

Compiling Business Processes: Untangling Unstructured Loops in Irreducible Flow Graphs

Wei Zhao

*Computer and Information Sciences Department
University of Alabama at Birmingham
Birmingham, AL 35294-1170, U.S.A.
zhaow@cis.uab.edu*

Rainer Hauser

*IBM Zurich Research
Saeumerstrasse 4, Rueschlikon 8803, Switzerland
rfh@zurich.ibm.com*

Kamal Bhattacharya

*IBM T.J. Watson Research
P O Box 218, Yorktown Heights, NY 10598
kamalb@us.ibm.com*

Barrett R. Bryant

*Computer and Information Sciences Department
University of Alabama at Birmingham
Birmingham, AL 35294-1170, U.S.A.
bryant@cis.uab.edu*

1. INTRODUCTION AND MOTIVATION

The evolution of programming languages (from machine languages, to assembly languages, high-level languages, and modeling languages) provides the technical foundation for the transition from machine-centric development to domain-centric development. The languages for machine-centric development are designed with machine specific considerations in mind such as cache, register, and imperative features. In domain-centric development, programming is performed with domain-specific concepts such as business tasks and business artifacts. Business process modeling [Fra04] is one type of domain-centric development that concerns the dynamic behavior of a business domain (an enterprise). A business process is composed by a set of atomic business tasks and/or sub-business processes. Business process modeling provides an appropriate programming abstraction for business level users.

Model Driven Architecture (MDA) [Fra03] is the result of programming language evolution. In the MDA framework, the business analysts create business process models, i.e. Platform Independent Models (PIMs). A business process PIM is not executable since the atomic tasks are abstract and non-executable semantic entities. A compilation engine can realize a PIM in different Platform Specific Models (PSMs) so that the process model can be executed. This paper discusses a compilation method called Regular Expression Language (REL) that compiles a business process PIM to a PSM, in which the atomic tasks are implemented as Web Services. Examples are based on a specific input PIM—UML activity diagram [UML03], and a specific output PSM—Business Process Execution Language for Web Services (BPEL4WS, in the rest of the paper, we will simply refer to it as BPEL) [BPEL03]. The contribution of REL is that a business process PIM that is irreducible with unstructured loops can be compiled into statements in structured PSM with controlled code complexity.

The rest of the paper is organized as follows. Section 2 presents an example and examines the problems we are addressing. Section 3 explains the details of REL. The implementation and experimental results are documented in

section 4. Section 5 compares our work to some related work. The paper concludes in section 6.

2. PROBLEM ANALYSIS

Throughout the paper, we will use the online product purchase system from [Koe05] as an example. The example, shown in figure 1, is simple and common enough to serve as an introduction, and yet complex enough to demonstrate various problems we are addressing. The description of the example is as follows:

From the initial page, an existing user has to logon to the system. If the logon succeeds, the authentication service is invoked. Authentication can fail if the user provides incorrect password, then the logon is repeated. If logon detects the user does not exist, he/she should be directed to the register function.

From the initial page, new users should register. If they succeed, they can go to the logon page. The registration page can be repeated in case of mistakes.

After authentication, the user can select and configure products. The process is flexible enough so that at each step, the user can go back to make changes. There is a verification service running in parallel to make sure the product purchase is secure.

The dynamic behavior of this business process is captured in the UML activity diagram in Figure 1. The activity diagram contains activity nodes and activity edges. The activity node includes executable nodes (represented as rounded rectangles denoting the atomic business tasks such as logon and register), the object nodes (an abstract activity node for defining object flow), and the control nodes (decision nodes, merge nodes, fork nodes, and join nodes). The executable node, the object node, and the merge node (node M in Figure 1) are undistinguished from the perspective of the compilation. The decision node (denoted as numbered D node, e.g., D1 tests whether the user is new) can be merged with the immediate predecessor executable node because the decision node provides no additional computational semantics as there is always a true semantics from an executable node to its immediate following decision node.

Therefore, for the purpose of compilation, Figure 1 can be normalized to Figure 2.

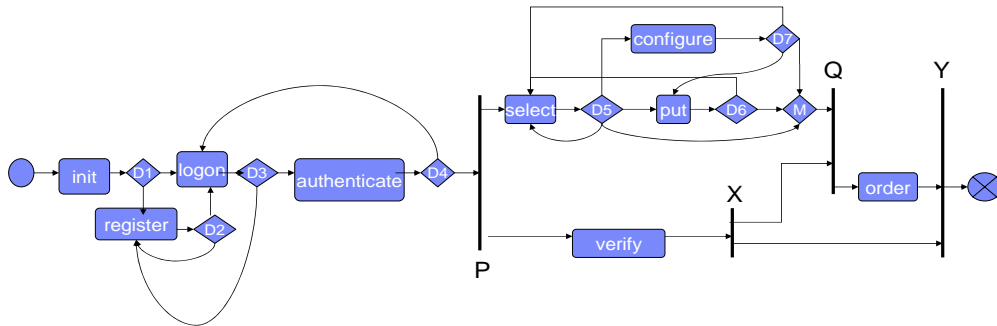


Figure 1. The activity diagram for the online purchase system.

