

Kỹ thuật dịch ngược

Ngày 5 tháng 12 năm 2016

Lời cam đoan

Lời cảm ơn

Tóm tắt luận văn

Mục lục

1	Giới thiệu	1
1.1	Kỹ thuật dịch ngược và ứng dụng	1
1.2	Bài toán đặt ra	3
1.3	Cấu trúc luận văn	5
2	Các kiến thức nền tảng và nghiên cứu liên quan	6
2.1	Trình biên dịch	6
2.2	Trình dịch ngược	8
2.3	Một số kỹ thuật tiêu biểu được sử dụng trong các công cụ dịch ngược	9
2.3.1	Lan truyền biểu thức	9
2.3.2	Loại bỏ mã chết	11
2.3.3	Mã SSA	11
2.4	Tình hình phát triển trình dịch ngược hiện nay	12
2.5	Cấu trúc của Boomerang	13
3	Kiểm tra kiểu - Type checking	17
3.1	Rút trích thông tin từ phần chú thích	17
3.2	Kiểm tra nguyên tắc sử dụng bộ biến	19
3.2.1	Xác định giá trị kiểu của thanh ghi <i>ACC</i> tại mỗi điểm của chương trình	19
3.2.2	Kiểm tra tính hợp lệ về kiểu của tác vụ với bit	25
4	Suy luận kiểu - Type inference	26
4.1	Xác định giá trị của thanh ghi <i>ACC</i> tại mỗi thời điểm của chương trình	26
4.2	Tìm kiếm mối quan hệ giữa các biến	31
4.2.1	Chuyển đổi giữa hằng số nguyên - biến byte tương ứng	31
4.2.2	Kiểm tra và ghi nhận mối quan hệ giữa các giá trị - biến bit	32
5	Kiểm thử	33
5.1	Hệ thống testcase	33
5.2	Phương thức kiểm thử	34
5.3	Kết quả chạy thử	34
6	Kết luận	35
6.1	Kết quả đạt được, khó khăn, điểm hạn chế	35
6.2	Hướng phát triển trong tương lai	35

Danh sách hình vẽ

1.1	Một ứng dụng của trình dịch ngược: chuyển đổi mã nguồn giữa các kiến trúc máy khác nhau	2
1.2	Các hướng tiếp cận được trình bày trong luận văn	5
2.1	Phân tích một câu lệnh thành các token	7
2.2	Cây cấu trúc cho một câu lệnh gán	7
2.3	Ví dụ về lỗi kiểu biến	7
2.4	Đoạn mã gốc và đoạn mã được dịch ngược bởi trình dịch ngược Boomerang	8
2.5	Các bước cơ bản của một trình dịch ngược	9
2.6	Một đoạn mã trước khi thực hiện lan truyền biểu thức	9
2.7	Đoạn mã ở hình 2.6 sau khi thực hiện lan truyền biểu thức	10
2.8	Đoạn mã trung gian với bốn câu lệnh gán đơn giản	10
2.9	Đoạn mã ở hình 2.8 sau khi loại bỏ mã chết	11
2.10	Đoạn mã trung gian với 3 lần định nghĩa biến a	11
2.11	Đoạn mã ở hình 2.11 đã được chuyển sang dạng mã SSA	12
2.12	Một đoạn mã được dịch ngược bởi trình dịch ngược ILSpy. Tên biến static abc được giữ nguyên như mã gốc	13
2.13	Một đoạn mã được dịch bởi Boomerang	13
2.14	Cấu trúc các khối lớn của Boomerang	14
2.15	Cấu trúc dữ liệu lưu trữ mã assembly trong Boomerang	15
2.16	Cách lưu trữ một chương trình dưới dạng mã trung gian của Boomerang	15
3.1	Sơ đồ các bước thực hiện giải pháp Kiểm tra kiểu	17
3.2	Một đoạn chương trình mẫu	19
3.3	Giải thuật tính Reaching definitions cho khối cơ bản	21
3.4	Giải thuật cho hàm findRegValue - phần mở rộng của phân tích Reaching definitions	23
3.5	Giải thuật của phân tích Type propagation	24
3.6	Quá trình kiểm tra một câu lệnh sử dụng bit	25
4.1	Giải thuật Constant propagation đã được điều chỉnh phù hợp với yêu cầu của trình dịch ngược	29
4.2	Các giải pháp chuyển đổi giá trị - biến byte	31
4.3	Các bước kiểm tra và ghi nhận dữ liệu vào danh sách UnionDefine	32

Listings

1.1	Một đoạn mã 8051	3
1.2	Một ví dụ vi phạm nguyên tắc sử dụng bộ biến	4
1.3	Một đoạn khai báo union	4
2.1	Đoạn mã mô tả cách biểu diễn giá trị của tham số trong Boomerang .	14
3.1	Mẫu khai báo bộ biến	17
3.2	Cấu trúc dữ liệu để lưu trữ một bộ biến	18
3.3	Đoạn mã parser cho phần khai báo bộ biến	18
3.4	Một số câu lệnh gán mà phương pháp Suy luận kiểu sử dụng Reaching definitions không xử lý được	22
3.5	Đoạn mã có nhiều câu lệnh khai báo cho ACC đến được một điểm của chương trình nhưng tất cả đều cùng giá trị	22
4.1	Trường hợp không thể xử lý được bằng các phương pháp phân tích dữ liệu trước	26
4.2	Đoạn mã trước khi thực hiện lan truyền hằng số	27
4.3	Đoạn mã sau khi thực hiện lan truyền hằng số cho biến y	27
4.4	Đoạn mã sau khi thực hiện lan truyền hằng số cho biểu thức trả về .	27
4.5	Đoạn mã ví dụ biến có giá trị là hằng số	27
4.6	Đoạn mã ví dụ biến có giá trị là bottom	27
4.7	Đoạn mã thể hiện class ConstantVariable	28
4.8	Một số câu lệnh gán cho ACC có giá trị về phải bằng nhau	30
4.9	Đoạn mã mới của class UnionDefine	32

Chương 1

Giới thiệu

Chương này nhằm giới thiệu về bài toán sẽ được giải quyết trong luận văn và các khái niệm liên quan. Đầu chương sẽ trình bày về kỹ thuật dịch ngược, các ứng dụng của nó và những khó khăn trong quá trình dịch ngược. Phần tiếp theo nêu bài toán đặt ra và các thách thức khi giải quyết bài toán. Phần cuối cùng sẽ tóm tắt cấu trúc của luận văn.

1.1 Kỹ thuật dịch ngược và ứng dụng

Trong khi kỹ thuật dịch phổ biến hiện nay là dịch từ mã viết bằng ngôn ngữ cấp cao xuống mã ngôn ngữ cấp thấp hơn, kỹ thuật dịch ngược thực hiện dịch từ mã ngôn ngữ cấp thấp lên mã ngôn ngữ cấp cao hơn. Kỹ thuật dịch ngược được sử dụng rất nhiều để hỗ trợ trong quá trình phát triển và sử dụng phần mềm:

- Vì một lý do nào đó, mã nguồn của một phần mềm bị mất đi. Để tiếp tục phát triển phần mềm hoặc bảo trì phần mềm đó, ta cần phải khôi phục lại mã nguồn. Nếu viết lại một chương trình mới hoàn toàn từ các tài liệu sẵn có sẽ rất mất thời gian và không đảm bảo sẽ tương đương được phần mềm cũ. Vì vậy một giải pháp phổ biến hiện nay là dựa vào file thực thi dịch ngược lại và hiệu chỉnh để có được mã nguồn mới hoàn chỉnh.
- Các phần mềm độc hại như virus, malware thường sẽ giấu kín mã nguồn. Kỹ thuật dịch ngược giúp sinh ra mã nguồn của chúng, qua đó, việc tìm ra phương pháp giải trừ sẽ dễ dàng hơn.
- Một số chương trình được viết để chạy trên các chip đã lỗi thời, như chip 8051, sẽ ngừng sản xuất trong tương lai gần, cần phải được chuyển đổi để chạy được trên những chip hiện đại đang được sản xuất. Một trong những giải pháp để giải quyết vấn đề này là dùng trình dịch ngược để chuyển chương trình viết trên chip lỗi thời sang ngôn ngữ cấp cao, sau đó dùng trình biên dịch để dịch thành mã của chip thay thế.
- Phần mềm viết bằng ngôn ngữ A cần phải chuyển đổi sang ngôn ngữ B để tiếp tục bảo trì và phát triển. Ngôn ngữ A có thể là một ngôn ngữ đã ra đời từ rất lâu (ví dụ: COBOL, Basic...), hiện nay không còn người hiểu biết về ngôn ngữ đó để lập trình phần mềm. Vì vậy, cần phải chuyển đổi phần mềm sang một ngôn ngữ khác mới hơn, có nhân lực để viết tiếp (ví dụ: Java, C#...). Quá



Hình 1.1: Một ứng dụng của trình dịch ngược: chuyển đổi mã nguồn giữa các kiến trúc máy khác nhau

trình này cũng được xem là dịch ngược, vì thường ngôn ngữ A ra đời trước sẽ có mức độ trừu tượng thấp hơn là các ngôn ngữ B được phát triển sau này.

Một trong những thách thức khó nhất của kỹ thuật dịch ngược là khôi phục thông tin. Do ở các ngôn ngữ cấp thấp không có phương tiện để lưu trữ một số thông tin cần thiết ở ngôn ngữ cấp cao, nên những thông tin đó sẽ bị mất đi trong quá trình dịch xuôi hoặc lập trình bằng ngôn ngữ cấp thấp. Một số thông tin cần khôi phục là:

- Kiểu dữ liệu của biến: Đối với các chương trình viết bằng ngôn ngữ cấp cao, kiểu dữ liệu của biến có thể xem như một ràng buộc khi gán giá trị cho biến và sử dụng biến. Ví dụ khi ta khai báo một biến có kiểu dữ liệu là integer, thì ta phải gán cho biến các giá trị là số nguyên (1, 2,...) và sử dụng biến trong các phép toán có toán tử là số nguyên. Nếu ta gán cho biến một giá trị khác số nguyên (số thực, chuỗi, boolean...) hoặc sử dụng biến trong các phép toán không chấp nhận toán tử là số nguyên thì trình biên dịch sẽ phát hiện lỗi ngay ở giai đoạn đầu. Tuy nhiên, đối với một số mã máy, kiểu dữ liệu không cần thiết và sẽ được loại bỏ trong quá trình biên dịch. Khi dịch ngược, nếu không khôi phục được kiểu dữ liệu thì sẽ không đủ thông tin để xây dựng mã đầu ra.
- Tên của biến: Ở ngôn ngữ cấp cao, tên biến mang ngữ nghĩa là công dụng của biến đó, và được dùng để truy xuất giá trị của biến. Còn ở ngôn ngữ cấp thấp, dữ liệu sẽ được lưu vào các thanh ghi có sẵn hoặc vùng nhớ được truy xuất bằng địa chỉ trực tiếp, vì vậy việc tên biến ở cấp độ này là không cần thiết và sẽ bị loại bỏ. Nếu trình dịch ngược không giữ được tên biến của chương trình gốc thì rất khó để phát triển và bảo trì. Giải pháp hiện nay của các trình dịch ngược là sinh ra tên biến tự động, sau đó dựa vào các tài liệu sẵn có để chỉnh sửa tên biến bằng tay ở chương trình đầu ra.
- Phân biệt giữa dữ liệu và mã điều khiển: Đặc điểm của một số mã máy (trừ mã máy chạy trên máy ảo) là dữ liệu và các câu lệnh điều khiển có cùng một

định dạng mã nhị phân và được lưu trong cùng một vùng nhớ. Vì vậy, khi dịch ngược từ mã máy lên cần phải phân biệt được phần nào của vùng nhớ là lưu các dữ liệu và phần nào là câu lệnh của chương trình.

Từ khái niệm của kỹ thuật dịch ngược, ta thấy có nhiều mức độ dịch ngược, tương ứng với những bài toán khác nhau cần giải quyết. Nếu lấy đầu ra của quá trình dịch ngược là một chương trình viết bằng ngôn ngữ cấp cao, thì đầu vào của nó có thể là: mã nhị phân, mã assembly hoặc mã của một ngôn ngữ lập trình cấp cao khác cần chuyển đổi. Tùy vào mức độ trừu tượng của ngôn ngữ đầu vào, các thông tin bị mất ở mã đầu vào sẽ khác nhau. Với mã máy thì tất cả các thông tin nêu trên đều không còn. Với mã assembly, tên biến vẫn xuất hiện trong chương trình vì một số assembler cho phép có các câu lệnh khai báo biến ở mã assembly. Còn với mã ngôn ngữ cấp cao thì gần như tất cả thông tin đều có ở chương trình gốc, và vấn đề cần giải quyết là tìm ra các cấu trúc tương đương ở ngôn ngữ đích.

1.2 Bài toán đặt ra

Như đã trình bày ở trên, kiểu dữ liệu là một thông tin quan trọng không xuất hiện ở mã cấp thấp nhưng cần phải có để xây dựng mã đầu ra của trình dịch ngược. Vì vậy, bài toán suy luận kiểu là một thách thức cấp bách cần phải giải quyết trong kỹ thuật dịch ngược. Hiện nay, các giải thuật cho vấn đề này đã tương đối đầy đủ, đưa ra kết quả kiểu dữ liệu chấp nhận được. Tuy nhiên, có vấn đề chưa được giải quyết thấu đáo là ở một số kiến trúc máy, người lập trình có thể vừa truy xuất một vùng nhớ lớn, vừa truy xuất một vùng nhớ nhỏ hơn thuộc vùng nhớ lớn đó, đồng thời có thể đặt tên các vùng nhớ. Vì việc thay đổi giá trị một vùng nhớ nhỏ cũng sẽ làm thay đổi giá trị của vùng nhớ lớn hơn, nên không thể xem vùng nhớ nhỏ như một biến độc lập, mà cần phải gom nhóm các biến đại diện cho vùng nhớ nhỏ và vùng nhớ lớn thành một cấu trúc dữ liệu phù hợp ở ngôn ngữ cấp cao.

