Metody Realizacji Języków Programowania

Obsługa wyjątków; zarządzanie pamięcią

Marcin Benke

MIM UW

15 stycznia 2015

Wyjątki

- Pojęcie wyjątek oznacza błąd (nietypową, niepożądaną sytuację).
- Obsługa wyjątków oznacza reakcję programu na wykryte błędy.
- Funkcja, która napotkała problem zgłasza (rzuca) wyjątek.
- Wyjątek jest przekazywany do miejsca wywołania funkcji, gdzie może być wyłapany i obsłużony albo przekazany wyżej. Innymi słowy poszukiwania bloku obsługi wyjątku dokonywane są po łańcuchu DL.
- Przy wychodzeniu z funkcji i bloków może zaistnieć potrzeba zwolnienia zaalokowanych w nich obiektów (np. wywołania destruktorów).

Składnia

Dla ustalenia uwagi przyjmiemy składnię C++ (składnia Javy jest w tej kwestii bardzo podobna)
Zgłoszenie wyjątku

```
throw <wyrażenie>
```

Obsługa wyjątków

```
try {
  <instrukcje>
} catch(<parametr 1>) {
  <obsługa wyjątku 1>
//...
} catch(<parametr n>) {
  <obsługa wyjątku n>
}
```

Semantyka

- Gdy któraś z instrukcji w części try przekazała wyjątek, przerywamy wykonanie tego ciągu i szukamy **catch** z odpowiednim parametrem.
- Jeśli znajdziemy, to wykonujemy obsługę tego wyjątku, a po jej zakończeniu instrukcje po wszystkich blokach **catch**.
- Jeśli nie znajdziemy, przechodzimy do miejsca wywołania (usuwając obiekty automatyczne bieżącej funkcji) i kontynuujemy poszukiwanie.
- Jeśli nie znajdziemy w żadnej z aktywnych funkcji, wykonanie programu zostanie przerwane.

Implementacja

Obsługę wyjątków można zrealizować na wiele różnych sposobów.

Bardzo istotne jest jednak to, aby narzut przy normalnym (tj. bez wystąpienia wyjątków) wykonaniu programu był minimalny, a w miarę możności zerowy.

Realizacje wymagające wykonania dodatkowych czynności na początku i końcu bloku try można uznać za nieefektywne.

Postępowanie w razie wyjątku

W momencie zgłoszenia wyjątku muszą zostać wykonane następujące czynności:

- stwierdzenie, czy nastąpiło ono wewnątrz bloku **try**,
- identyfikacja aktywnych bloków **try** — może być więcej niż jeden,
- rozpoznanie typu zgłoszonego wyjątku,
- próba dopasowania do typu wyjątku jednego z bloków **catch**,
- w wypadku powodzenia wykonanie tego bloku,
- w przeciwnym wypadku przekazanie wyjątku w górę DL .

Identyfikacja aktywnych bloków try

- W czasie wykonania programu musi być dostępna informacja (struktura danych), która dla każdej instrukcji pozwoli ustalić czy i jakie bloki **try** ją otaczają.
- Jeżeli chcemy uniknąć narzutu dla ’prawidłowego’ przebiegu programu, informacja taka musi być w całości wygenerowana w czasie kompilacji.
- Powszechnie stosowaną metodą jest uzycie tablicy indeksowanej adresami (zakresami adresów) instrukcji.
- Elementami tej tablicy będą listy odpowiednich bloków **try** lub listy odpowiednich bloków **catch**.

Przykład

```
void h()
{
  try {
    f();
    try {
      g(); // tu wyjątek
    }
    catch E1 { i1(); }
  }
  catch E2 { i2(); }
  i3();
}
```

Przykład c.d.

Załóżmy przy tym, że wygenerowany dla niej został następujący kod maszynowy:

```
0: enter
1: call f
2: call g
3: jmp 7
4: call i1
5: jmp 7
6: call i2
7: call i3
8: leave
9: ret
```

Skoku w instrukcji 3 mozna uniknąć, generujac kod np. tak

```
 h: enter
    call f
    call g
 R: call i3
    leave
    ret
C1:
    call i1
    jmp R
C2:
    call i2
    jmp R
```

Ale komplikuje to tablice wyjątków, więc dla przejrzystości zostaniemy przy pierwotnej wersji.

Przykład c.d.