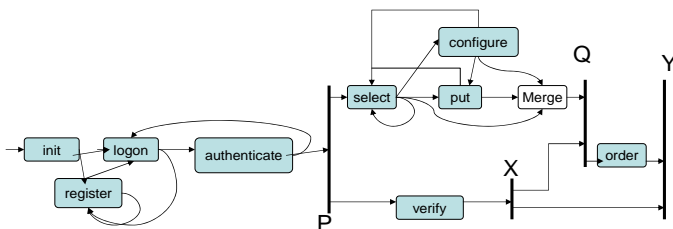


Figure 2. The normalized process model of Figure 1.

The activity edges in an activity diagram include the control flow and the object flow. The semantic content (the evaluation of a test condition, the flow of a particular type of object, or a condition of an event) of each edge is irrelevant for compilation. The compiler assigns each edge a unique id (e.g., the alphabetic letters in Figure 3). A hash table with the edge's id as the index contains the next state for this edge and the semantic content of the edge.

However, the semantics of how each transition edge is activated is important to the compiler. The action nodes, object nodes, decision nodes, and merge nodes have the XOR semantics on the associated edges, i.e., exactly one incoming edge is activated, and the activity node enables exactly one outgoing edge. The fork node has the AND semantics on its outgoing edges; the join node has AND semantics on its incoming edges.

It is clear that a UML activity diagram is formed by a set of "gotos". "Goto" is desirable at the business level modeling because of its simplicity. However, the target PSM, BPEL, is a web service orchestration language with explicit structured control flow constructs such as "switch", "while", and "flow" (for concurrency). The main task of the compiler is to translate the unstructured "goto" flows into well-structured statements in BPEL. We identify two problems in this translation, detailed in the following sections.

2.1 The Problem with Unstructured Loops

Figure 3 is an abstract representation for the part of the process before the fork node P in Figure 2. Each node is named by a letter. It is easy to see Figure 3 is a finite automaton (FA). A non-concurrent process model with a single start node and with the XOR transition semantics can always be normalized into an FA.

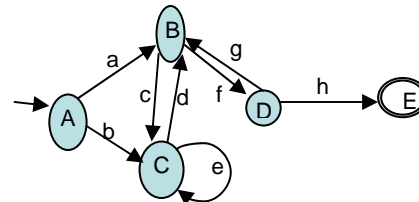


Figure 3. The FA for the authentication portion of Figure 2.

A=init; B=logon; C=register; D=authenticate; E=P

A naïve transformation algorithm could traverse the graph and output two types of code at each node: first, a conditional statement such as "switch" or "if-else"; second, a loop statement such as "while" if the node is the entry of the loop. In order to write out the loop statement, a standard loop detection algorithm, e.g., back-edge detection [Aho86] has to be adopted as a pre-process to detect the loop entry. Suppose we have detected B as the entry of the loop involved with B and D, we can output the pseudo BPEL code for Figure 3 as follows:

```
invoke A
switch
  case a, invoke B
    while f
      invoke D
      switch
        case h, invoke E, exit
        case g, invoke B
  case b, invoke C
  .....?
```

Each invoke statement represents a web service invocation. We notice that the loop involved with B and C cannot be represented because the loop body crosses both branches coming out of A. This type of loop is called an unstructured loop.

Definition 1: An *unstructured loop* is a loop that has more than one entry. For example, the loop BC has two entries B and C.

Definition 2: Node N *dominates* node M if every path from the initial node of the flow graph to M goes through N [Aho86].

Definition 3: A flow graph is *irreducible* iff it contains an unstructured loop of which one entry does not dominate all other entries.

The flow graph of Figure 3 is an irreducible graph because neither B nor C dominates each other. We refer to this type of unstructured loop as an irreducible loop. An unstructured loop that does not cause irreducibility is a *natural loop* (a loop that has a single dominating entry [Aho86]). The loop BD is a natural loop where B is the dominating entry (or the header). All the loops presented in a reducible graph are natural loops.

We consider this new definition of irreducible graphs suitable in our problem context. It allows us to identify the irreducibility of a graph by just looking at the graph without going through complex algorithms such as T1-T2 reduction or interval analysis. See [Aho86] for more discussion on irreducible loops.