Một kiến trúc máy có những tính chất trên là 8051. Như trong ví dụ 1.1, có 2 biến được khai báo là *OPTIONS* với giá trị là **38H**, và *TESTSUPS* đại diện cho bit thứ nhất của thanh ghi *ACC*. Ở đoạn mã thực thi bên dưới, thanh ghi *ACC* được gán cho giá trị ở vùng nhớ có địa chỉ là giá trị của biến *OPTIONS* (câu lệnh số 1) và sau đó giá trị của *TESTUPS* được chuyển thành 1, nghĩa là bit đầu tiên của thanh ghi *ACC* bật lên. Qua thấy được thanh ghi *ACC* chỉ là trung gian, và thực chất *TESTUPS* đại diện cho bit đầu tiên của vùng nhớ quy định bởi biến *OPTIONS*. Việc sử dụng này đưa đến kết luận là biến *OPTIONS* và *TESTUPS* là cùng một bộ biến vì đều truy xuất đến toàn bộ hoặc một phần vùng nhớ có địa chỉ là **38H**.

Listing 1.1: Một đoạn mã 8051

```
#DEFINE OPTIONS, #38H
#DEFINE TESTSUPS ACC.1
public AA
AA:
    MOV A, OPTIONS #1
    SETB TESTUPS #2
```

Trong bộ biến này, *OPTIONS* được gọi là **biến byte**, vì nó truy xuất đến vùng nhớ có độ lớn là 1 byte. Còn các biến truy xuất đến một bit nào đó của vùng nhớ

như *TESTSUPS* sẽ được gọi là **biến bit**. Thông thường, người ta luôn sử dụng các biến byte - biến bit này theo bộ và việc sử dụng các bộ biến này phải tuân thủ nguyên tắc sau:

- Chỉ khi thanh ghi *ACC* đang mang giá trị của vùng nhớ có địa chỉ quy định bởi biến byte, thì các biến bit cùng bộ mới được sử dụng.
- Mỗi biến bit chỉ thuộc một bộ duy nhất. Nói cách khác, nếu mỗi thời điểm sử dụng biến bit, thanh ghi *ACC* đang mang giá trị của những vùng nhớ khác nhau, thì việc sử dụng này đã vi phạm nguyên tắc.
- Tại mỗi vị trí bit của một bộ biến chỉ có một biến bit duy nhất tồn tại. Một ví dụ vi phạm nguyên tắc này là ở đoạn mã 1.2. Khi thanh ghi *ACC* đang mang giá trị vùng nhớ quy định bởi *OPTIONS*, có hai biến bit cùng truy xuất đến bit đầu tiên của vùng nhớ được sử dụng là *TESTSUPS* và *TESTSUPS1*. Như vậy, ta không thể chọn được biến bit nào sẽ đại diện cho bit đầu tiên của vùng nhớ giữa hai biến trên.

Listing 1.2: Một ví dụ vi phạm nguyên tắc sử dụng bộ biến

```
#DEFINE OPTIONS #38H
#DEFINE TESTSUPS ACC.1
#DEFINE TESTSUPS1 ACC.1

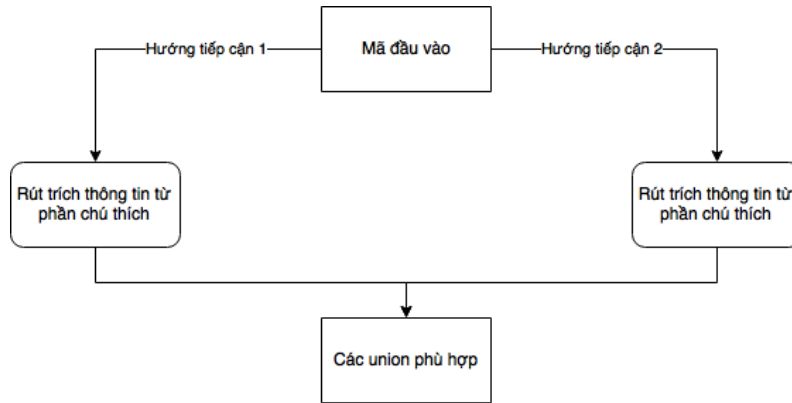
public AA
AA:
MOV A, OPTIONS
SETB TESTSUPS
CLR TESTSUPS1
```

Vì 8051 đáp ứng đầy đủ các yêu cầu về một kiến trúc máy cho phép truy xuất vùng nhớ ở nhiều cấp độ và cho phép đặt tên biến cho vùng nhớ, nên các phân tích sau này của luận văn sẽ chủ yếu dựa vào các đoạn mã của 8051. Tuy nhiên, các nghiên cứu này hoàn toàn có thể áp dụng cho các kiến trúc máy có tính chất tương tự như vậy. Về sau trong luận văn này, khi đề cập đến nguyên tắc sử dụng bộ biến, nghĩa là đang nói đến các nguyên tắc nêu trên.

Union là một cấu trúc dữ liệu có nhiều thành phần với những kiểu dữ liệu khác nhau, nhưng cùng chia sẻ một vùng nhớ. Trong ví dụ 1.3, hai thành phần của union *ci* là ký tự *c* và số nguyên *i* đều có vùng nhớ giống nhau. Khi thay đổi một thành phần như ở câu lệnh số 3, thành phần còn lại của union cũng thay đổi theo. Giá trị *ci.i* được xuất ra ở câu lệnh số 2 là **0** (giá trị mặc định của một số nguyên chưa khai báo), còn ở câu lệnh số 4 là **97**. Như vậy, tính chất này của union là tương tự với tính chất truy xuất vùng nhớ đã nêu ở trên của một số kiến trúc máy.

Listing 1.3: Một đoạn khai báo union

```
union {char c; int i;} ci;
cout << i;
ci.c = 'a';
cout << i;
```



Hình 1.2: Các hướng tiếp cận được trình bày trong luận văn

Tóm lại, luận văn sẽ giải quyết vấn đề gom nhóm biến truy xuất toàn bộ một vùng nhớ và những biến truy xuất một phần vùng nhớ đó, đưa chúng về dưới dạng cấu trúc dữ liệu union ở ngôn ngữ cấp cao.

Để giải quyết vấn đề này, có hai hướng tiếp cận được đề ra trong luận văn. Đó là:

- Dựa trên một số thông tin trong phần chú thích của người lập trình viên. Trong quá trình lập trình, để biết được các biến nào cùng một bộ với nhau, người lập trình phải có chú thích ở phần khai báo của các biến đó. Giải pháp đề ra là đưa ra một mẫu quy định các chú thích này, yêu cầu người lập trình sửa chữa chú thích của mình theo mẫu đó, sau đó dùng parser rút trích thông tin từ chú thích để tìm ra các bộ biến và kiểm tra việc sử dụng biến bên dưới (Kiểm tra kiểu - Type checking).
- Dựa vào quá trình sử dụng các biến trong chương trình. Việc chú thích là không bắt buộc với lập trình viên, nên một số đoạn mã có thể bị thiếu chú thích hoặc chú thích không đầy đủ. Cách giải quyết khả thi là xem xét quá trình sử dụng các biến, phân tích luồng đi của dữ liệu và tìm ra được các union phù hợp. (Suy luận kiểu - Type reference).

1.3 Cấu trúc luận văn

Luận văn gồm có 6 chương. Chương tiếp theo sẽ trình bày về các kiến thức nền tảng, cũng như một số nghiên cứu liên quan đến kỹ thuật dịch ngược. Ngoài ra, chương này cũng chỉ rõ kiến trúc của trình dịch ngược Boomerang, trình dịch ngược được dùng để hiện thực các giải pháp của bài toán. Chương 3 giới thiệu hướng tiếp cận đầu tiên của bài toán là rút trích thông tin từ các chú thích của người lập trình. Chương 4 sẽ làm rõ hướng tiếp cận tiếp theo là dựa vào cách sử dụng biến trong đoạn mã đầu vào. Chương 5 nêu ra một số thiết lập cần thiết trên trình dịch ngược Boomerang để kiểm tra kết quả của hai hướng tiếp cận trên, cách thiết lập các mẫu thử (testcase) và kết quả chạy thử. Chương 6 kết luận về kết quả đạt được của luận văn và đề ra các hướng nghiên cứu tiếp theo.

Chương 2

Các kiến thức nền tảng và nghiên cứu liên quan

Chương này sẽ trình bày về một số kiến thức nền tảng như: trình biên dịch, trình dịch ngược, một số kỹ thuật thường được sử dụng trong trình dịch ngược và kiến trúc của trình dịch ngược Boomerang. Việc vì sao lựa chọn trình dịch ngược Boomerang để hiện thực các giải pháp cho bài toán cũng sẽ được bàn đến trong chương này.

2.1 Trình biên dịch

Trình biên dịch (compiler) là một chương trình hoặc một bộ chương trình máy tính, có nhiệm vụ biến đổi mã nguồn được viết bằng một ngôn ngữ lập trình này (ngôn ngữ lập trình gốc) sang một ngôn ngữ lập trình khác (ngôn ngữ lập trình đích), thường sẽ có dạng nhị phân và thực thi được.

Các bước của trình biên dịch gồm có:

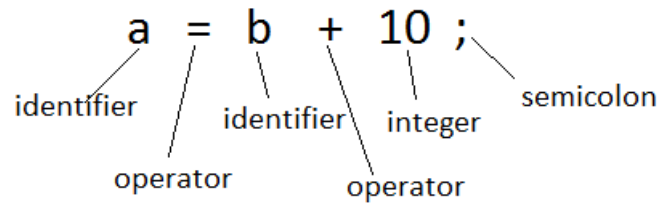
- Phân tích từ vựng (lexical analysis): Đây là quá trình chuyển hóa một chuỗi các ký tự (ví dụ như các câu lệnh trong một chương trình máy tính) thành chuỗi các từ tố (token), ví dụ như: định danh, số nguyên, số thực... Giai đoạn này kiểm tra các lỗi về từ vựng, chính tả của chương trình.

Trong hình 2.1, các ký tự a , b được nhận diện là các định danh (identifier), 10 được nhận diện là số nguyên (integer), $=$ và $+$ được nhận diện là các phép tính (operator), $;$ là ký tự đặc biệt kết thúc một câu lệnh.

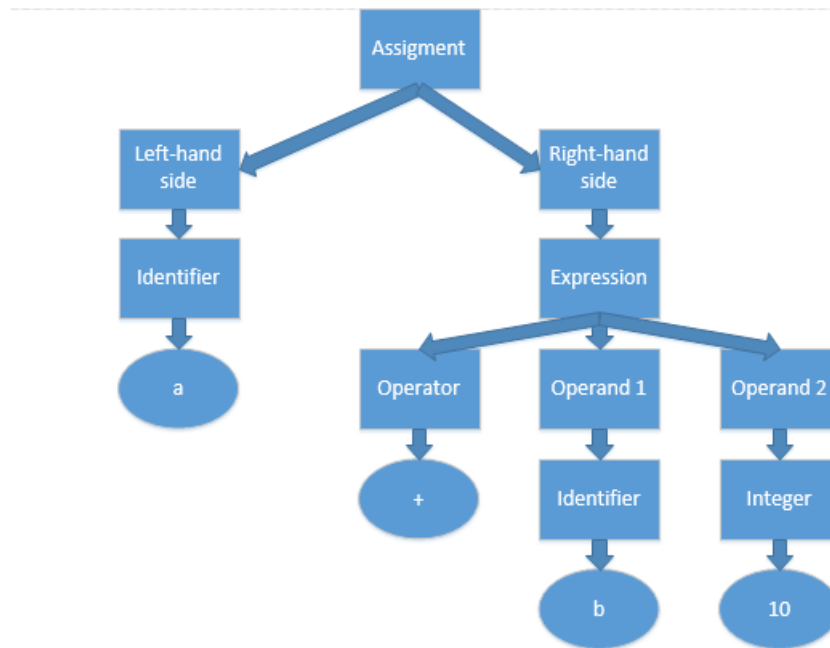
- Phân tích cú pháp (syntax analysis): Từ chuỗi các từ tố được tạo ra ở giai đoạn trên, một chương trình gọi là parser sẽ tạo ra một cấu trúc dữ liệu, thường là parse tree hoặc abstract syntax tree. Giai đoạn này sẽ kiểm tra các lỗi về cấu trúc ngữ pháp.

Câu lệnh gán từ ví dụ trên sẽ được phân tích thành cây cấu trúc như hình 2.2

- Phân tích ngữ nghĩa (semantic analysis): Trong giai đoạn này, từ cây cấu trúc đã có, trình biên dịch sẽ áp dụng các luật về ngữ nghĩa để kiểm tra tính đúng đắn của chương trình. Thường sẽ là các luật về kiểu dữ liệu, kiểm tra tầm vực của biến và object binding.



Hình 2.1: Phân tích một câu lệnh thành các token



Hình 2.2: Cây cấu trúc cho một câu lệnh gán

Tiếp tục ví dụ ở trên, trong giai đoạn này, trình biên dịch sẽ kiểm tra xem các biến `a` và `b` đã được khai báo chưa, phạm vi của các biến có phủ tới vị trí của câu lệnh không (ví dụ: có những biến được khai báo ở hàm `A` thì sẽ không có phạm vi ở bên ngoài hàm `A`), kiểu của biến có phù hợp với câu lệnh gán không (ví dụ: nếu `b` có kiểu là **string** thì câu lệnh trên không hợp lệ).