Tablica, o której mowa wyglądać będzie następująco

```
Od      Do      Bloki catch
1       1       C2
2       2       C1, C2
3       5       C2
```

Ponadto dla każdego bloku catch potrzebujemy informacji o typie obsługiwanego wyjątku oraz adresie jego kodu:

```
Catch   Typ     Adres
C1      E1      4
C2      E2      6
```

Tablice te mogą łatwo zostać wygenerowane w czasie kompilacji. Uzbrojeni w nie, możemy przejść do następnego etapu: dopasowania bloku catch do typu wyjątku

Dopasowanie bloku catch do typu wyjątku

W wielu językach wyjątkiem może być dowolna wartość (obiekt). Dla każdego obiektu musi zatem istnieć możliwość stwierdzenia w czasie wykonania, czy jest on określonego typu.

Języki z wyjątkami zwykle udostępniają informacje o typach w czasie wykonania (ang. Run Time Type Information, RTTI)

Rozważmy nasz przykład poszerzony o następujące definicje:

```
class E1 {};
class E2 {};
class E3 : public E2 {};
class K {};
void g() {
   K k;
   throw (new E3());
}
```

Dopasowanie bloku catch do typu wyjątku

Catch	Typ	Adres
C1	E1	4
C2	E2	6

Wywołanie funkcji `g` powoduje zgłoszenie wyjątku. Nazwijmy jego wartość *e*.
W poprzedniej fazie ustaliliśmy, że aktywne są bloki C1, C2.
Przystępujemy zatem do dopasowania typów:

- C1 obsługuje typ E1; czy *e* jest typu E1? NIE.
- C2 obsługuje typ E2; czy *e* jest typu E2? TAK
  (jest klasy E3, która jest podklasą E2).

Wykonany powinien zostać blok C2, czyli skok pod adres 6.

Zwijanie stosu

Jeśli nie został odnaleziony żaden pasujący do typu wyjątku blok catch (w szczególności, jeśli nie byliśmy w żadnym bloku try), kontynuujemy poszukiwanie wzdłuż łańcucha DL, usuwając po drodze wszystkie obiekty automatyczne.

W naszym przykładzie:

```
void g(){
   K k;
   throw (new E3());
 }
```

należy usunąć obiekt `k` i kontynuować poszukiwania w miejscu wywołania funkcji g (czyli w funkcji h).

Proste zwijanie stosu

Najprostszą metodą realizacji takiego zachowania jest ustawienie flagi oznaczającej wyjątek, a potem zachowanie takie, jak przy powrocie z funkcji.

Kod dla wywołania funkcji musi po powrocie sprawdzić flagę wyjątku i w razie potrzeby podjąć poszukiwania bloku obsługi dla tego wyjątku.

Rozwiązanie to wprowadza pewien dodatkowy koszt także w sytuacjach, kiedy nie został zgłoszony żaden wyjątek (flagę wyjątku trzeba sprawdzać po każdym wywołaniu funkcji).

Koszt ten jest jednak dość niewielki (1-2 instrukcje procesora na wywołanie).

Pełne zwijanie stosu

Jeżeli chcemy uniknąć tego kosztu, musimy zaimplementować pełne zwijanie stosu.

W tym celu musimy przechowywać (poza stosem maszynowym) listę obiektów automatycznych.

Przy zgłoszeniu wyjątku poszukujemy po łańcuchu DL ramki stosu zawierającej odpowiedni blok catch (patrzymy na ślad powrotu) po czym usuwamy kolejno wszystkie obiekty aż do tej ramki.

Pewnej staranności wymaga rozstrzygnięcie, które obiekty z tej ostatniej ramki powinny zostać usunięte.

Oczywiście sprawa jest prostsza w językach z automatycznym zarządzaniem pamięcią (odśmiecaniem).

Zwijanie stosu a wskaźnik ramki

Powyższy algorytm (przejście po łańcuchu DL) zakłada istnienie łańcucha DL i wskaźnika ramki

Co zrobić gdy protokół wywołania pomija wskaźnik ramki? (np. `-fomit-frame-pointer`)

Możemy odtworzyć łańcuch DL przy pomocy wskaźnika stosu i śladów powrotu.

W każdym punkcie kodu musimy wiedzieć jak głęboko leży ślad powrotu

Zwijamy stos zgodnie z wytycznymi bieżącej funkcji i przechodzimy do poprzedniej ramki

Ślad powrotu wskazuje wytyczne dla poprzedniej ramki.

Zwijanie stosu a wskaźnik ramki

Przykład

```
   .cfi_startproc
// esp -> slad powrotu, frame-esp=4b
  pushl   %ebx
  .cfi_def_cfa_offset 8  // frame-esp=8b
  subl    $40, %esp
  .cfi_def_cfa_offset 48 // frame-esp=48b
  movl    $1, (%esp)
  .cfi_offset 3, -8      // ebx zapisany pod frame-
  call    _Znwj
```

NB dyrektywy CFI (Call Frame Information) są również uzywane przez debuggery.

