostych w każdych warunkach. Na przykład kod przedstawiony na listingu 6.1 zostanie skompilowany do ciągu operacji na mniej efektywnym typie `ImmT`. Spowodowane jest to zmianą typu zmiennej `x` w trakcie działania programu, przez co nie można jej przypisać żadnego z typów prostych.

Listing 6.1: Zmienna kompilowana do typu `ImmT`

```
var x;  
x = 'a';  
x = 1;  
x++;
```

6.2. Typy złożone

Dla każdego typu złożonego z przestrzeni `ptd` został utworzony odpowiadający mu typ w przestrzeni `own`. Dla każdego konkretnego typu, zdefiniowanego przez użytkownika i korzystającego z typów `own`, w kodzie wynikowym w C generowane są oddzielne definicje typu oraz funkcji na nim operujących. Jest to główna różnica w stosunku do typów z przestrzeni `ptd`, dla których zdefiniowany był jeden polimorficzny typ `ImmT`. Takie podejście umożliwia bardziej efektywną alokację pamięci przez kompilator (ponieważ rozmiar wielu struktur jest z góry znany) oraz zapewnia statyczną kontrolę typów także przez kompilator `gcc`, pozwalając wykrywać wiele błędów w kodzie generowanym przez kompilator NianioLanga już na etapie kompilacji.

6.2.1. Tablice

Definicje typów `own::arr` kompilowane są do struktury zawierającej rozmiar tablicy, jej maksymalną pojemność i wskaźnik na jej pierwszy element. Do programu wynikowego dodawane są również funkcje, pozwalające na dodawanie nowego elementu na koniec tablicy, pobieranie wskaźnika do elementu o danych indeksie oraz na pobieranie rozmiaru tablicy. Operacja dodawania elementu na koniec tablicy została zaimplementowana w sposób opisany na przykład w książce *Wprowadzenie do algorytmów* [3]: po przekroczeniu połowy dostępnego miejsca tablica zostaje powiększona dwukrotnie. Dzięki takiej implementacji operacja ta działa w amortyzowanym czasie stałym.

Dla każdego typu tablicowego zostaje także wygenerowana funkcja zwalniania pamięć zajmowaną przez tablicę. Wywołuje ona odpowiednią funkcję zwalniania pamięć dla każdego elementu, a następnie zwalnia pamięć, w której znajdowały się elementy tablicy.

6.2.2. Rekordy

Rekord był typem, od którego zaczęła się idea typów `own`. Było tak, ponieważ rekordy jako jedyny typ złożony posiadają bezpośrednie odzwierciedlenie w typach języka C, czyli `struct`. Każda definicja typu rekordowego z rodziny `own` jest kompilowana do definicji typu `struct` w języku C. Każde pole rekordu odpowiada osobnemu polu struktury o odpowiadającym typie i nazwie. Istotne jest, że typy pola struktury są wartościami, a nie wskaźnikami, co pozwoliło uniknąć niepotrzebnych dynamicznych alokacji pamięci i umożliwiło alokację struktur w całości na stosie.

Dzięki takiej reprezentacji rekordów w pamięci możliwe było skompilowanie dostępów do pól rekordu wykorzystując operatory `.` i `->` języka C, co jest znacznym usprawnieniem w stosunku do poprzedniej implementacji, używającej w tym celu tablic haszujących.

6.2.3. Typy wariantowe

Typ wariantowy, inaczej nazywany tagowaną unią (*tagged union*), jako jedyny pozwala wprowadzić w NianioLangu typy rekurencyjne. O ile rekurencyjne rekordy są niedozwolone ze względu na niemożliwość ich pełnej inicjalizacji, o tyle warianty pozwalają przezwyciężyć rekurencję na pewnym poziomie, tym samym unikając problemu nieskończonych struktur (przykład takiej definicji znajduje się na listingu 6.2). Ta cecha wymusiła przechowywanie danych w wariancie wyłącznie jako wskaźniki na właściwe obiekty. W celu oszczędności pamięci przy zachowaniu sprawdzania poprawności typów przez kompilator C definicje typów wariantowych kompilowane są do rekordów, zawierających całkowitoliczbowe pole określające, który z wariantów jest przechowywany i unię wskaźników na typy poszczególnych możliwych wartości. Zastosowanie liczby całkowitej jako wartości określającej konkretny wariant umożliwiło bardziej efektywne rozpoznawanie wariantu i skróciło czas wykonania instrukcji `match` (poprzednia implementacja do odróżniania wariantów używała nazwy wariantu z kodu, co wymuszało kosztowne porównywanie łańcuchów znaków w trakcie działania programu).