<code>int a, b;</code>	<code>int a;</code>
<code>b = 20;</code>	<code>string b = "hello";</code>
<code>a = b + 10;</code>	<code>a = b + 10;</code>

Hình 2.3: Ví dụ về lỗi kiểu biến

Trong hình 2.3, cả hai đoạn mã đều hợp lệ tính đến cuối giai đoạn phân tích cú pháp. Tuy nhiên, giai đoạn phân tích ngữ nghĩa sẽ phát hiện ra đoạn mã ở bên phải không hợp lệ vì nó vi phạm các ràng buộc về kiểu.

- Tạo ra mã trung gian: Sau khi trải qua các giai đoạn phân tích và kiểm tra,

trình biên dịch sẽ tiến hành sinh mã trung gian từ mã nguồn. Đặc điểm của mã trung gian là đơn giản và rất gần với mã đích, tuy nhiên con người vẫn có thể đọc và hiểu được. Việc sinh mã trung gian nhằm giảm thiểu chi phí cho trình biên dịch khi phải sinh mã đích cho nhiều kiến trúc máy khác nhau. Thay vì với mỗi kiến trúc máy, trình biên dịch phải tạo ra công cụ riêng để dịch từ mã nguồn sang mã đích, thì ở đây chỉ cần tạo ra công cụ để dịch từ mã trung gian - vốn đã rất gần với mã đích.

- Tạo mã đích: Từ mã trung gian, tùy vào kiến trúc máy sẽ thực thi chương trình, trình biên dịch sẽ tạo ra mã đích tương ứng. Giai đoạn này sẽ thực hiện các công việc như: lựa chọn câu lệnh trung gian sẽ thực hiện, quyết định các giá trị được lưu trong thanh ghi, sắp xếp thứ tự thực hiện các câu lệnh. Đầu ra của giai đoạn là mã máy có thể thực thi được.
- Tối ưu mã đích: Để tăng tốc độ thực hiện chương trình cũng như giảm các chi phí chạy chương trình, giai đoạn tối ưu mã đích sẽ kiểm tra và áp dụng các kỹ thuật nhằm loại bỏ mã chết, tối ưu vòng lặp, loại bỏ dư thừa... Giai đoạn này không nhất thiết chỉ thực hiện ở cuối quá trình biên dịch mà có thể nằm ở bất cứ đâu.

2.2 Trình dịch ngược

Mục tiêu của trình dịch ngược là chuyển đổi chương trình được viết bằng một ngôn ngữ cấp thấp (thường là mã máy) lên một ngôn ngữ cấp cao hơn (như C, C++...). Vì vậy, trình dịch ngược (decompiler) có thể xem như một quá trình đảo ngược của trình biên dịch (compiler). Chương trình đầu ra phải thực hiện được những chức năng tương đương như chương trình đầu vào.

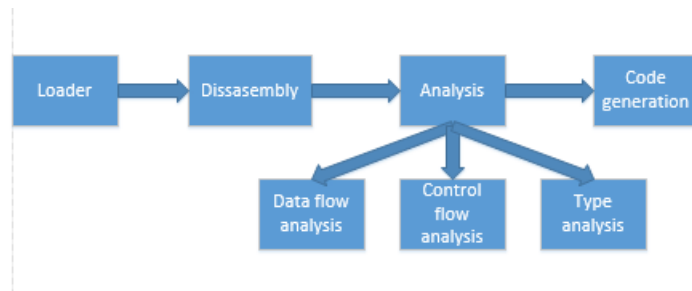
<pre>#include<stdio.h> main() { printf("Hello World"); }</pre>	<pre>// address: 0x804841d int main(int argc, char *argv[], char *envp[]) { __size32 eax; // r24 printf("Hello World"); return eax; }</pre>
---	--

Hình 2.4: Đoạn mã gốc và đoạn mã được dịch ngược bởi trình dịch ngược Boomerang

Quá trình dịch ngược có thể chia thành các giai đoạn sau:

- Loader: Load file cần dịch ngược, đọc từ file ra các thông tin như: loại file, loại kiến trúc máy... và xác định được ngõ vào của chương trình (tương đương với hàm main trong C).
- Disassembly: Mã gốc sẽ được chuyển thành mã trung gian, mã trung gian là gì thì tùy vào trình dịch ngược. Ví dụ Boomerang sẽ dùng mã trung gian là Register Transfer Language.

- **Analysis:** Sau khi đã chuyển sang mã trung gian, chương trình sẽ đi qua các bước phân tích để khôi phục lại thông tin đã mất trong quá trình biên dịch. Các phân tích thường phải có là: lan truyền biểu thức, loại bỏ mã chết, xác định nguyên mẫu hàm (function prototype), xác định kiểu dữ liệu...
- **Code generation:** Trải qua các kỹ thuật phân tích để xác định được thông tin cần thiết về dữ liệu, kiểu và luồng điều khiển chương trình, giai đoạn cuối cùng của dịch ngược là sinh ra mã chương trình bằng ngôn ngữ bậc cao.



Hình 2.5: Các bước cơ bản của một trình dịch ngược

2.3 Một số kỹ thuật tiêu biểu được sử dụng trong các công cụ dịch ngược

2.3.1 Lan truyền biểu thức

Lan truyền biểu thức (Expression propagation) là biến đổi phổ biến nhất trong quá trình dịch ngược một đoạn code. Nguyên tắc truyền biểu thức cũng rất đơn giản: Với các câu lệnh sử dụng giá trị của một biến nào đó, ta có thể thay tên biến đó bằng biểu thức nằm bên phải câu lệnh gán biến đó.

Hình 2.6 và 2.7 là một ví dụ cho lan truyền biểu thức. Trong hình 2.6, ta có các câu

```

0 esp0 := esp          ; Save esp; see text
80483b0 1 esp := esp - 4
        2 m[esp] := ebp    ; push ebp
80483b1 3 ebp := esp
80483b3 4 esp := esp - 4
        5 m[esp] := esi    ; push esi
80483b4 6 esp := esp - 4
        7 m[esp] := ebx    ; push ebx
80483b5 8 esp := esp - 4
        9 m[esp] := ecx    ; push ecx
80483b6 10 tmp1 := esp
        11 esp := esp - 8   ; sub esp, 8
80483b9 13 edx := m[ebp+8]   ; Load n to edx
  
```

Hình 2.6: Một đoạn mã trước khi thực hiện lan truyền biểu thức

lệnh ở dạng mã trung gian trước khi thực hiện lan truyền biểu thức. Hình 2.7 là kết quả sau khi thực hiện lan truyền biểu thức. Ở đây ta giả sử có một biến đặc biệt là


```

0 esp0 := esp
80483b0 1 esp := esp0 - 4
2 m[esp0-4] := ebp
80483b1 3 ebp := esp0-4
80483b3 4 esp := esp0 - 8
5 m[esp0-8] := esi
80483b4 6 esp := esp0 - 12
7 m[esp0-12] := ebx
80483b5 8 esp := esp0 - 16
9 m[esp0-16] := ecx
80483b6 10 tmp1 := esp0 - 16
11 esp := esp0 - 24
80483b9 13 edx := m[esp0+4]

```

Hình 2.7: Đoạn mã ở hình 2.6 sau khi thực hiện lan truyền biểu thức

$esp0$ được gán giá trị là giá trị ban đầu của biến esp . Ta sẽ thực hiện một thay thế đặc biệt ở câu lệnh số 1, thay vế phải của câu lệnh gán này - esp - bằng biến tương đương với nó là $esp0$. Sau đó, ở các câu lệnh tiếp theo, ta sẽ tiếp tục thay thế biến esp bằng các biểu thức tương đương. Ví dụ: Ở câu lệnh số 2, biểu thức tương đương của esp là $esp0 - 4$, còn ở câu lệnh số 5, biểu thức tương đương của esp là $esp0 - 8$ (do esp đã được gán một giá trị mới ở câu lệnh số 4). Tuy nhiên, biểu thức $esp0 - 8$ không thể được dùng để thay thế cho biến esp ở câu lệnh số 7 được, vì lúc đó esp đã mang giá trị khác.

Như vậy, qua ví dụ trên, ta có thể thấy việc lan truyền biểu thức từ câu lệnh a có dạng $x := exp$ đến một câu lệnh b chỉ có thể được thực hiện nếu đáp ứng hai điều kiện sau:

- a phải là câu lệnh gán có vế trái là x ở gần b nhất. Nói cách khác, giữa a và b không được có bất cứ câu lệnh gán nào khác có vế trái là x
- Trên tất cả các luồng đi của chương trình từ a tới b , không có câu lệnh gán nào có vế trái là bất kỳ biến nào được sử dụng trong a

```

1      a := b + 10;
2      m[b] := c * 20;
3      c := m[a] - 4;
4      d := m[b] + 3;

```

Hình 2.8: Đoạn mã trung gian với bốn câu lệnh gán đơn giản

Ở đoạn code hình 2.8, ta có thể thực hiện lan truyền biểu thức với biến a ở câu lệnh số 3, kết quả sẽ là $c := m[b + 10] - 4$; và câu lệnh gán biến a sẽ được loại bỏ bằng kỹ thuật loại bỏ mã chết được bàn ở phần tiếp theo. Tuy nhiên, ta không thể thực hiện lan truyền biểu thức với $m[b]$ ở câu lệnh số 4, vì biến c được sử dụng trong câu lệnh gán số 2 đã được sử dụng làm vế trái trong câu lệnh gán số 3.

Với mã trung gian như trên, để kiểm tra hai điều kiện thỏa mãn việc lan truyền biểu thức phải mất rất nhiều thời gian, ta phải xét hết tất cả các luồng chương trình

từ câu lệnh gán đến câu lệnh sử dụng biến, kiểm tra tất cả các biến được sử dụng trong câu lệnh gán. Tuy nhiên, với mã SSA - sẽ được nói đến ở mục 2.3.3, hai điều kiện trên sẽ được tự động thỏa mãn và không cần bất kỳ kiểm tra gì thêm.

2.3.2 Loại bỏ mã chết

Mã chết bao gồm các câu lệnh gán mà biến ở vế trái của nó không bao giờ được dùng. Cần phân biệt mã chết (dead code) với mã không bao giờ được chạy (unreachable code), là những câu lệnh mà không có bất kỳ luồng điều khiển hợp lệ nào của chương trình đi qua (ví dụ: câu lệnh ở dưới một vòng lặp vô hạn). Việc lan truyền biểu thức sẽ dẫn đến việc có một số biến không được sử dụng, từ đó sinh ra mã chết.

Từ đoạn mã đã được lan truyền biểu thức ở hình 2.7, ta thấy biến *esp* không được sử dụng ở bất kỳ câu lệnh nào. Vì vậy, các câu lệnh gán có *esp* ở vế trái sẽ được xem là mã chết và được loại bỏ. Kết quả là hình 2.9.

```
80483b0 2 m[esp0-4] := ebp
80483b3 5 m[esp0-8] := esi
80483b4 7 m[esp0-12] := ebx
80483b5 9 m[esp0-16] := ecx
80483b9 13 edx := m[esp0+4]
```

Hình 2.9: Đoạn mã ở hình 2.8 sau khi loại bỏ mã chết

Để kiểm tra xem một biến có được sử dụng hay không, ta phải xem xét tất cả các luồng chạy hợp lệ của chương trình từ câu lệnh gán biến đến cuối chương trình, điều này phức tạp và mất nhiều thời gian. Tuy nhiên, mã SSA sẽ giúp việc kiểm tra mã chết dễ dàng hơn.

2.3.3 Mã SSA

Mã SSA là một dạng mã trung gian có tính chất là: Mỗi biến hoặc vùng nhớ được định nghĩa duy nhất một lần trong toàn bộ chương trình.

Để chuyển từ mã RTL sang mã SSA, các biến cần phải được thay đổi tên, thường là sẽ được đánh số thứ tự đằng sau tên biến gốc. Ví dụ, nếu biến *a* xuất hiện ở vế trái của 3 câu lệnh gán, thì sẽ được đánh số lần lượt là *a1*, *a2* và *a3* như ví dụ ở hình 2.11

```
1      a := b + c;
2      m := a;
3      a := e + f;
4      n := a;
5      a := 10;
6      k := a;
```

Hình 2.10: Đoạn mã trung gian với 3 lần định nghĩa biến *a*

Với tính chất của mã SSA, việc lan truyền biểu thức và loại bỏ mã chết sẽ được hiện thực rất dễ dàng.

```

1      a1 := b0 + c0;
2      m1 := a1;
3      a2 := e0 + f0;
4      n1 := a2;
5      a3 := 10;
6      k1 := a3;

```

Hình 2.11: Đoạn mã ở hình 2.11 đã được chuyển sang dạng mã SSA

Đối với lan truyền biểu thức, hai điều kiện đã được tự động thỏa mãn. Điều kiện đầu tiên thỏa mãn vì mỗi biến đều được định nghĩa duy nhất một lần, không có việc có nhiều định nghĩa cho cùng một tên biến (nếu ở mã gốc có việc đó, thì khi chuyển sang mã SSA, biến đó sẽ được đánh số để trở thành những biến khác nhau ở mỗi câu lệnh gán). Điều kiện thứ hai thỏa mãn vì chắc chắn từ câu lệnh gán một biến đến bất kỳ câu lệnh nào sử dụng biến đó, biến sẽ không được định nghĩa lại.