A natural loop can be translated into structured statements as demonstrated in the example. However, the common techniques (node splitting [Coc69]) for converting an irreducible graph to a reducible graph require at least 2^{n-1} state duplication where n is the number of nodes in the original irreducible graph [Car03]. This becomes unmanageable for even a relatively small business process. Therefore, a method is needed to translate the irreducible graph directly.

2.2 Problem with Directed Acyclic Graphs

Even if the process model is cycle free, the naïve transformation method still has problems. Suppose we delete the edges c, e, and g in Figure 3, the remaining graph is a Directed Acyclic Graph (DAG). We may translate the remaining graph to the following code. We notice that the bold-face code has been repeated.

```

invoke A
switch
case a, invoke B
    if f, invoke D
        if h, invoke E
case b, invoke C
    if d, invoke B
        if f, invoke D
            if h, invoke E

```

After the analysis of the problems, we conclude that the naïve method only performs well for a tree-structured process model. Next, we will discuss REL on compiling a non-concurrent business process into BPEL code. The algorithm is customized to support the compilation of concurrent flows as well, but is out of scope of this paper. With REL we can solve the problems discussed in section 2.

3. THE REL ALGORITHM

One notable difference of a business process compiler from a traditional programming language compiler is that this compiler is a linearization tool that translates a graphic model in two-dimensional space to a well-structured textual language in one-dimensional space. By “one-dimensional”, we mean that a structured textual language defines a sequence of statement execution. By linearizing¹ a graph, we translate the logic of how each state is visited in a graph to a linear sequence of state visiting. Our approach transforms the graphic model first to an intermediate representation that is textual and theoretically equivalent to the source graphical model. We use the Regular Expression (RE) for this purpose. RE has structured control flow constructs: concatenation, or, and star. The alphabet set Σ of the RE in our context is the set of all edge-IDs in the process graph.

Traditionally, an RE is treated as a definition of a language (an instance of the regular language). We place a slightly different view: an RE sentence is a program written in a Regular Expression Language (REL). Since a grammar can be defined for REL, the well established compiler techniques (syntax directed translation) can be leveraged to translate this intermediate representation to the target language. Therefore, the compilation is comprised of two steps: 1) the translation from FA to RE; 2) the compilation from REL to BPEL. The first step obtains a logically correct linear representation of the source model; and the second step brings the intermediate logic representation to a correct format.

Our compilation method overall is thus termed REL. The REL can also be easily customized and extended to support specific features of our compilation system such as compiling concurrent processes.

Section 3.1 gives the details on how to convert an FA to an RE.

3.1 Linearization of Finite Automata

The set equation algorithm [Den78] is used to convert an FA to an RE. In the process of solving the equations, the strategies and optimizations are used to ensure the resulting RE is optimal. Section 3.1.1 discusses how the equations are extracted from an FA, and how they are solved. Detailed explanation and reasoning on RE optimizations and the strategies are presented in section 3.1.2 and 3.1.3

¹ We should be careful that a direct XML encoding of a graph is not a linearization of the graph as the encoding is still semantically two dimensional.

respectively. Section 3.1.4 summarizes how the problems we introduced in section 2 are solved.

3.1.1 Extract and Solve set equations

Definition 4: Let M be an FA, the *end set* $E(q)$ of a state q of M is the collection of input strings that can lead from q to an accepting state of M [Den78].

For example, the end set $E(A)$ of Figure 3 is the set of strings that lead the FA from state A to the accepting state E . A string in $E(A)$ consist of an “a” followed by the end set $E(B)$, or a “b” followed by the end set $E(C)$. Thus the definition of $E(A)$ can be written as:

$$E(A) = a E(B) + b E(C)$$

The “+” sign means “XOR”. For the readability, the previous equation is simplified to:

$$A = aB + bC \quad (1)$$

Similarly, we can define the end set for each state:

$$B = cC + fD \quad (2)$$

$$C = eC + dB \quad (3)$$

$$D = gB + hE \quad (4)$$

Since E is the final state:

$$E = \bullet \quad (5)$$

It is easy to see the language recognized by the FA is precisely those strings in the end set of FA’s initial state. That is, $L(\text{FA}) = E(A)$. The equations (1) through (5) comprise a system of right-linear set equations.

Proposition 1: The end sets of a finite automata $M=(Q, \Sigma, \delta, q_0, F)$ satisfy a system of right-linear set equations: for each $q \in Q$

$$E(q) = \sum_{q' \in Q} V(q, q') E(q') + W(q)$$

where

$$V(q, q') = \{x \in \Sigma \mid M \text{ has } q \xrightarrow{x} q'\}$$

and

$$W(q) = \begin{cases} \varepsilon & \text{if } q \in F \\ \emptyset & \text{otherwise} \end{cases}$$

We will be particularly interested in the solution of A , i.e., $E(A)$, that is the regular expression for the FA.