Zarządzanie pamięcią

- Przydział pamięci
  - lista wolnych bloków; znajdowanie bloku o odpowiednim rozmiarze
  - fragmentacja wolnej pamięci
  - kompaktyfikacja
  - buddy-systems
- Zwalnianie pamięci
  - jawne (np. C)
  - automatyczne (Python, Smalltalk, Java, .NET)
  - odśmiecanie (garbage collection).

Przydział pamięci a informacje administracyjne

Często musimy przydzielić większy blok niż zamówiono

Przydzielony blok musi przechowywać pewne informacje administracyjne, np.

- rozmiar
- łącze na liście zajętych/wolnych bloków
- licznik odwołań (dla potrzeb odśmiecania)

Taki nagłówek ma zwykle stały rozmiar, powiedzmy h.

Użyteczna sztuczka: przydzielamy blok pod adresem a, na początku umieszczamy nagłówek, do programu przekazujemy adres $b = a + h$.

Nagłówek bloku b jest pod adresem $b - h$.

Jawne zwalnianie pamięci

Zalety:

- Prosta implementacja
- Dobrze określony moment wywołania destruktora (ważne jeśli ma zwalniać inne zasoby np. zamykać pliki czy połączenia)

Wady:

- Wycieki pamięci
- Trudne do wykrycia błędy
- Dodatkowy koszt programowania

Odśmiecanie

Nieużywane (niedostępne) bloki pamięci muszą być rozpoznane i zwolnione.

Podstawowe metody:

- Zliczanie odwołań (reference counting)
- Metody śledcze (tropią dostępne bloki):
  - Kopiowanie
  - Mark-sweep

Synchronizacja

Inny ważny podział metod odśmiecania:

- synchroniczne (zatrzymajcie świat, ja odśmiecam)
- asynchroniczne (równolegle z działającym programem)

Metody synchroniczne zatrzymują program na czas odśmiecania — w najlepszym wypadku dyskomfort użytkownika.
Metody asynchroniczne są zaś trudne w implementacji (i zwykle mniej skuteczne).

Konserwatywność odśmiecania

- Z punktu widzenia poprawności algorytm odśmiecania musi zagwarantować, że nigdy nie usunie dostepnego obiektu.
- Idealny odśmiecacz usuwa wszystkie niedostepne obiekty natychmiast gdy stają się niedostepne.
- Realne odśmiecacze nie usuwają wszystkich śmieci, przynajmniej nie od razu.
- Z tej przyczyny mówimy o **konserwatywności** odśmiecaczy (zachowują niektóre śmieci) i jej stopniach (jeden algorytm zachowuje więcej śmieci niż drugi).

Zliczanie odwołań

- Każdy obiekt przechowuje licznik wskaźników, które doń prowadzą.
- Każde przypisanie wskaźnika modyfikuje odpowiednie liczniki; gdy licznik dojdzie do 0 — zwalniamy obiekt.
- Zalety:
  - prosta w implementacji metoda asynchroniczna.
  - dobrze określony moment wywoływania finalizatorów
- Wady:
  - narzut czasowy i pamięciowy
  - niezwalnianie niedostępnych cykli (np. listy dwukierunkowe).
- Czasem stosowane w połączeniu z innymi metodami.

Odśmiecanie śledcze

Poczynając od zbioru korzeni (np stos, zmienne globalne) śledzimy które obiekty są (a raczej mogą być) używane. Pozostałe są śmieciami.
Problemy:

- Wybór korzeni (które komórki stosu? rejestry?)
- Rozpoznawanie wskaźników
- Narzuty czasowe (przejście całej zaalokowanej pamięci) lub pamięciowe
- Lokalność (stronicowanie, cache)

Odśmiecanie kopiujące

- Dostępną pamięć dzielimy na dwa równe obszary;
- w jednym z nich alokujemy nowe obiekty, drugi pusty.
- Gdy zabraknie pamięci, wykrywamy dostępne obiekty i przenosimy je do drugiego obszaru.
- Po zakończeniu przenosin w pierwszym obszarze pozostają same śmieci, więc możemy uznać go za pusty...
- ...i zamienić obszary rolami.

Cena:

- tylko połowa dostępnej pamięci jest rzeczywiście używana;
- za to koszt czasowy proporcjonalny do rozmiaru dostępnych obiektów.
- dwupoziomowe wskaźniki (lub trudne mechanizmy zmiany wskaźników).