Việc loại bỏ mã chết cũng có thể thực hiện dễ dàng nhờ vào hành động thu thập thông tin về định nghĩa và sử dụng của một biến. Trong quá trình biến đổi từ mã RTL sang SSA, ta có thể xây dựng nên một bảng vị trí câu lệnh gán của một biến và vị trí các câu lệnh sử dụng biến đó. Trải qua các quá trình phân tích, nhất là lan truyền biểu thức, bảng này sẽ được cập nhật lại. Đến cuối cùng, các biến được định nghĩa nhưng không được sử dụng ở bất kỳ câu lệnh nào sẽ được xác định và loại bỏ các câu lệnh gán dư thừa đi.

2.4 Tình hình phát triển trình dịch ngược hiện nay

Hiện nay, có rất nhiều trình dịch ngược đã và đang được phát triển. Hầu hết đều hỗ trợ việc dịch ngược từ mã máy và có thể chia làm hai loại: trình dịch ngược nhận đầu vào là mã máy chạy trên máy ảo (ví dụ: mã máy được dịch từ các chương trình viết bằng Java, C#) và trình dịch ngược nhận đầu vào là mã máy chạy trên máy thật. Loại thứ nhất có số lượng nhiều hơn, lý do có thể vì mã máy ảo còn lưu giữ được khá nhiều thông tin từ chương trình gốc, điển hình như tên biến toàn cục (xem hình 2.12). Vì vậy, bài toán cần giải quyết để xây dựng một trình dịch ngược dạng này không nhiều. Ngược lại, mã máy thật đã bị mất hầu hết các thông tin từ chương trình gốc, nên việc khôi phục thông tin ở các trình dịch ngược từ mã máy thật khá phức tạp. Trong hình 2.13, tên biến ở chương trình gốc đã bị mất đi. Ngoài ra, cấu trúc vòng lặp cũng thay đổi từ *for* sang *while*. Đó là do ở mã máy, cấu trúc vòng lặp ở ngôn ngữ cấp cao đã được dịch thành các câu lệnh kiểm tra điều kiện và jump, nên Boomerang phải sử dụng các thuật toán phân tích luồng điều khiển để tạo ra cấu trúc vòng lặp mới, đôi khi có thể không trùng khớp với cấu trúc vòng lặp ban đầu.

Một số trình dịch ngược phổ biến có thể kể đến là:

- dec: Là một trình dịch ngược từ mã máy, đây được xem là một trong những trình dịch ngược đầu tiên và vẫn còn được phát triển tới bây giờ.
- ILSpy: Là một trình dịch ngược cho .NET, input là các file assembly được

```

using System;

namespace TestILSpy
{
    internal class Program
    {
        public static string abc = "Hello, World";

        private static void Main(string[] args)
        {
            Console.Write(Program.abc + 9);
            Console.ReadLine();
        }
    }
}

```

Hình 2.12: Một đoạn mã được dịch ngược bởi trình dịch ngược ILSpy. Tên biến static *abc* được giữ nguyên như mã gốc

```

#include<stdio.h>

main()
{
    int a=0;
    for (int i=0; i<10; i++){
        a++;
    }
}

// address: 0x80483ed
int main(int argc, int argv, int envp) {
    __size32 local0; // m[esp - 12]
    int local1; // m[esp - 8]

    local0 = 0;
    local1 = 0;
    while (local1 <= 9) {
        local0++;
        local1++;
    }
    return 0;
}

```

Hình 2.13: Một đoạn mã được dịch bởi Boomerang

dịch từ chương trình .NET, được phát triển bởi isharpcode. Hiện nay ILSpy vẫn đang được tiếp tục phát triển và thêm các tính năng mới.

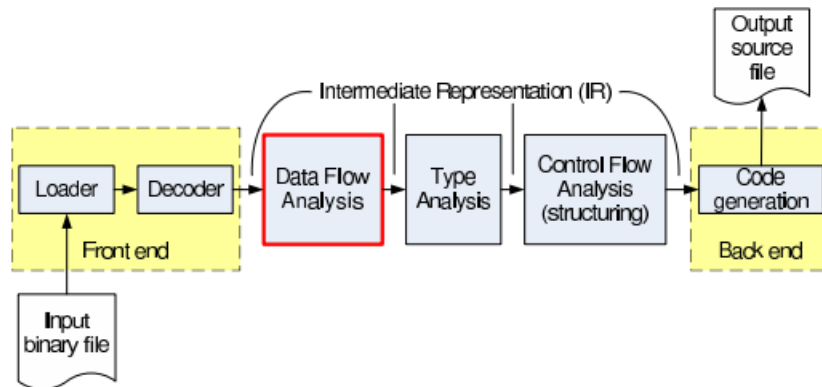
- Procyon: Là một trình dịch ngược cho Java. Trước đây lựa chọn hàng đầu để dịch ngược mã Java là JAD (Java decompiler), tuy nhiên hiện nay JAD đã ngừng phát triển và mã nguồn không còn mở nữa. Một số trình dịch ngược khác được phát triển và Procyon là một đại diện tiêu biểu.
- Boomerang: Là trình dịch ngược từ mã máy, mục tiêu là tạo ra một trình dịch ngược không quan tâm tới ngôn ngữ viết ra chương trình gốc. Boomerang đã ngừng phát triển từ năm 2006 do hai lập trình viên chính bắt đầu làm việc cho một công ty mà lĩnh vực nghiên cứu của họ trùng lặp với Boomerang.

2.5 Cấu trúc của Boomerang

Ở giai đoạn thực tập tốt nghiệp, sau quá trình nghiên cứu một số trình dịch ngược hiện nay, người viết đã quyết định sử dụng trình dịch ngược Boomerang vì sử dụng nhiều kỹ thuật phân tích giúp chất lượng mã đầu ra cao và có cấu tạo kiểu module

dễ thay đổi, thêm mới. Phần này sẽ giới thiệu về cấu trúc code của Boomerang, giúp ích cho việc trình bày các giải pháp của bài toán ở chương kế.

Về mặt tổng thể, Boomerang gồm có các phần sau:



Hình 2.14: Cấu trúc các khối lớn của Boomerang

Khi đọc vào một chương trình assembly, Boomerang sẽ lưu trữ chúng dưới cấu trúc sau:

Lưu ý, trong `AssemblyArgument`, giá trị thực sự của tham số được lưu vào một union có tên là `Arg`. Giá trị này có thể là một chuỗi (đối với trường hợp thanh ghi hoặc tên biến), một số nguyên, một số thực hoặc một structure đại diện cho bit (bao gồm tên thanh ghi và vị trí của bit).

Listing 2.1: Đoạn mã mô tả cách biểu diễn giá trị của tham số trong Boomerang

```

struct bits{
    char* reg;
    int pos;
}
union Arg{
    int i;
    float f;
    char* c;
    bits bit;
}

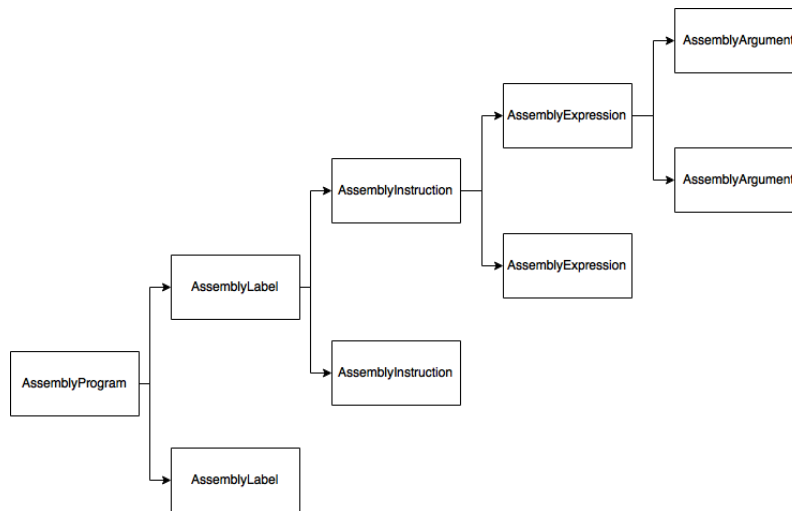
```

Khi giải mã lên ngôn ngữ trung gian, cấu trúc Boomerang dùng để thể hiện là:

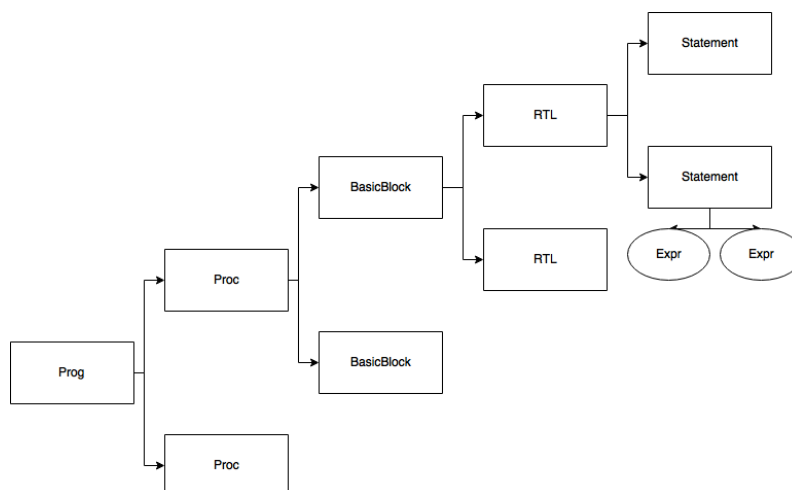
`Prog` là tương ứng với toàn bộ chương trình. Một `Proc` là một hàm, `BasicBlock` đại diện cho một khối cơ bản mà ở đó không có một câu lệnh rẽ nhánh nào (ví dụ như `if`, hoặc vòng lặp...). `Statement` là một câu lệnh và `Expr` là các biểu thức trong chương trình. Ngoài ra còn có các class đại diện cho kiểu dữ liệu. Việc thực hiện các phân tích chủ yếu diễn ra tại `Proc`, vì vậy, các thay đổi trong luận văn này cũng chủ yếu được thực hiện bằng các hàm của `Proc`.

Như vậy, để thực hiện việc xử lý mã 8051, ta cần phải chỉnh sửa các phần sau đây của Boomerang:

- Phần parser để chấp nhận câu lệnh `#DEFINE`, cũng như lưu trữ biến và giá trị của biến vào một bảng dữ liệu.



Hình 2.15: Cấu trúc dữ liệu lưu trữ mã assembly trong Boomerang



Hình 2.16: Cách lưu trữ một chương trình dưới dạng mã trung gian của Boomerang

- Phần giải mã từ mã assembly lên mã trung gian để đưa các biến không phải thanh ghi thành một biểu thức trung gian phù hợp.
- Phần chuyển đổi từ biểu thức trung gian thành tên biến để giải quyết các trường hợp biến không phải thanh ghi. Hiện tại, Boomerang chỉ cho phép người dùng định nghĩa sẵn một danh sách thanh ghi của kiến trúc máy và dựa vào đó để chuyển đổi. Cần có cơ chế để Boomerang có thể chuyển đổi các biến ngoài danh sách thanh ghi đó.
- Phần phân tích dữ liệu của Boomerang để thêm vào một phân tích hỗ trợ cho việc nhận biết các bộ biến. Mỗi giải pháp sẽ có một phương pháp phân tích khác nhau và được trình bày ở các chương kế.

Các thay đổi này sẽ được trình bày kỹ hơn ở mục ??.

Chương 3

Kiểm tra kiểu - Type checking

Giải pháp Kiểm tra kiểu gồm có các bước sau đây:

- Rút trích thông tin từ phần chú thích trong đoạn mã chương trình đầu vào để biết được các biến byte - biến bit nào thuộc cùng một bộ.
- Thực hiện phân tích dữ liệu để biết được giá trị của thanh ghi ACC tại một thời điểm của chương trình.
- Tại các câu lệnh sử dụng biến bit, tiến hành các bước kiểm tra nguyên tắc sử dụng bộ biến.

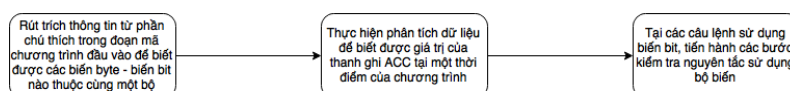
Các phần tiếp theo của chương sẽ trình bày lần lượt các bước này.

3.1 Rút trích thông tin từ phần chú thích

Khi lập trình bằng ngôn ngữ assembly và có sử dụng tên biến, các lập trình viên thường có phần chú thích bên cạnh để biết được các biến này dùng cho mục đích gì. Đối với các tên biến bit, người lập trình còn phải chỉ rõ ra họ sẽ dùng biến bit này với biến byte nào. Như vậy, nếu đọc các phần chú thích này, những người lập trình khác sẽ biết được các biến byte - biến bit nào là cùng một bộ với nhau và sử dụng đúng theo nguyên tắc. Tuy nhiên, để làm cho trình dịch ngược cũng đọc được thông tin này, cần phải có một quy định chung về việc chú thích cho các tên biến. Mẫu khai báo này được trình bày ở đoạn mã 3.1. Vì các thay đổi nằm ở phần chú thích, nên sẽ không ảnh hưởng tới nội dung chính của chương trình đầu vào.

Listing 3.1: Mẫu khai báo bộ biến

```
;BEGIN DEFINE  
;BYTE VAR  
[byte var declare]  
;BIT VAR
```



Hình 3.1: Sơ đồ các bước thực hiện giải pháp Kiểm tra kiểu


```
[eight bit var declares]
;END DEFINE
```

Như vậy, cần chỉnh lại phần parser của trình dịch ngược để chấp nhận cấu trúc mới này. Ngoài ra, trong trình dịch ngược, cần có một cấu trúc mới để lưu trữ thông tin của các bộ biến byte và biến bit này. Trong đoạn mã 3.2, có một trường `byteVar` để lưu tên biến byte, và một sơ đồ lưu thứ tự của biến bit và tên biến bit tương ứng. Hiện tại, cấu trúc này đã đủ cho giải pháp Kiểm tra kiểu. Tuy nhiên, nó sẽ được mở rộng ở giải pháp sau.