There are 3 types of rules to solve the set of equations (1) through (5).

1. Standard algebraic substitution: substitute an unknown with its value.
2. Standard RE algebraic laws:
 - a. Commutative +: $R+S = S+R$
 - b. Associative +: $R+(S+T) = (R+S)+T$
 - c. Associative concatenation: $R(ST) = (RS)T$
 - d. Distributive +: $R(S+T) = RS+RT$, $(S+T)R = SR+TR$
 - e. Anti-distributive is the inverse of d.
3. Arden’s rule for the removal of self-recursion. For example, $C = eC + dB = e^*dB$. Please refer to [Den78] for the proof of Arden’s rule.

Given a set of equations, there exists multiple syntactically different but equivalent REs as the solution for the start state. Since the RE is an intermediate step through which we generate the target code, the complexity of the RE logic directly reflects the complexity of the target code. An effective way we can generate an optimal RE from an FA is to apply the strategy during the process of equation solving. There are two strategies:

1. Arden’s rule is applied only before the node is going to be substituted.
2. There is a specific order based on which the nodes are substituted. The order is from bottom up in a dominator tree [Aho86] of a graph. Node m dominates node n if and only if m is an ancestor of n in the dominator tree. The dominator tree of Figure 3 is shown in Figure 4.

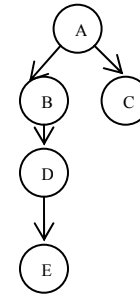


Figure 4. The dominator tree for the graph of Figure 3.

In section 3.1.3, we will show how the strategies are formulated and why they are important. Based on the strategies aforementioned, we solve the set of equations as follows:

We leave E out of consideration at the moment. Thus D is at the bottom of the dominator tree, and should be selected first. After D is substituted, the rest of the equations become:

$$\begin{aligned} A &= aB + bC \\ B &= cC + f(gB+hE) \\ C &= eC + dB \end{aligned}$$

Next, C should be substituted before B (the reason is explained in section 3.1.3). Before C can be eliminated, Arden’s rule should be applied to C to remove the recursion: $C = eC + dB = e^*dB$. After the substitution of C , A and B become:

$$\begin{aligned} A &= aB + be^*dB = (a+be^*d) B \\ B &= ce^*dB + f(gB+hE) = ce^*dB + fgB + fhE = (ce^*d + fg) B + fhE = (ce^*d + fg) * fhE \end{aligned}$$

Finally, after B is eliminated:

$$A = (a+be^*d) (ce^*d+fg) * fhE.$$

E is the final state in Figure 3 and it should be first chosen and substituted by nothing. However, in the complete process graph, E is only a place holder for the RE of the rest of the graph. The RE for Figure 3 is: $(a+be^*d) (ce^*d+fg) * fh$

3.1.2 Discussion on RE optimizations

Besides the semantic equivalence to the input model, the efficiency of the generated code is the most important consideration for every compiler. By taking an intermediate step, we apply the optimization of logic complexity to the RE level. The shorter the RE, the less number of loops, the less number of “or” constructs, the more optimal the RE is. The strategies and the optimization rules guarantee the optimal RE to be generated from an FA. Please note that all FAs in our problem domain are deterministic, therefore, all the REs in the solution set are deterministic. By “optimal”, we mean the optimal RE in the solution set. There might exist a more optimal RE than the “optimal RE” in the solution set. For example, a non-deterministic RE is always simpler than the REs solved from the equivalent deterministic FA. A non-deterministic RE is an RE from which only a non-deterministic FA can be built directly, e.g. $(a + b) * abc$ with $\Sigma = \{a, b, c\}$. It will be an interesting research topic to find out the optimal non-deterministic RE (if it exists) for a given deterministic FA, but it is not the concern of this paper.

There are two types of optimizations: 1) general RE optimizations; 2) loop look-ahead common sub-expression elimination (short for loop exit optimization).