Listing 6.2: Rekurencyjna definicja struktury drzewiastej

```
def tree::node() {
    return own::var({
        node => own::rec({
            left => @tree::node,
            right => @tree::node,
            value => ptd::int(),
        }),
        leaf => ptd::none(),
    });
}
```

6.2.4. Tablice asocjacyjne

Tablice asocjacyjne, nazywane w NianioLangu hashami, to struktury podobne do tablic, ale indeksowane dowolnymi ciągami znaków. Są one zdecydowanie bardziej wysokopoziomowe od wcześniej opisanych struktur i nie mają oczywistego odpowiednika w języku C, więc zysk ze skompilowania ich jako typy `own` nie jest istotny. Niemniej jednak zostały one uwzględnione, ponieważ ze względu na ograniczenia typów `own` nie mogą one być zawarte w typach `ptd`, więc dodanie typu `own::hash` było jedynym sposobem na efektywną kompilację struktur zawartych wewnątrz tablic asocjacyjnych.

Sama implementacja dostępów do elementów tablic asocjacyjnych nie zmieniła się znacząco w porównaniu do poprzedniej implementacji. Główną różnicą jest generowanie osobnej funkcji dostępu dla każdego zdefiniowanego typu, co zapewnia zachowanie typowania na poziomie C i umożliwia pominięcie utrzymywania licznika referencji tam, gdzie jest on zbędny.

6.3. Konwersja `own` \rightarrow `im`

Konwersja typów `own` na typ `ptd::im`, który jest uniwersalną reprezentacją typów z przestrzeni `ptd`, jest naturalną potrzebą języka. Obiekty o typie z przestrzeni nazw `own` są ograniczone w zasięgu, więc chcąc zwrócić wartość takiego obiektu lub przekazać do funkcji, która przyjmuje argumenty z przestrzeni nazw `ptd`, musimy go przekonwertować.

Analiza przypadków użycia konwersji pokazała, że konwersja z typów `ptd` na typy `own` nie ma szczególnego zastosowania i nie musi być przez język obsługiwana, przede wszystkim dlatego, że tworzenie typu statycznego z typu dynamicznego w pewnym sensie mija się z celem

korzystania z typów statycznych w ogóle, gdyż narzut wydajnościowy związany z tworzeniem obiektów typu `ptd::im`, a potem konstruowanie z nich obiektów typu `own` byłby zbyt duży. W drugą stronę zaś konwersja pozwala na łączenie nowego kodu ze starym oraz odejście od statycznego typowania w chwili, gdy chcemy korzystać z funkcji obsługujących wiele typów, na przykład implementując kontenery.

Istnieje wiele sposobów na implementację konwersji. Pierwszy z nich polega na zapisaniu w bibliotece C języka ogólnej funkcji konwertujących obiekty `own` na obiekty `ptd::im`. Podejście to ma jednak wiele wad, przede wszystkim trudność implementacji, dużą podatność na błędy oraz wątpliwą przenośność rozwiązania. Szczególnie trzeci argument jest ważny, gdyż dalsze kierunki rozwoju języka prawdopodobnie przewidują przywrócenie do kompilatora możliwości kompilowania na inne języki niż C.

Drugi sposób polega na stworzeniu funkcji-wydmuszki przyjmującej w argumencie obiekt `own`, która w trakcie kompilacji podmieniana jest na właściwy kod dokonujący konstrukcji obiektu `ptd::im`. Duże problemy stwarza jednak implementacja dla typów rekurencyjnych (warianty) oraz puchnięcie kodu wynikowego, zawierającego bardzo podobne, potencjalnie bardzo długie sekcje.

Trzeci sposób, który został zrealizowany, jest niejako połączeniem dwóch poprzednich – dla każdego typu nazwanego oraz dla każdego typu, który korzysta z funkcji-wydmuszki, generowana jest funkcja, podpinana w miejsce wydmuszki. Można to w pewnym sensie porównać do generowania szablonów w C++, aczkolwiek ten system jest bardziej wyspecjalizowany.