Listing 3.2: Cấu trúc dữ liệu để lưu trữ một bộ biến

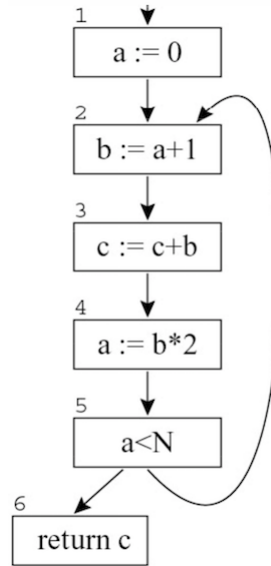
```
class UnionDefine{
    public:
        char* byteVar;
        map<int, char*>* bitVar;
    }
};
```

Như vậy, ta cần phải điều chỉnh lại phần parser của trình dịch ngược, sao cho nó vừa chấp nhận mẫu khai báo biến trên, vừa đưa từng bộ biến được khai báo vào trong một thực thể của class *UnionDefine*. Kết quả là, sau khi kết thúc phần parser, ta sẽ thu được một danh sách các *UnionDefine* tương ứng với các khai báo các bộ biến. Đoạn mã 3.3 là phần parser viết dựa trên thư viện Bison++ để thực hiện chức năng trên.

Listing 3.3: Đoạn mã parser cho phần khai báo bộ biến

```
definebit: BEGINDEFINE END_LINE DEFINEBYTE END_LINE
define DEFINEBITS END_LINE defineeachbit defineeachbit
defineeachbit defineeachbit defineeachbit defineeachbit
defineeachbit defineeachbit ENDDEFINE END_LINE
{
    UnionDefine* ut = new UnionDefine();
    $5 -> expList -> pop_back();
    ut->byteVar = $5 -> expList -> back() -> argList.back()->value.c;
    ut->bitVar = bitVar;
    unionDefine1 -> push_back(ut);
    bitVar = new map<char*, int>();
}
;
defineeachbit: DEFINE ID bit END_LINE {
    std::string temp($3->value.c);
    char c = temp.at(temp.size()-1);
    int num = c - '0';
    (*bitVar)[$2] = num;
}
;
```

Như vậy, sau giai đoạn parser, ta đã thu được bộ biến theo khai báo của người dùng. Tuy nhiên, nguyên tắc sử dụng bộ biến này là không bắt buộc trong quá trình lập



Hình 3.2: Một đoạn chương trình mẫu

trình 8051, và các assembler cũng không hề kiểm tra việc người dùng có tuân thủ nguyên tắc này không. Vì vậy, trước khi chuyển hoá các bộ biến này sang cấu trúc union ở mã đầu ra, ta cần phải kiểm tra đoạn mã của người dùng sử dụng các bộ biến như thế nào.

3.2 Kiểm tra nguyên tắc sử dụng bộ biến

3.2.1 Xác định giá trị kiểu của thanh ghi *ACC* tại mỗi điểm của chương trình

Để kiểm tra được nguyên tắc sử dụng bộ biến, trước hết, ta phải biết được tại mỗi thời điểm của chương trình, thanh ghi *ACC* đang mang giá trị gì. Có hai phương pháp phân tích được áp dụng trong giải pháp này, đó là Reaching definitions kết hợp phần mở rộng và Type propagation.

Phân tích Reaching definitions kết hợp phần mở rộng

Mục đích của phân tích Reaching definitions là biết được ở một thời điểm của chương trình, các câu lệnh khai báo nào đang còn hiệu lực, hay nói cách khác là giá trị của các biến đang được khai báo bởi những câu lệnh nào. Phần mở rộng của Reaching definition sẽ giúp xử lý thêm một số trường hợp, khi thanh ghi được truyền giá trị thông qua một thanh ghi trung gian khác. Ví dụ như ở đoạn chương trình 3.2, ta cần biết được giá trị của biến *a* ở câu lệnh số 2 được khai báo ở câu lệnh nào. Đối với con người thì rất dễ dàng biết được là biến *a* được sử dụng có thể khai báo ở câu lệnh số 1 hoặc câu lệnh số 5. Tuy nhiên, cần có một phương pháp phân tích để cho máy tính cũng biết được điều đó, và đó chính là phương pháp Reaching definitions. Như vậy, khi áp dụng vào trình dịch ngược, ta sẽ biết được tại thời điểm sử dụng biến bit, giá trị của thanh ghi *ACC* đang được định nghĩa ở câu lệnh nào. Từ đó tiến hành các bước kiểm tra tiếp theo.

Để thực hiện Reaching definitions, ta phải làm quen với các định nghĩa sau:

- Nếu một biến được khai báo ở câu lệnh *def1*, sau đó được khai báo lại ở câu lệnh *def2* sau đó, thì ta nói là câu lệnh *def1* đã **bị giết (killed)** bởi câu lệnh *def2*.
- Nếu có một đường thực thi chương trình đi từ câu lệnh khai báo *def1* đến một điểm *p* của chương trình, mà trên đó *def1* không bị giết bởi bất kỳ câu lệnh nào, thì ta nói là *def1* đã **đến được (reach)** điểm *p*. Khái niệm một câu lệnh đến được một khối cơ bản cũng tương tự như vậy.

Ngoài ra, ta phải quy định một số khái niệm mới cho một khối cơ bản *B*:

- *REACHin(B)*: Tập hợp các câu lệnh khai báo đến được đầu vào (entry) của *B*.
- *REACHout(B)*: Tập hợp các câu lệnh khai báo đến được ngõ ra (exit) của *B*.
- *GEN(B)*: Tập hợp các câu lệnh khai báo xuất hiện trong *B* và có thể đến được ngõ ra (exit) của *B*, nghĩa là biến được khai báo trong câu lệnh đó không được khai báo lại ở các câu lệnh đằng sau nó.
- *KILL(B)*: Tập hợp các câu lệnh khai báo mà biến được khai báo đã được khai báo lại trong *B*.

Như vậy, mục tiêu của phân tích Reaching definitions ở cấp độ khối cơ bản là tìm ra được tập hợp *REACHin* và *REACHout* của từng khối. Công thức được áp dụng là;

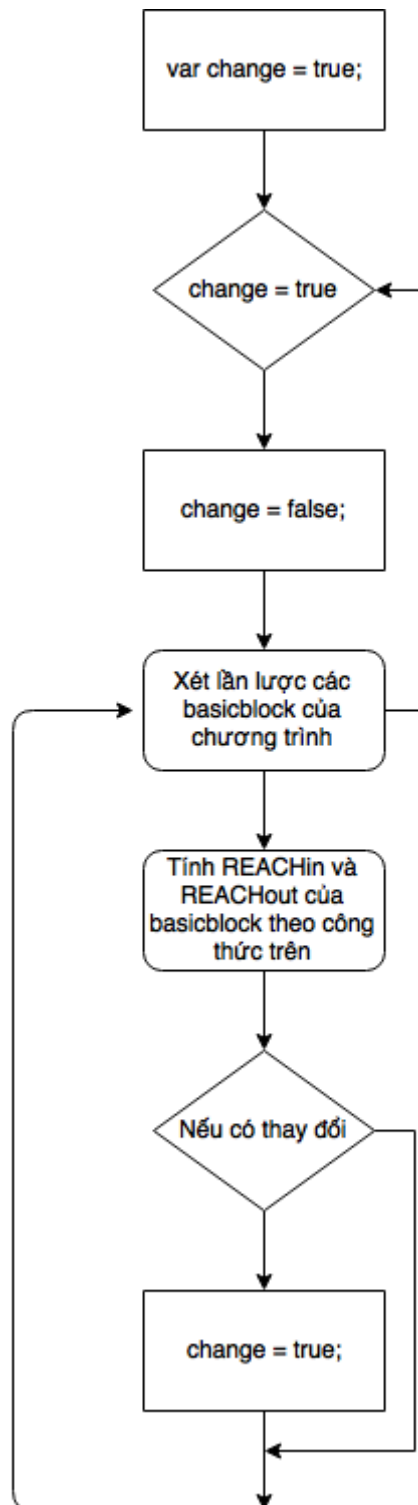
$$REACHout(B) = GEN(B) \cup (REACHin(B) - KILL(B)) \quad (3.1)$$

$$REACHin(B) = \cup_{j \in Pred(B)} REACHout(j) \quad (3.2)$$

Từ hai công thức 3.1 và 3.2, ta thấy ở phân tích này, tập hợp các giá trị ra (*REACHin*) được quyết định bởi các giá trị vào (*REACHout*), ngược lại với phân tích liveness (tìm tập hợp biến đang sống tại một thời điểm của chương trình). Như vậy, luồng đi của phân tích là cùng chiều với luồng đi của chương trình. Tương tự một số phương pháp phân tích dữ liệu khác, ta sẽ lần lượt tính toán các tập hợp vào và tập hợp ra của từng khối cơ bản cho đến khi không còn thay đổi nào được ghi nhận. Xem sơ đồ khối hình 3.3.

Tuy nhiên, trong trường hợp của bài toán cần giải quyết, ta cần phải biết tập hợp ra vào của từng câu lệnh một, chứ không chỉ của toàn bộ khối cơ bản, vì vậy, khi ứng dụng vào Boomerang, giải thuật sẽ được điều chỉnh lại để tìm tập *REACHin* và *REACHout* của từng câu lệnh. Việc điều chỉnh này là khá nhỏ, và các bước vẫn sẽ giữ nguyên, không thay đổi nhiều.

Khuyết điểm lớn nhất của phân tích Reaching definitions chỉ cho biết được câu lệnh khai báo có hiệu lực của một biến tại một thời điểm chương trình, chứ không cho biết được giá trị thực sự của biến đó. Cụ thể, với thanh ghi *ACC*, nếu về phải của câu lệnh khai báo này chỉ đơn giản là trở đến một vùng nhớ có địa chỉ được quy định bởi một biến byte thì biến byte đó sẽ được ghi nhận là đại diện cho vùng nhớ hiện thời *ACC* đang lưu trữ. Nhưng ngoài ra, biểu thức quy định địa chỉ vùng nhớ được gán cho *ACC* có thể là các trường hợp sau đây:



Hình 3.3: Giải thuật tính Reaching definitions cho khối cơ bản

- Một hằng số.
- Một thanh ghi, giá trị của thanh ghi có thể được khai báo ở các câu lệnh trước đó.
- Một biểu thức có hai vế, các vế của biểu thức có thể là một biến, một thanh ghi hoặc một hằng số.

Listing 3.4: Một số câu lệnh gán mà phương pháp Suy luận kiểu sử dụng Reaching definitions không xử lý được

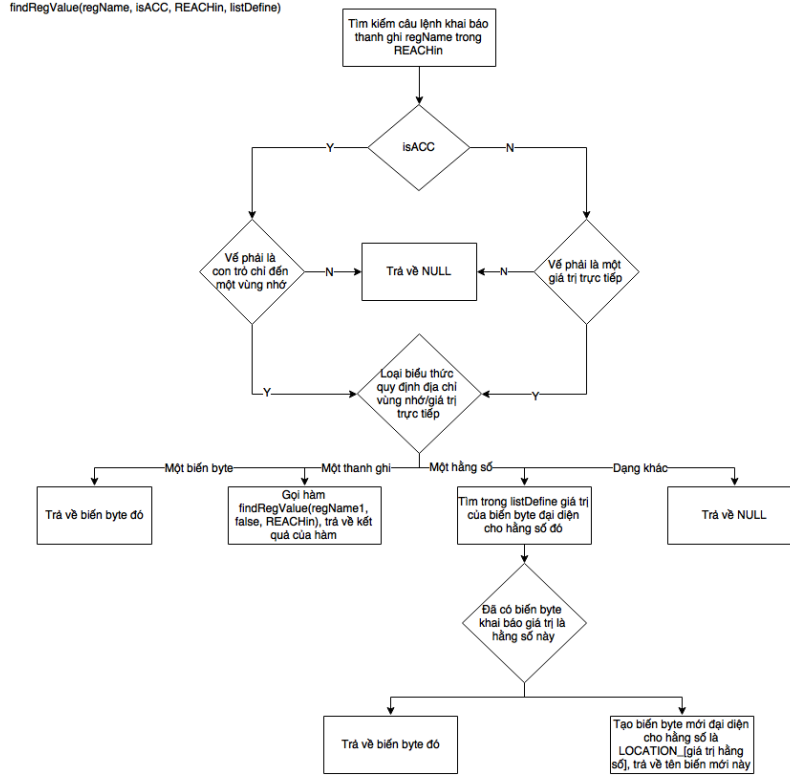
```
MOV A, 38H #1
MOV A, @DPTR #2
MOV A, OPTION+1 #2
```

Các trường hợp phức tạp nêu trên được trình bày trong đoạn mã 3.4. Phương pháp Reaching definitions sẽ không thể xử lý được khi gặp các câu lệnh gán này. Ở câu lệnh số 1, phân tích có thể lấy được giá trị **38H**, nhưng không thể xác định được biến byte nào đã được khai báo giá trị **38H**. Ở câu lệnh số 2, Reaching definitions không thể biết được giá trị của thanh ghi *DPTR* là bao nhiêu. Ở câu lệnh số 3, việc xử lý lại càng phức tạp hơn, vì nếu chỉ đơn giản lấy vế phải của khai báo ra, không thể nào biết được giá trị thực sự của nó là bao nhiêu. Ngoài ra, với trường hợp trong tập hợp *REACHin* của câu lệnh có nhiều câu lệnh khai báo cho thanh ghi *ACC*, phương pháp này sẽ không thể kiểm tra được vế phải của tất cả các câu lệnh khai báo đó có cùng một giá trị hay không mà chỉ đơn giản xử lý là câu lệnh đã vi phạm nguyên tắc sử dụng bộ biến. Ví dụ như đoạn mã 3.5, tại thời điểm câu lệnh sử dụng biến bit *TESTSUPS*, có hai câu lệnh khai báo biến *a* (là biến đại diện cho thanh ghi *ACC* tại ngôn ngữ trung gian). Đối với phương pháp Reaching definitions, nó sẽ xem như có hai giá trị mà biến *a* có thể mang, và sẽ báo lỗi vì vi phạm nguyên tắc sử dụng bộ biến. Tuy nhiên, nếu xét kỹ hơn, thì sẽ thấy là cả hai giá trị đó đều là biến *OPTIONS*, và thực chất tại thời điểm này *a* chỉ mang một giá trị, cho dù luồng đi của chương trình có như thế nào. Như vậy, phương pháp Reaching definitions sẽ bỏ qua những trường hợp như thế này và báo lỗi, dẫn đến việc độ chính xác sẽ không được cao.