General RE optimizations include common sub-expression elimination and loop normalizations. Common sub-expression elimination can be achieved by the anti-distributive law. A normalized loop after loop normalization is a loop in the form of:

$$\left(\sum_{x \in R} x \right)^*$$

where R is a regular expression. Some sample normalization rules are as follows ($u, v, w \in R$):

1. $(v^*)^* \Rightarrow v^*$
2. $(v^* + w^*)^* \Rightarrow (v + w)^*$
3. $(v^* w^*)^* v^* \Rightarrow (v + w)^*$
4. $(v^* w^*)^* \Rightarrow (v + w)^*$
5. $v^*(w v^*)^* \Rightarrow (v + w)^*$
6. $(v^* u^* + w)^* \Rightarrow (v + u + w)^*$

Loop exit optimization is necessary because of the natural differences of loops between structured languages and FAs. In RE and BPEL, there is only one point where the loop enters and exits. That point is called the *home state* of the loop. Some other structured languages allow loops to have multiple exits, e.g., multiple “break” statements can be used in a single “while” statement in Java. A loop with multiple exits can be easily simulated by a loop with a single exit by introducing an additional conditional variable. On the other hand, arbitrary loops in an FA can have multiple entry points and multiple exit points, and the entries and the exits are not necessarily the same. The problem of multiple entries will be discussed in section 3.1.3. The loop exit optimization enables correct translation of loop exits from FA to RE and BPEL.

To see an example, the loop involved with B and D in Figure 3 has two exits: B and D. When this loop is directly represented in the RE shown in section 3.1.1, the path inside the star operator starts and ends at the home state B (the node where Arden’s rule is applied). The alternative exit is represented as “fh” following the star operator. Multiple alternative exiting paths would be represented by an or operator following the star. However, the path following the star has an overlap “f”, called the loop look-ahead common sub-expression, with the body of the loop. The overlap results in the non-determinism at the node B. The loop exit optimization rule deletes the loop look-ahead common sub-expression in the strings that follow the star. A new syntactical marker “[]” is used to explicitly mark the loop exits (the letter that immediately follows the deleted string) inside the body of the loop. We thus get the following RE:

$$(a+be*d) (ce*d+f [h] g) * h$$

In the definition of REL, “[h]” indicates a special type of literal called loop exit. The loop exit optimization fails only with the non-deterministic RE because in that case the exit indicator could be the same as the looping condition. For example, the letter “a” in $(a+b)*abc$ is both the loop exit and the looping condition, in other words, two outgoing edges from one node are both named “a”. Luckily, since all edges are given a distinct ID in the FA in our problem context, i.e., all outgoing edges for any node are marked distinctly, no non-deterministic RE will be generated from the equation solver.

3.1.3 Discussion on equation solving strategies

We now discuss how the two strategies lead us to the optimal RE. Strategy 1 guarantees the minimum number of loops resulted from the equation solver. The number of loops in the final RE is equal to the number of applications of Arden’s rule. If Arden’s rule is applied not immediately before the node is to be substituted, there is always a chance that it will be applied again at the same node. For instance, if Arden’s rule is applied to B after D and before C are substituted, the set of equations become:

$$\begin{aligned} A &= aB + bC \\ B &= (fg)^* (cC + fhE) \\ C &= eC + dB = e*dB \end{aligned}$$

After C is substituted and Arden’s rule is applied to B again, we get:

$$\begin{aligned} A &= (a+be*d) B \\ B &= ((fg)^* ce*d) * (fg)^* fhE \end{aligned}$$

As can be seen, we obtain two extra loops in B. Although based on the loop normalization rule 3, $(fg)^* ce*d)^* (fg)^*$ is equivalent with $(fg+ce*d)^*$, it is better to get the simplest RE in the first round.

Strategy 2 is the most important factor that controls the complexity of the RE solution. To explain this strategy, we consider the following three cases.

Case 1, the flow graph is a DAG. The RE solution will be optimal disregarding the substitution order.

Proof. Because there is no loop in the graph, no nodes can be repeated on a path from the start node to the end node. On the other hand, an RE is a way to denote all possible ways to traverse a graph; thence, each node in a graph has to be visited as least once. Therefore, a path without node repetition is considered optimal. Multiple paths that share some nodes can be optimized by RE anti-distribution law. Furthermore, because there is no cycle in the graph, Arden's rule is not applicable during equation solving. Among all the equation solving rules, only Arden's rule is affected by the substitution order in producing solutions of different complexity. This will be explained more in cases 2 and 3. The RE algebraic laws do not distinguish the unknowns from their values and are therefore independent from the substitution order.

Case 2, the flow graph is a reducible graph and thus contains only natural loops. If and only if the header is substituted after the loop body, the RE solution is optimal.

Proof. Because of the nature of a loop, any arbitrary node on the loop can be treated as the start (the home state) of the loop. Any node substituted after all other nodes on the loop gets a recursion in its equation and therefore requires the application of Arden's rule. Whenever Arden's rule is applied to a node, that node becomes the home state of the loop.