“Zaznacz i zamieć” (mark and sweep)

- Odznaczamy wszystkie obiekty
- Poczynając od korzeni, obchodzimy graf dostępnych obiektów
- Które rejestry zawierają wskaźniki?
- Krawędziami są wskaźniki, jak je wykryć?
  - znaczniki
  - jeśli wygląda jak wskaźnik, bezpiecznie założyć że jest wskaźnikiem (konserwatywność)
- Wykrywamy dostępne obiekty i zaznaczamy je jako dostępne
- Po zakończeniu zaznaczania, wszystkie niezaznaczone obszary są śmieciami.
- Przechodzimy wszystkie obiekty i niezaznaczone usuwamy.

“Zaznacz i zamieć” (mark and sweep)

Koszty:

- przechowywanie listy wszystkich zaalokowanych obiektów;
- znaczniki dostepności (w nagłówku lub odrębna tablica)
- stos/kolejka dla obejścia grafu (lub odwracanie wskaźników)
- koszt czasowy proporcjonalny do łącznego rozmiaru pamięci.

Dla skrócenia przerw, odznaczanie i zamiatanie może wykonywać alokator.

Wariant “ze zgniataniem”: dla uniknięcia fragmentacji przesuwamy dostępne obiekty do spójnego obszaru.

Rozpoznawanie wskaźników

Jak odróżnić wskaźnik od obiektu pierwotnego, np int?

- Etykietowanie — int ma swoją etykietę, wskaźnik inną
  narzuty czasowe, niepełny zakres int
- Mapa stosu
  w silnie typowanych językach można statycznie określić
- Konserwatywne — jesli coś wygląda jak wskaźnik, uznajemy za wskaźnik
  gdy nie wiemy co jest wskaźnikiem nie możemy przesuwać obiektów

Odśmiecanie generacyjne

Hewitt, 1987:

- większość obiektów umiera młodo
- nowe obiekty częściej zawierają wskaźniki do starszych niż na odwrót

Pomysł:

- “Szkółka” dla młodych obiektów z odśmiecaniem kopiującym (tzw. małe odśmiecanie)
- Szkółka jest mała i w większości zaśmiecona (obiekty umierają młodo)
- Gdy mało miejsca w szkółce po odśmiecaniu, przenieś obiekty do dużej przestrzeni

Wymaga specjalnego traktowania wskaźników od starych do młodych obiektów.

Odśmiecanie przyrostowe (współbieżne, asynchroniczne)

- Często zatrzymanie programu (na bliżej nieokreślony czas) dla przeprowadzenia odśmiecania nie jest akceptowalne.
- Wtedy odśmiecanie musi być synchronizowane z działaniem programu.
- **Problem:** Podczas gdy Odśmiecacz przechodzi graf dostępnych obiektów, Program moze ten graf zmieniać.
- Z tej przyczyny z punktu widzenia odśmiecacza, program nazywany jest Zmieniaczem (ang. mutator).
- Odśmiecacz też może zmieniać fragmenty grafu, których używa Zmieniacz — konieczna pełna współbieżność.

Trójkolorowe zaznaczanie (Dijkstra, Lamport et al.)

Algorytmy odśmiecania mogą być opisane jako proces obchodzenia i kolorowania grafu obiektów:

- obiekty podlegające odśmiecaniu mają kolor biały
- na końcu odśmiecania obiekty dostępne mają mieć kolor czarny
- Dla synchronizacji Odśmiecacza i Zmieniacza wprowadzamy trzeci kolor: szary
- Obiekt jest szary, jeśli został już odwiedzony, ale jego potomkowie niekoniecznie.
- **Niezmiennik:** żaden czarny obiekt nie moze zawierać wskaźnika do białego.
- Postęp obejścia odbywa się w “szarej strefie” (białe obiekty stają się szare, szare stają się czarne).
- Gdy nie ma już szarych obiektów, białe są śmieciami

Synchronizacja odśmiecania — bariery

- Bariera odczytu: odczyt wskaźnika do białego obiektu powoduje, że staje się on szary:
  - jest dostępny,
  - Zmieniacz nigdy nie dostanie wskaźnika do białego obiektu.
- Bariera zapisu: różne mechanizmy zapewniające zachowanie niezmiennika przy zapisie wskaźników, np:
  - Steele: zapis wskaźnika do białego obiektu zmienia go w szary (ale teraz trzeba udowodnić postęp algorytmu)
  - Dijkstra: zapis wskaźnika zmienia obiekt wskazywany w czarny (oczywisty postęp algorytmu, ale bardziej konserwatywne).
  - Nowe obiekty mogą być oznaczane jako czarne (bardziej konserwatywne) lub białe (licząc na to, ze nowe obiekty żyją krócej niż stare).