Listing 3.5: Đoạn mã có nhiều câu lệnh khai báo cho *ACC* đến được một điểm của chương trình nhưng tất cả đều cùng giá trị

```
if (...) {
    a = *(OPTIONS);
    ...
} else {
    a = *(OPTIONS);
    ...
}
TESTSUPS = 1;
```

Để khắc phục phần nào khuyết điểm của Reaching definitions, một phần mở rộng được thêm vào để giải quyết một trong những trường hợp mà phương pháp này không thể giải quyết được. Đó là trường hợp sử dụng một biến trung gian và sử dụng hằng (câu lệnh số 1 và số 2 ở đoạn mã 3.4). Các trường hợp này được giải



Hình 3.4: Giải thuật cho hàm findRegValue - phần mở rộng của phân tích Reaching definitions

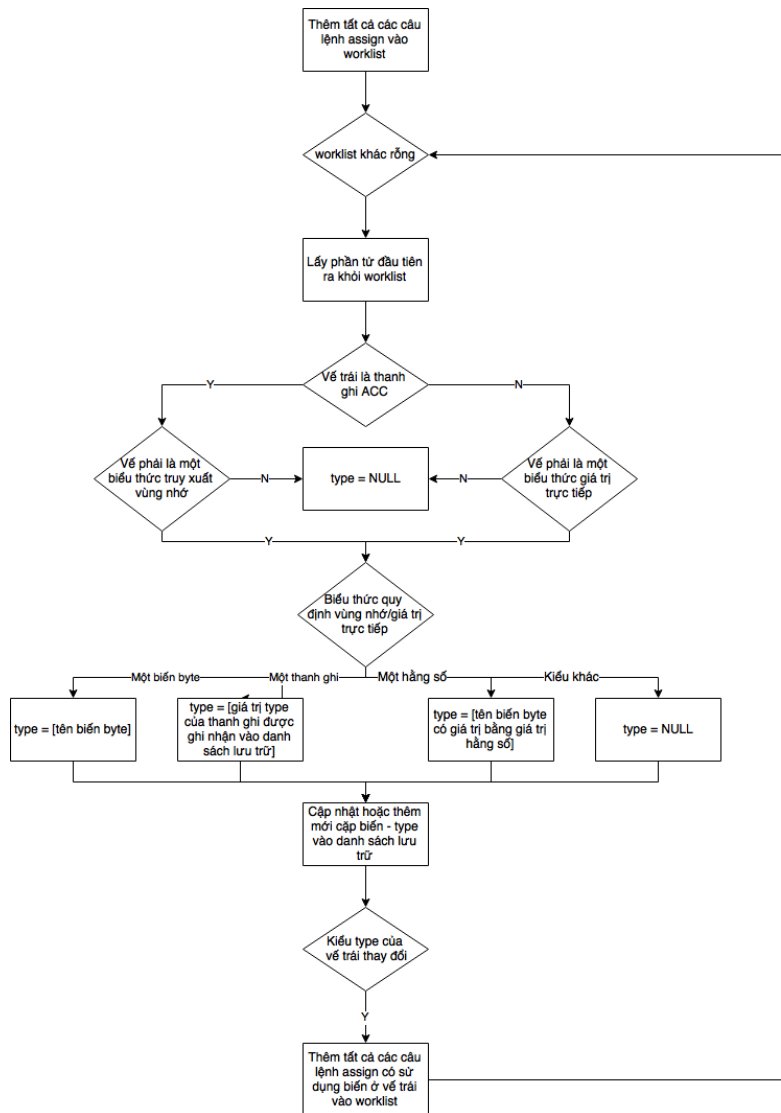
quyết nhờ vào việc lưu trữ các khai báo *#DEFINE* ở đoạn mã đầu vào, cũng như trong quá trình phân tích, tất cả các câu lệnh khai báo của tất cả các biến đến được một điểm trong chương trình đều được lưu giữ chứ không chỉ riêng của thanh ghi *ACC*. Cụ thể các bước của phần mở rộng được trình bày ở hình 3.4.

Theo giải thuật nêu trên, để tìm được giá trị của thanh ghi *ACC* tại một thời điểm của chương trình, cần truyền vào hàm *findRegValue* tên thanh ghi ("*ACC*"), giá trị true để hàm biết đang tìm kiếm giá trị cho thanh ghi *ACC*, tập *REACHin* tại thời điểm đó của chương trình và danh sách các cặp tên biến - giá trị biến được khai báo ở phần đầu của mã đầu vào. Qua các bước của giải thuật, hàm sẽ trả về giá trị là biến byte quy định vùng nhớ được load dữ liệu vào *ACC* nếu tìm được biến này, còn nếu trả về giá trị **NULL**, nghĩa là giải thuật đã gặp phải trường hợp không xử lý được, đó là trường hợp biểu thức vế phải gồm hai toán hạng (câu lệnh số 3 của đoạn mã 3.4).

Tuy đã xử lý được hầu hết các trường hợp của phép gán cho *ACC*, nhưng việc có thêm một phần mở rộng này sẽ làm cho tốc độ xử lý trình dịch ngược giảm đi. Cộng thêm việc bản thân giải thuật Reaching definitions đã có độ phức tạp cao, tổng thời gian xử lý cho bước này của trình dịch ngược là khá lớn. Vì vậy, một phương pháp phân tích khác đã được thực hiện để tìm ra giá trị của thanh ghi *ACC* với độ phức tạp thấp hơn.

Phân tích Type propagation

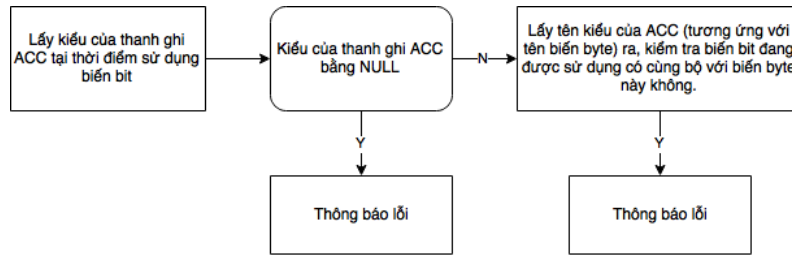
Phân tích Type propagation có mục đích sẽ tìm ra được kiểu của một thanh ghi bằng cách lan truyền kiểu của những thanh ghi khác. Áp dụng vào trường hợp cụ



thể của luận văn, các tên biến byte có thể được xem là một kiểu, vì khi chuyển đổi bộ biến thành một kiểu union ở ngôn ngữ cấp cao, tên biến byte này sẽ được dùng để đặt tên cho kiểu union đó. Như vậy, tác dụng của phân tích Type propagation này là tìm ra được thanh ghi ACC đang mang kiểu union nào tại một thời điểm của chương trình. Các bước của giải thuật phân tích Type propagation được trình bày ở hình 3.5

Trong giải thuật đã trình bày ở trên, không xét đến trường hợp một biến được khai báo lại nhiều lần và tại mỗi thời điểm sẽ có giá trị khác nhau là do tận dụng được mã trung gian SSA. Hầu hết các trình dịch ngược đều sử dụng mã SSA tại một giai đoạn nào đó trong quá trình phân tích dữ liệu vì các lợi ích của loại mã này (chi tiết về mã SSA được trình bày trong phần 2.3.3). Với mã SSA, mỗi biến chỉ được khai báo một lần duy nhất trong chương trình, nên không bị ảnh hưởng bởi trường hợp khai báo lại biến.

Như vậy, phương pháp Type propagation này sẽ làm giảm đáng kể thời gian xử lý của chương trình so với phương pháp Reaching definitions. Khi có một thay đổi



Hình 3.6: Quá trình kiểm tra một câu lệnh sử dụng bit

nào đó, chương trình phân tích chỉ cần tính toán lại các câu lệnh khai báo chịu ảnh hưởng của sự thay đổi đó, chứ không cần phải tính toán lại hết tất cả các câu lệnh như giải thuật của Reaching definitions. Phương pháp này cũng không cần một phần mở rộng mà ngay trong quá trình lan truyền, nó đã tính toán được các trường hợp biểu thức về phải của khai báo là một thanh ghi hoặc một hằng số. Khuyết điểm của Type propagation là vẫn chưa xử lý được trường hợp biểu thức về phải có hai toán hạng. Để giải quyết được vấn đề này, cần có một phân tích khác mạnh hơn và phân tích đó sẽ được giới thiệu ở chương tiếp theo.

3.2.2 Kiểm tra tính hợp lệ về kiểu của tác vụ với bit

Sau khi đã tìm được thông tin về kiểu của thanh ghi ACC tại mỗi điểm của chương trình, bước tiếp theo của giải pháp Kiểm tra kiểu là kiểm tra tính hợp lệ khi sử dụng các biến bit. Như đã trình bày ở phần 1.2, việc sử dụng các biến byte - biến bit trong mã chương trình 8051 cần tuân thủ các quy định sau:

- Chỉ khi thanh ghi *ACC* đang mang giá trị của vùng nhớ có địa chỉ quy định bởi biến byte, thì các biến bit cùng bộ mới được sử dụng.
- Mỗi biến bit chỉ thuộc một bộ duy nhất.
- Tại mỗi vị trí bit của một bộ biến chỉ có một biến bit duy nhất tồn tại.

Như vậy, ở phần này, chương trình sẽ chạy vòng lặp kiểm tra lần lượt từng câu lệnh sử dụng bit để xem xét việc sử dụng này có vi phạm quy định nào trong những quy định trên hay không. Các bước kiểm tra này được trình bày ở hình 3.6. Nếu có một câu lệnh sử dụng bit nào đó vi phạm nguyên tắc trên, chương trình kiểm tra vẫn tiếp tục chạy để kiểm tra những câu lệnh sau đó. Điều này giúp người lập trình biết được tất cả các lỗi trong chương trình đầu vào của mình để có sự chỉnh sửa phù hợp. Tuy nhiên, khi có vi phạm xảy ra, trình dịch ngược sẽ không sinh được mã đầu ra, vì không có cách thể hiện union phù hợp.

Như vậy, với giải pháp Kiểm tra kiểu, ta đã có sẵn thông tin về bộ biến ngay từ đầu và chỉ cần kiểm tra xem người lập trình có tuân thủ đúng quy tắc không trước khi sinh ra mã ở ngôn ngữ cấp cao. Giải pháp này yêu cầu can thiệp vào trình dịch ngược ít và hiện thực dễ dàng. Tuy nhiên, giải pháp còn nhiều hạn chế như phương pháp phân tích dữ liệu chưa đạt độ chính xác cao, cần người dùng phải chỉnh sửa lại chú thích theo mẫu quy định... Chính vì vậy, giai đoạn sau của luận văn đã phát triển một hướng tiếp cận mới có độ chính xác cao hơn và không cần chỉnh sửa mã đầu vào của người dùng, đó là giải pháp Suy luận kiểu. Giải pháp này sẽ được trình bày ở chương tiếp theo.

Chương 4

Suy luận kiểu - Type inference

Giải pháp Kiểm tra kiểu bắt buộc người lập trình chương trình đầu vào phải để lại chú thích ở phần khai báo biến, tuy nhiên, trong vì một số lý do, có thể phần chú thích này sẽ không xuất hiện hoặc không thể hiện đầy đủ thông tin, vì vậy cần có một giải pháp khác để xử lý trường hợp này. Giải pháp được giới thiệu ở chương này là Suy luận kiểu, với giải pháp này, bằng các phép phân tích dữ liệu, trình dịch ngược sẽ tự động tìm ra được các bộ biến được sử dụng trong chương trình. Các bước của giải pháp Suy luận kiểu gồm có:

1. Xác định giá trị của thanh ghi ACC tại mỗi thời điểm của chương trình.
2. Thông qua quá trình sử dụng biến của chương trình, lấy ra được mối quan hệ của các biến.

4.1 Xác định giá trị của thanh ghi ACC tại mỗi thời điểm của chương trình

Tương tự như ở chương trước, bước đầu tiên của giải pháp Suy luận kiểu là xác định giá trị của thanh ghi ACC tại mỗi thời điểm của chương trình. Các giải pháp Reaching definitions mở rộng và Type propagation đã trình bày đều không thể xử lý hết tất cả các trường hợp xảy ra của phép gán thanh ghi ACC. Cụ thể, trường hợp không thể xử lý được là khi biểu thức quy định địa chỉ vùng nhớ của thanh ghi ACC là một biểu thức có hai toán hạng (xem ví dụ ở đoạn mã 4.1).