In Figure 5 (a), three arbitrary nodes on a natural loop are denoted as A, B, and C. A is the header. The solid arrows starting from the loop show three arbitrary exits. If A is the home state, we get the following unique RE:

$$|QA| (|ABCA|)^* (a + |AB| b + |ABC| c)$$

The vertical bar denotes the path among a set of nodes. After the loop exit optimization, the above RE becomes:

$$|QA| (|AB [b] C [c] A|)^* (a+b+c)$$

The exit points are indicated explicitly along the path of the loop. As can be seen, each node in this graph is visited only once except the indication of the exits. Therefore, we claim, the above RE is the most optimal one for this graph.

However, if B is the home state, we will get the following string:

$$|QA| |AB| (|BC [c] A [a] B|)^* (b+c+a)$$

It is clear the path $|AB|$ is repeated before and in the star. The deeper B is below A, the bigger the repetition is. This shows that we get the optimal RE only if the loop header is the home state.

If a graph contains multiple natural loops, they either disjoint, or nest within another, or share the same loop head [Aho86]. In either case, the analysis of a single loop remains valid. For the loops that share the same loop head, the generated RE star loops are in the normal form.

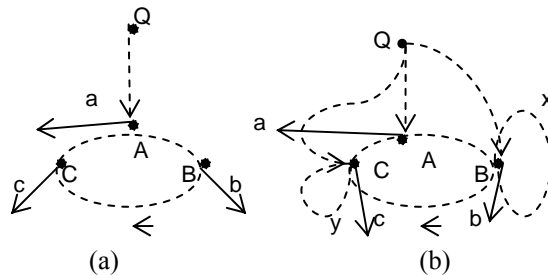


Figure 5. The natural loop (a) and the irreducible loop (b).

Dotted line: a path evolved with multiple nodes. **Solid line:** a path between two nodes. **Q** is the *common immediate dominator* for the entry (entries) of the loop. **The dotted circles are loops.** **The small arrow under a circle denotes the rotation direction of the loop.**

Case 3, the process graph is irreducible. If and only if the entry node that has the largest sum of end set and pre-set (if it is also the exit of the loop) is the home state of the loop, the RE solution is optimal.

We introduce two definitions before we present the proof.

Definition 5: An *Irreducible Loop Region (ILR)* consists of the Common Immediate Dominator (CID) of the loop entries and a set of nodes that can reach the loop exits without going through the CID. The CID of an ILR is called the header of the ILR. We can see that some nodes in an ILR are not on the loop.

Definition 6: The *pre-set* of a node in an ILR is the path from the CID to that node without going through the home state of the loop.

Proof. Figure 5 (b) is an example of the generalized ILR. The loop ABCA is an irreducible loop. A, B, and C are the entries to the loop. Loop x is a natural loop led by B; and loop y is a natural loop led by C. a, b, and c are exits at the entry points. Exits that are not also the entries of the loop should be handled the same way as in a natural loop, so are ignored in this case. Suppose B is the home state, the set equation algorithm generates the following pattern for this ILR.

$$\frac{(|QC|y^*|CA| + |QA|) |AB| + |QB|}{(x + |BC|y^*[c]|CA| [a] |AB|)^* (b+c+a)} + \frac{(|QC|y^*c + (|QC|y^*|CA| + |QA|) a)}{(|QC|y^*|CA| + |QA|) a}$$

The first underlined portion is how B is reached from Q; the second underlined part is the loop itself; and the third underlined part is how the loop can be exited without going through B. The repeated portion is $y^*|CA| |AB|$ in the first part and the whole string in the third underlined part. If we reorganize the repeated code, we see that $|AB| + a$ is the end set of A; $|CA| + y^* + c$ is the end set of C; $(|QC|y^*|CA| + |QA|)$ is the pre-set of A; $|QC|y^*$ is the pre-set of C. From the third underlined part, we know that: if an entry node is not also the exit of the loop, its pre-set will not cause the repetition. Therefore, the repetition of

the generated code for an ILR is the end sets of the entry nodes that are not the home state, and the pre-sets of the entry nodes that are also the exits and are not the home state of the loop. Thus if and only if the entry node that has the largest sum of end set and pre-set (if it is also the exit of the loop) is the home state of the loop, the RE solution is optimal.

To give an example, suppose we solve the equation of Figure 3 in a different order so that B is substituted earlier than C, then C becomes the home state, and the resulting RE is:

$$(a (fg) *c + b) (\underline{e+d(f[h]g) *c}) *h + a (f [h] g) *h$$

This RE is more complex than that we obtained previously. From Figure 3, the pre-set of B is $a(fg)^*$; the pre-set of C does not count since C is not the exit of the irreducible loop. The end set of B is larger than that of C (the next several paragraphs explain how the size of end set can be determined). Hence, choosing C instead of B as the home state results in a larger RE.