Pierwszym elementem rozwiązania jest zebranie informacji o tym, jakie funkcje należy wygenerować, co przeprowadzane jest w dwóch miejscach. Ze względu na możliwość używania typów z innych modułów, niż w danym momencie wykonywany, każdy moduł udostępniać musi komplet funkcji do rzutowania typów, które definiuje. Najpierw, po załadowaniu drzewa AST modułu do type checkera, dokonywany jest obchód po funkcjach definiujących typy sprawdzający, czy nie definiują one typów `own`. Miejsce to zostało wybrane, gdyż jest to pierwsze miejsce, gdzie pojawia się konkretna wiedza o typach zmiennych. Następnie, podczas sprawdzania typów funkcji, za każdym razem, gdy napotykana jest funkcja-wydmuszka, zapamiętywany jest typ argumentu, z jakim została ona wywołana. Typy te są również rekurencyjnie sprawdzane, czy nie zawierają innych typów `own`, do których także należy stworzyć stosowne funkcje.

Drugim elementem rozwiązania jest generowanie kodu. Tworzona jest zmienna tekstowa, która następnie wypełniana jest instrukcjami tworzącymi obiekt `ptd::im`. Ponieważ typy mogą być zagnieżdżone, w razie potrzeb wywołują one inne funkcje konwertujące wewnętrzne obiekty `own` na `im`. Warto podkreślić, że w przypadku typów `own::rec` oraz `own::var` funkcje rzutujące nawet podobne typy są istotnie różne. Rzutując rekordy należy explicite przypisać wszystkie pola, nie jest możliwe zrobienie tego w ogólny sposób pętlą, zaś rzutując warianty wykonać należy pełną instrukcję `match`.

Trzecim elementem jest wstrzyknięcie kodu. Sprawdzanie typów rozdzielone jest od wpisywania ich do drzewa AST i w trakcie tego wpisywania podmieniane są wywołania funkcji-wydmuszki na wywołania odpowiedniej funkcji o automatycznie generowanej nazwie. Generowana nazwa jest tworzona na podstawie nazwy dla typów nazwanych, oraz na podstawie zawartości dla typów anonimowych – uzyskujemy w ten sposób deduplikację kodu, generując w kodzie wynikowym C tylko jedną funkcję dla danego typu.

Pozostaje jeszcze problem dodania definicji nowych funkcji do drzewa AST. Wygenerowany kod jest parsowany do samodzielnego modułu, który przechodzi przez type checker w sposób podobny do normalnego kodu. Po przejściu przez type checker moduły są łączone w jeden. Gdy cały program przejdzie przez type checker dalsza kompilacja przebiega tak, jak gdyby użytkownik sam utworzył sztucznie wygenerowane funkcje.

Generując funkcje, a z nich drzewo AST, można mieć wątpliwości, czy nie lepiej byłoby od razu tworzyć gotowe do podłączenia poddrzewa, czy nawet kod wynikowy w C, zamiast marnować czas procesora na przejście przez parser i type checker. Zautomatyzowane tworzenie drzewa AST daje gwarancję, że w przypadku dalszej ewolucji kompilatora nie trzeba będzie pamiętać o tym, aby każda zmiana miała odzwierciedlenie w module generowania kodu. Ze względu na dużą dynamikę rozwoju języka jest to cecha pożądana. Ponadto zmniejszanie czasu kompilacji programów nie było głównym celem projektu, a analiza przypadków użycia wykazała, że narzut czasowy jest w pełni akceptowalny.

Rozdział 7

Efekty optymalizacji i wnioski

7.1. Porównanie czasu wykonania programów

Rozdział 8

Wkład poszczególnych członków zespołu

Co zrobiliśmy w rozbiciu na osoby

Bibliografia

- [1] LK, *Wzorzec „Nianio” przykład* - 2006.
- [2] NianioLang Compiler - GitHub
- [3] Thomas H. Cormen, Charles E. Leiserson, Ron Rivest, Clifford Stein *Wprowadzenie do algorytmów*, Wydawnictwo Naukowe PWN (2013)