Listing 4.1: Trường hợp không thể xử lý được bằng các phương pháp phân tích dữ liệu trước

```
MOV ACC, OPTIONS+1
```

Để mở rộng khả năng xử lý của trình dịch ngược, cần phải tìm ra một phương pháp khác tốt hơn. Phương pháp đạt yêu cầu phải tính toán được chính xác giá trị hiện có của thanh ghi ACC cho dù biểu thức bên phải của phép gán là gì. Cụ thể, chỉ các trường hợp giá trị ở thanh ghi ACC là một giá trị cố định, có thể tính toán được trước khi thực thi chương trình mới được xét đến vì nếu thanh ghi ACC có thể mang nhiều giá trị vùng nhớ khác nhau thì nguyên tắc sử dụng bộ biến sẽ bị vi phạm. Từ các yêu cầu trên, phương pháp phân tích phù hợp nhất trong trường hợp này là Lan truyền hằng số - Constant propagation. Phương pháp này cho phép tính toán giá trị của các biến, cho biết được giá trị đó có phải là một hằng số tại một

thời điểm của chương trình hay không. Ví dụ như đoạn mã ban đầu 4.2, chương trình phân tích có thể thấy giá trị của biến x là 14, nhưng không biết được giá trị thực sự của biến y , cũng như biểu thức trả về là bao nhiêu. Nhờ vào việc lan truyền hằng số, các giá trị này sẽ được tính toán, như trong đoạn mã 4.3 và 4.4.

Listing 4.2: Đoạn mã trước khi thực hiện lan truyền hằng số

```
int x = 14;
int y = 7 - x / 2;
return y * (28 / x + 2);
```

Listing 4.3: Đoạn mã sau khi thực hiện lan truyền hằng số cho biến y

```
int x = 14;
int y = 0;
return y * (28 / x + 2);
```

Listing 4.4: Đoạn mã sau khi thực hiện lan truyền hằng số cho biểu thức trả về

```
int x = 14;
int y = 0;
return 0;
```

Với phương pháp này, một biến có thể có ba giá trị sau:

- Top: Nghĩa là chưa biết được biến có giá trị gì.
- Hằng số: Nghĩa là đã xác định được giá trị của biến là một hằng số.
- Bottom: Nghĩa là biến có thể mang những giá trị khác nhau, tùy thuộc vào luồng chạy của chương trình.

Ở bước khai báo ban đầu của giải thuật, tất cả các biến đều được truyền vào giá trị top (chưa biết), sau đó, trải qua quá trình phân tích thì giá trị của một biến có thể được xác định là hằng số hoặc bottom. Trong ví dụ 4.5, dễ dàng thấy được ở câu lệnh số 2, biến a được gán giá trị là hằng số 4. Và vì không có câu lệnh khai báo biến a nào xuất hiện ở giữa, nên ở câu lệnh số 3, giá trị của biến a cũng vẫn là hằng số 4. Còn ở ví dụ 4.7, giá trị của biến b được người dùng nhập vào, nên không thể biết trước được giá trị chính xác của nó là gì. Vì vậy, cũng không thể xác định chính xác luồng đi của chương trình như thế nào. Trong quá trình thực thi, chương trình có thể đi theo nhánh #1, cũng có thể đi theo nhánh #2 tùy thuộc vào người dùng nhập gì cho biến b . Kết quả cuối cùng là biến a có thể mang những giá trị khác nhau ở câu lệnh dòng thứ 9. Hay nếu dựa vào định nghĩa 3 loại giá trị của biến đã nêu trên, giá trị của a tại câu lệnh return này là bottom.

Listing 4.5: Đoạn mã ví dụ biến có giá trị là hằng số

```
int a;
a = 4;
b = a*4;
```

Listing 4.6: Đoạn mã ví dụ biến có giá trị là bottom

```
int a;
int b;
```

```

cout<<"Enter_b:_";
cin >> b;
if (b>15) #1
    a = 4;
else #2
    a = 5;
return a;

```

Như vậy, khi áp dụng vào trình dịch ngược Boomerang, mục tiêu của giải thuật này là để tìm ra được ở mỗi điểm của chương trình, thanh ghi ACC có mang giá trị của một vùng nhớ duy nhất hay không, và nếu có thì giá trị thật sự của địa chỉ vùng nhớ đó là gì.

Có nhiều cách thực hiện Constant propagation, và như đã trình bày ở chương trước, thường có một giai đoạn code trung gian trong các trình dịch ngược được giữ ở dạng SSA, nên cách phân tích Sparse constant propagation sẽ được lựa chọn để giảm thiểu thời gian xử lý. Và việc phân tích này sẽ được thực hiện ở cuối giai đoạn SSA, khi các phân tích khác đã hoàn tất. Giải thuật của phân tích Sparse constant propagation gồm có các bước được trình bày ở hình 4.1

Để thể hiện giá trị của một biến có thể thuộc ba loại là top, hằng số hoặc bottom, cần tạo ra một class mới để lưu trữ loại giá trị, đồng thời lưu trữ giá trị thực sự nếu biến đó là một hằng số. Class này được đặt tên là ConstantVariable và có khai báo được trình bày ở đoạn mã 4.7

Listing 4.7: Đoạn mã thể hiện class ConstantVariable

```

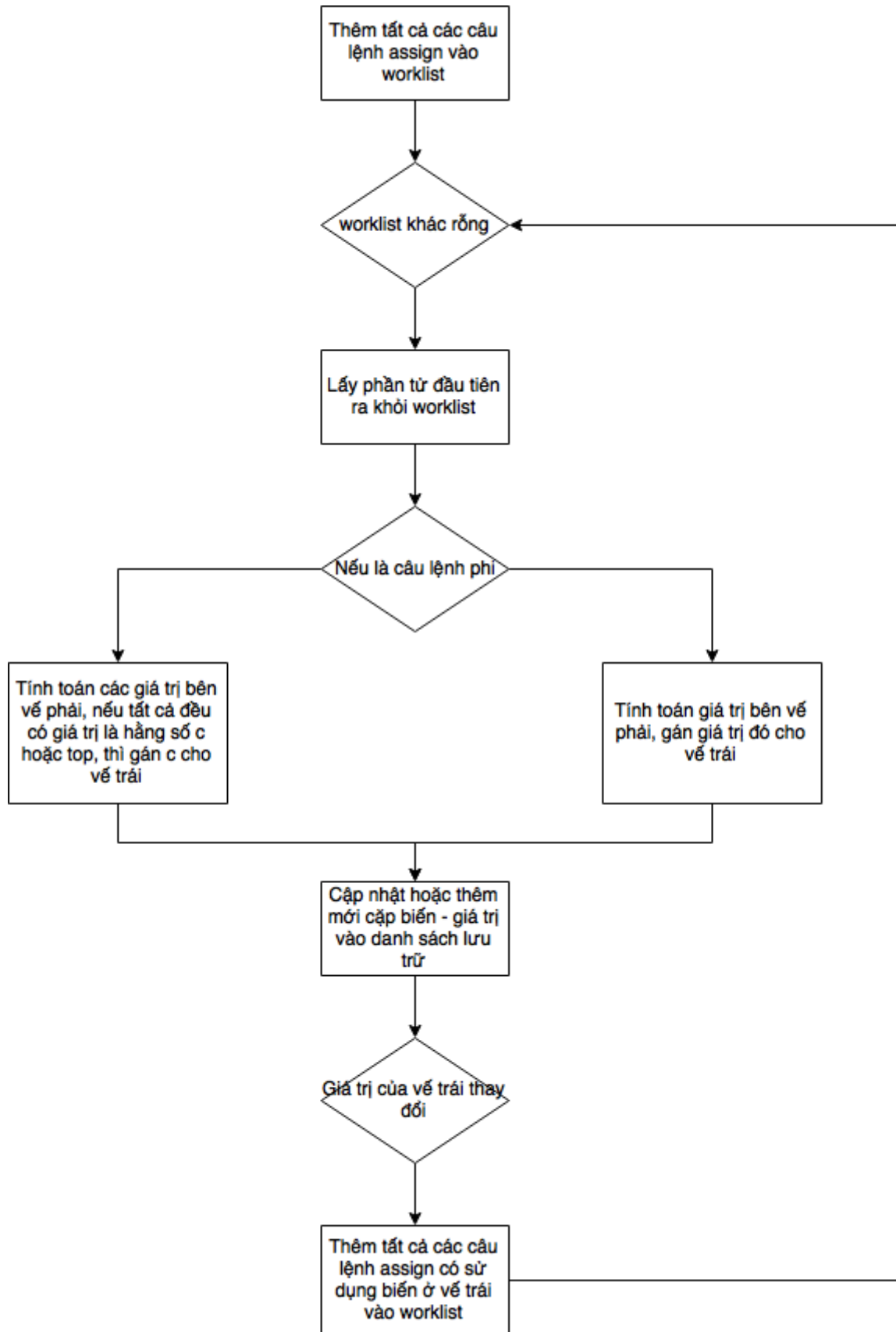
class ConstantVariable{
public:
    int type; //1: top, 2: constant, 3: bottom
    Exp* variable;
    ConstantVariable(){
        type = 3;
    }
};

```

Như vậy, mục tiêu của giải thuật này là tạo ra được một sơ đồ liên kết giữa một biến SSA và một thực thể ConstantVariable thể hiện giá trị của biến đó.

Ngoài ra, để tính toán được giá trị thực sự của một biến, cần phải có một hàm chức năng nhận vào một biểu thức và trả về được giá trị của biểu thức đó. Có rất nhiều cách để hiện thực chức năng này, sau một quá trình xem xét, visitor sẽ được sử dụng cho việc tính toán giá trị của biểu thức. Visitor là một pattern design trong các chương trình lập trình hướng đối tượng, nó được dùng để tách một thuật toán ra khỏi cấu trúc dữ liệu mà thuật toán đó sử dụng. Lợi ích đạt được là người lập trình có thể thêm những chức năng mới cho cấu trúc dữ liệu đó mà không cần thay đổi kiến trúc của nó. Điều này phù hợp với nhu cầu hạn chế tối đa việc thay đổi các mã có sẵn khi hiện thực các giải pháp của luận văn trên một trình dịch ngược nào đó.

Để hiện thực pattern design này, cần tạo ra một class visitor và viết hàm visit cho các loại biểu thức. Các hàm visit này chính là nơi tính toán giá trị của biểu thức và trả về chúng. Vì biểu thức ở mức assembly thường được viết rất đơn giản, nên trong class visitor này chỉ cần có một số hàm visit cho các loại biểu thức sao đây:



Hình 4.1: Giải thuật Constant propagation đã được điều chỉnh phù hợp với yêu cầu của trình dịch ngược

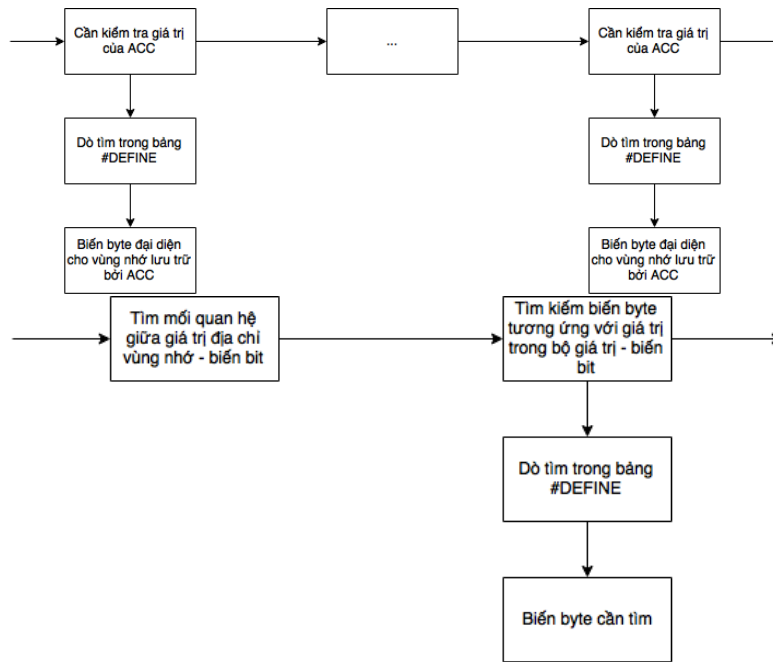
- **Const:** Là biểu thức hằng số. Hàm visit này chỉ đơn giản trả về giá trị hằng số nếu đây là một hằng số nguyên.
- **Binary:** Là biểu thức có 2 vế. Hàm visit sẽ visit từng vế của biểu thức, và nếu cả hai vế đều là hằng số, thì sẽ thực hiện phép tính cộng trừ nhân chia hai hằng số đó để ra được kết quả cuối cùng.
- **RefExp:** Loại biểu thức này chứa một biểu thức khác, kèm theo câu lệnh khai báo biểu thức đó. Trong giới hạn nhu cầu của bài toán, chỉ những RefExp chứa biểu thức là một biến hoặc một thanh ghi được tính toán tiếp, còn những loại biểu thức kia sẽ mặc định trả về giá trị là bottom. Tên biến hoặc biểu thức sẽ được dò tìm trong bảng lưu trữ dữ liệu hằng số và bảng lưu trữ dữ liệu của câu lệnh `#DEFINE` để tìm ra được giá trị thực sự của chúng và trả về.
- **TypedExp:** Loại biểu thức để ép kiểu một biểu thức nào đó thành kiểu mong muốn. Với trường hợp này, giá trị trả về của biểu thức ép kiểu chính là giá trị của biểu thức con bên trong.