This strategy attempts to find the best solution. However, when each entry node is weighted the same, there will be no optimal solution as each solution will be equally complex. For instance, all REs for a strongly connected component are equally complex.

Knowing the correct order of the substitution, next, we discuss how this order is computed. If a graph is reducible, a depth-first ordering is sufficient since the dominating entry is always numbered lower than its dominated loop body [Aho86]. However, if the graph is not reducible, one pass of depth-first search is not enough to determine the correct order of substitution because depth-first search does not guarantee the ordering for nodes that do not dominate each other. For example, depth-first search could give the ordering for Figure 3 as ABDEC (numbered from 1 to 5). As C does not dominate BDE, so C could also be positioned before B, or after B and before DE.

Since we do not know the reducibility of the graph before we start to solve the set of equations, we use a uniform way to calculate the substitution order for both cases. First of all, a depth-first ordering of the nodes is computed followed by the algorithm presented in [Coo01] for finding the dominator set for each node. For efficiency, the finding dominator algorithm has to be built on top of depth-first ordering. After the dominator set for each node is determined, a dominator tree may be constructed.

If the graph is reducible, substituting the nodes bottom up based on a dominator tree guarantees the loop entry is substituted after the loop body. The bottom-up order ensures that the end set of the node that is to be substituted at each step contains only literals except the dominating loop entry (or entries in case of irreducible graph). This gives us the benefit that RE optimizations can be performed at each step to prevent the strings from getting too complex.

The irreducibility of the graph can be detected automatically when the equations are solved bottom-up based on a dominator tree. At the time the nodes at the same layer of the tree are to be substituted, if the end sets of the nodes do not contain each other, the graph is reducible and hence the nodes at the same layer can be substituted in any order. Otherwise, this indicates an irreducible loop, and we need to use the strategy laid out in case 3. At that particular situation, the size of the end set can be determined simply by measuring the length of the equation of the corresponding node because the equation would contain only literals except the other loop-entries. For example, in the example of section 3.1.1, the end set of B is larger than that of C after D has been substituted. The pre-set of an entry node can be computed by counting the hops from the CID, which is this node's direct parent in the dominator tree, to that node.

3.1.4 Summary on how the problems with DAG and unstructured loop are solved.

Removal of code repetition in a DAG can be achieved by the removal of common sub-expressions using the anti-distributive law. For the FA in Figure 3 with edges c, e, and g removed, the regular expression is $a fh + b d fh$, which is optimized to $(a + b d) fh$.

Regarding a loop, there are 4 missions: how to detect a loop; how to scope a loop; where the entries are; and where the exits are. The first two missions are automatically solved wherever Arden's rule is applied. The algorithm for calculating the substitution order selects a right loop entry. The loop exit optimization can detect loop exits and mark the exits explicitly inside the body of the loop.

3.2 Compilation from REL to BPEL

The mapping from REL to BPEL is straightforward. Table 1 shows the mapping from REL constructs to BPEL constructs. The pseudo BPEL code for the RE of Figure 3 is shown in Figure 6. The "if" statements in the pseudo code will be changed to conditions based on the evaluation of BPEL variables or will be deleted if there is no condition test and no event or data passed on that edge. The state information can be obtained by looking up the edge hash table.

Table 1. The mapping of syntactical constructs between REL and BPEL

construct language	sequential	alternative	loop	action node	flow control
RE	concatenation	+	*	state	input symbol
BPEL	sequence	switch case	while	web service invocation	conditions events data

```

invoke A
switch
  case a: invoke B
  case b: invoke C
    while e
      invoke C
    if d, invoke B
thisExit=false
while ((c or f) and not thisExit)
{ switch
  case c: invoke C
    while e
      invoke C
    if d, invoke B

  case f: invoke D
    if h, thisExit=true.
    if g, invoke B
}
if h, invoke E.

```

**Figure 6. The pseudo BPEL code for Figure 3.
The RE of Figure 3 is: (a+be*d)(ce*d+f [h] g)*h**

```

R0 ::= (R1)
{ :R0.code=R1.code;
R0.condition=R1.condition; :}

|R1R2
{ :R0.code=R1.code || R2.code;
R0.condition=R1.condition; :}

|R1+R2
{ :R0.code= "switch" || "case:" || R1.code
|| "case:" || R2.code;
R0.condition=R1.condition || "or" ||
R2.condition; :}

|R1*
{ :R0.code= loops[i] || "Exit=false" ||
"while" || R1.condition || "and not" ||
loops[i] || "Exit" || R1.code;
R0.condition=R1.condition;
i++; :}

|input
{ :R.code= "if" || input.lexeme ||
"invoke" || input.nextState;
R.condition=input.lexeme; :}

|[exit]
{ :R.code="if" || exit.lexeme || loops[i]
|| "Exit=true"; :}

```

Figure 7. The attribute grammar of REL for pseudo code generation.