Như vậy, với phương pháp phân tích này, vấn đề về phải của phép gán thanh ghi là những biểu thức phức tạp có hai toán hạng đã được giải quyết. Ngoài ra, phân tích này còn nhận biết được các biểu thức có giá trị giống nhau mặc dù hình thức bên ngoài khác nhau. Xem ví dụ các câu lệnh ở đoạn mã 4.8. Câu lệnh gán số 1 và số 2 thực chất đều gán cho ACC giá trị vùng nhớ có địa chỉ quy định bởi biến `OPTIONS`. Nếu thực hiện phân tích *Reaching definitions* ở giải pháp trước, trình dịch ngược sẽ không thể biết được điều này. Tuy nhiên, ở giai đoạn này, vì trình dịch ngược sẽ tính toán được ở cả hai câu lệnh, ACC đều mang giá trị của vùng nhớ có địa chỉ là 38. Và ở những bước tiếp theo, trình dịch ngược sẽ đối chiếu giá trị 38 với bảng lưu trữ dữ liệu và biết được biến `OPTIONS` đại diện cho giá trị đó.

Listing 4.8: Một số câu lệnh gán cho ACC có giá trị về phải bằng nhau

```
#DEFINE OPTIONS #38
#DEFINE CLAMP #37
...
MOV ACC, OPTIONS
MOV ACC, CLAMP+1
```

Phân tích Constant propagation sẽ trả về được giá trị thực sự của một biến, tuy nhiên, thanh ghi ACC là một trường hợp đặc biệt. Khi gặp vế trái của câu lệnh khai báo là thanh ghi ACC, giá trị thực sự không được quan tâm, mà chỉ giá trị của địa chỉ vùng nhớ đang được ACC lưu giữ mới được xét đến. Như vậy, chỉ có các câu lệnh dạng `MOV ACC, [biểu thức]` sẽ được tính toán còn khi gặp câu lệnh gán có dạng `MOV ACC, #[biểu thức]` thì đoạn mã phân tích sẽ xem như giá trị của ACC không phải là hằng số (bottom). Như vậy, với các biến khác, giá trị lưu trong thực thể `ConstantExpression` tương ứng với biến đó là giá trị thực sự của biến, còn riêng với thanh ghi ACC, giá trị đó được hiểu là giá trị địa chỉ vùng nhớ mà thanh ghi ACC được load vào.



Hình 4.2: Các giải pháp chuyển đổi giá trị - biến byte

4.2 Tìm kiếm mối quan hệ giữa các biến

4.2.1 Chuyển đổi giữa hằng số nguyên - biến byte tương ứng

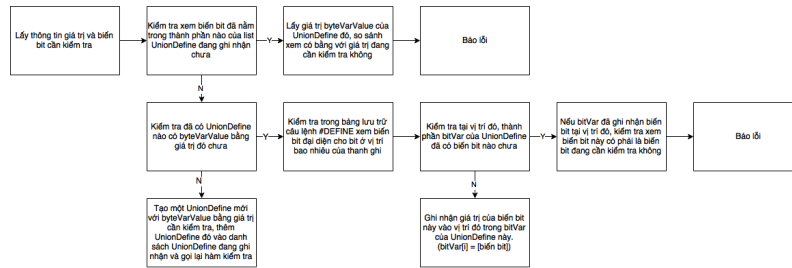
Trước khi bước vào giai đoạn kiểm tra và ghi nhận mối quan hệ giữa các biến, trình dịch ngược cần phải giải quyết kết quả trả về của phương pháp phân tích Constant propagation ở trên. Như đã trình bày, phương pháp này cho biết chính xác giá trị địa chỉ vùng nhớ được lưu trữ bởi thanh ghi ACC. Tuy nhiên, cần phải chuyển đổi con số này thành một biến byte có giá trị bằng nó vì mục đích cuối cùng vẫn là tìm mối quan hệ giữa các biến byte - biến bit. Có hai giải pháp cho vấn đề này:

- Ở bất kỳ vị trí nào cần biết được biến byte đang quy định vùng nhớ lưu bởi ACC, lấy giá trị thực sự của địa chỉ vùng nhớ đó và dò tìm trong bảng lưu trữ dữ liệu các câu lệnh `#DEFINE` để tìm ra biến byte tương ứng với giá trị đó.
- Tạm thời chấp nhận thay vì tìm kiếm mối quan hệ giữa biến byte - biến bit thì sẽ tìm kiếm mối quan hệ giữa giá trị trực tiếp - biến bit. Giá trị này chính là địa chỉ vùng nhớ mà thanh ghi ACC đang lưu trữ tại thời điểm sử dụng biết bit. Sau khi tìm ra các bộ giá trị - biến bit, thêm vào trình dịch ngược một bước chuyển đổi từ giá trị sang biến byte tương ứng dựa vào các câu lệnh khai báo giá trị biến byte ở chương trình đầu vào.

Hai giải pháp trên được mô tả ở hình 4.2. Qua hai sơ đồ trên, dễ dàng nhận ra giải pháp thứ hai sẽ giúp tiết kiệm thời gian xử lý chương trình hơn, nhờ vào việc chỉ cần truy xuất bảng dữ liệu ở cuối giai đoạn tìm kiếm. Còn giải pháp đầu tiên không hiệu quả do mỗi lần cần kiểm tra giá trị của thanh ghi ACC lại phải tìm kiếm dữ liệu. Như vậy, giải pháp thứ hai sẽ được áp dụng.

4.2.2 Kiểm tra và ghi nhận mối quan hệ giữa các giá trị - biến bit

Nếu như ở giải pháp Kiểm tra kiểu, mối quan hệ biến byte - biến bit đã được cho trước và trình dịch ngược chỉ cần kiểm tra lại thì ở giải pháp này, vì không có thông tin từ phần chú thích của người lập trình, trình dịch ngược phải tự tìm kiếm các mối quan hệ giữa các giá trị địa chỉ vùng nhớ và biến bit, kiểm tra tính hợp lệ của chúng và ghi nhận vào danh sách dữ liệu. Các bước này được thể hiện ở hình 4.3. Giai đoạn kiểm tra tính hợp lệ nhằm đảm bảo là các bộ biến được sử dụng theo đúng nguyên tắc đã giới thiệu ở chương đầu.



Hình 4.3: Các bước kiểm tra và ghi nhận dữ liệu vào danh sách UnionDefine

Như đã trình bày ở phần 4.2.1, thay vì tìm kiếm các bộ biến byte - biến bit thì ở giai đoạn này sẽ tìm kiếm các bộ giá trị địa chỉ vùng nhớ - biến bit, vì vậy, cấu trúc lưu trữ dữ liệu tìm thấy được cũng phải được chỉnh sửa để ghi nhận mối quan hệ mới này. Cấu trúc UnionDefine được giới thiệu ở chương trước vẫn tiếp tục được sử dụng, tuy nhiên, được mở rộng thêm một trường dữ liệu mới là byteVarValue để ghi nhận giá trị địa chỉ vùng nhớ của bộ biến. Đoạn mã mới được trình bày bên dưới.

Listing 4.9: Đoạn mã mới của class UnionDefine

```

class UnionDefine{
public:
char* byteVar;
map<int, char*>* bitVar;
int byteVarValue;
};
  
```

Sau khi đã quét hết các câu lệnh ở các procedure, trước khi chuyển đổi các UnionDefine thành các khai báo ở ngôn ngữ C và thêm vào danh sách các biến toàn cục của program như giải pháp trước, cần phải thêm vào một bước chuyển đổi từ giá trị thành biến byte đại diện cho giá trị đó. Điều này có thể được thực hiện bằng cách chạy vòng lặp qua bảng lưu trữ các câu lệnh #DEFINE đã được thiết lập từ quá trình parse mã đầu vào. Nếu như có một giá trị nào đó chưa được khai báo ở câu lệnh #DEFINE, một biến byte mới sẽ được sinh ra để đại diện cho giá trị ấy. Mẫu tên biến byte sẽ là LOCATION_[giá trị của biến byte], ví dụ như LOCATION_38.

Như vậy, giải pháp này đã giải quyết được các vấn đề đặt ra của luận văn.

Chương 5

Kiểm thử

Bất kỳ một sản phẩm nào đều cần phải được kiểm tra trước khi công bố, luận văn này cũng không phải là một ngoại lệ. Để đảm bảo chất lượng được đánh giá một cách khách quan nhất, một hệ thống testcase với các loại tình huống được phân bổ một cách khoa học nhất sẽ được đưa ra, sau đó cho chạy thử qua cả 2 giải pháp của luận văn. Phần một của chương này sẽ trình bày về các testcase đó, phần hai sẽ trình bày cách kiểm tra chúng trên 2 giải pháp, và phần cuối sẽ là kết quả kiểm tra.

5.1 Hệ thống testcase

Có các tiêu chí phân loại testcase như sau:

- Loại biểu thức được gán vào thanh ghi ACC (tiêu chí I)
- Cách truy xuất bit của thanh ghi (tiêu chí II)
- Có vi phạm nguyên tắc sử dụng bộ biến hay không (tiêu chí III)

Với mỗi tiêu chí, ta sẽ có các trường hợp sau đây:

Tiêu chí I:

1. Một giá trị trực tiếp
2. Giá trị ở một vùng nhớ có địa chỉ là một biến byte
3. Giá trị ở một vùng nhớ có địa chỉ là một giá trị trực tiếp
4. Giá trị ở một vùng nhớ có địa chỉ là một thanh ghi
5. Giá trị ở một vùng nhớ có địa chỉ là một biểu thức 2 vế. Mỗi vế có thể là một biến byte, một thanh ghi, hoặc một giá trị trực tiếp

Tiêu chí II:

1. Truy xuất dựa vào một biến bit.
2. Truy xuất bằng cấu trúc truy xuất trực tiếp một bit của thanh ghi. Ví dụ: ACC.5

Tiêu chí III:

1. Không vi phạm nguyên tắc sử dụng. Nghĩa là: mỗi một biến bit chỉ được sử dụng khi thanh ghi ACC đang mang giá trị vùng nhớ có địa chỉ quy định bởi một biến byte duy nhất. Và không có hai biến bit nào cùng vị trí được sử dụng khi thanh ghi ACC đang mang giá trị của một biến byte nào đó.
2. Vi phạm nguyên tắc sử dụng. Một biến bit được sử dụng ở nhiều chỗ, trong các chỗ đó thanh ghi ACC mang giá trị của những biến byte khác nhau.
3. Vi phạm nguyên tắc sử dụng. Một biến bit được sử dụng ở một vị trí, tại vị trí đó thanh ghi ACC có thể mang giá trị của nhiều vùng nhớ khác nhau, không thể xác định trước khi thực thi chương trình.
4. Vi phạm nguyên tắc sử dụng. Một biến byte được sử dụng ở nhiều vị trí, sau câu lệnh gán biến byte, có 2 biến bit cùng một vị trí được sử dụng.

Dựa vào các tiêu chí và trường hợp trên, ta sẽ có tổng cộng $5 \times 2 \times 4 = 40$ loại testcase. Ngoài ra, sẽ có một testcase phức tạp được lấy từ một đoạn chương trình thực của doanh nghiệp được đưa vào kiểm thử, nhằm đảm bảo tính thực tế của luận văn này.

5.2 Phương thức kiểm thử

Vì số lượng testcase không quá lớn, và output ra của trình dịch ngược có khá nhiều thông tin khác ngoài phạm vi luận văn, nên ta sẽ dùng cách kiểm tra bằng tay. Hai bộ source code sẽ được đưa vào vòng lặp, chạy từ testcase 1 đến testcase 50. Output ở console sẽ được lưu vào file có định dạng [số thứ tự testcase]console.txt và đoạn mã đầu ra (nếu có) sẽ được lưu vào file có định dạng [số thứ tự testcase]code.txt. Sau đó người viết sẽ trực tiếp kiểm tra hai file này để xác định kết quả có đúng như mong muốn hay không.

5.3 Kết quả chạy thử

Kết quả chạy thử của 2 phương pháp được thể hiện ở bảng dưới.

Như vậy, có thể thấy giải pháp đầu tiên ra kết quả không chính xác rất nhiều, còn giải pháp Suy luận kiểu thì ra được kết quả chấp nhận được. Điều này đã được dự báo trước vì giải pháp Kiểm tra kiểu còn nhiều hạn chế.

Chương 6

Kết luận

Chương này sẽ tổng kết lại các kết quả đã đạt được của luận văn và đưa ra hướng phát triển trong tương lai.

6.1 Kết quả đạt được, khó khăn, điểm hạn chế

Nhìn chung, luận văn đã hoàn thành mục tiêu đề ra ban đầu, giải quyết được bài toán về kiểu dữ liệu bit và câu lệnh xử lý bit trong mã assembly của 8051. Ngoài ra, luận văn đã chứng minh được tính thực tiễn của đề tài, khả năng áp dụng vào thực tế của các doanh nghiệp có nhu cầu dịch ngược. Cuối cùng, luận văn cũng đưa ra một phương pháp lập testcase và kiểm thử khoa học, đảm bảo đưa ra được các trường hợp có thể xảy ra ngoài thực tế và đặc biệt có một testcase là một đoạn mã thật của doanh nghiệp. Kết quả kiểm thử trên bộ testcase dành cho phương pháp Suy luận kiểu là chấp nhận được.

6.2 Hướng phát triển trong tương lai

Các hướng phát triển trong tương lai của trình dịch ngược gồm có:

- Tiếp tục mở rộng khả năng dịch ngược cho nhiều máy khác nhau
- Phân tích và sửa lỗi sai của giải thuật phân tích dòng dữ liệu
- Cải tiến chức năng nhận dạng kiểu của Boomerang

Tài liệu tham khảo