Syntax-directed translation is used to generate the code of Figure 6. The attribute grammar specification for this translation is shown in Figure 7. The lexical symbols of REL shown in bold-face are: “()”, “[]”, “*”, and “+”. During the translation, code and looping-condition are synthesized attributes. The evaluations of the two attributes are shown as the semantic actions enclosed by “{ :; }”. For

the simplicity, only the synthesis of pseudo code is shown. “[]” means concatenation of code fragments. The syntax of the grammar is in BNF.

A global array variable “loops” is used to ensure that the loop exit information is bound to a correct scope. loops[i] means the current loop.

4. EXPERIMENTATION

The algorithm has been completely implemented in Java. The mathematical notations of RE are implemented as an object-oriented tree structure. For example, the object-oriented syntax tree of A’s equation is shown in figure 8. BPEL code generation is achieved by using the visitor pattern [Gam95] to walk through the RE syntax tree.

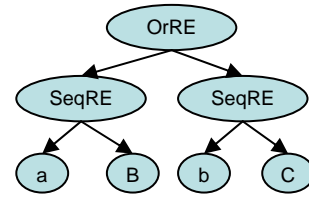


Figure 8. Equation for A=aB+bC

The first part of the REL algorithm, converting FA to RE, is presented in Figure 9.

Input: graph G with n nodes and m edges.

Output: the optimal RE that describes the same process.

Algorithm:

```

DepthFirstOrderOfG ← depthFirstSearch (G);
calculateDominatorSetForEachNode
(DepthFirstOrderOfG);
extractEquationsForEachNodeAndBuildDominatorTree;
Stack ← breadthFirstSearchOfDominatorTree;
while (Stack not empty) {
  Top ← Stack.firstElement();
  TopLevel ← Top.getDepthInDominatorTree();
  for (all nodes on TopLevel) {
    calculateWeightsOfNodes;
  }
  reorderNodesOnTheStackBasedOnWeights;
  for (all nodes on TopLevel) {
    substitute (Stack.pop() );
  }
}

```

Figure 9. The algorithm for converting FA to REL

Among all these operations, substitution is the most expensive one and takes up 80% computation in this algorithm. Other equation solving rules will be invoked when needed by the substitution operation. For example, Arden’s rule will be used before the substitution can take place if the equation is self recursive; rules such as distribution and anti-distribution are necessary preparation in order to apply substitution for some equations. Loop exit

optimization and loop normalization are performed after Arden's rule. During equation solving, no matter what rules are used, all equations should keep a very important mathematical invariant (right linear). However, this invariant will break in the practical implementation data structure after some manipulation of the equations. Therefore, some other normalization rules are also used after the application of each rule to keep the implementation data structure of the equation conforming to the mathematical properties. Two examples of these normalization rules are: if a $SequenceRE_1$ has a $SequenceRE_2$ as one of $SequenceRE_1$'s child, lift all $SequenceRE_2$'s children up as children of $SequenceRE_1$; if a $SequenceRE$ has a single child, replace the $SequenceRE$ with its child.

In terms of time complexity, `depthFirstSearch` takes $O(n)$ time. `calculateDominatorSetForEachNode` is based on the iterative dominator algorithm described in [Coo01]. Kam and Ullman [Kam76] proved that an iterative algorithm traversing the graph in depth first order will halt in no more than $d(G) + 3$ passes of G , where $d(G)$ is the depth of G . Knuth [Knu71] show that average flow graphs have depth 2.75. Therefore, `calculateDominatorSetForEachNode` takes $O(n)$ on average and $O(m*n)$ in the worst case. Plus, we have applied some implementation tricks to improve the performance of this operation: ordered vector rather than the set is used as the data structure; set union is implemented by appending to the end of the vector; set intersection is implemented as pair wise comparison of two vectors because the order of nodes in two vectors are always consistent; the set copy is simply an assignment of object reference as we use Java as our implementation language.

`extractEquationsForEachNode-AndBuildDominatorTree` takes $O(n)$. In the while loop, nodes will be processed layers by layers bottom-up in the dominator tree. `substitute()` will be performed exactly $O(n)$ times. Since `substitute()` is the core operation of the algorithm in figure 9, the algorithm overall takes $O(n)$.

For the experimentation, we start by using a freely available UML tool "Poseidon for UML"² to draw UML activity diagrams. Our algorithm takes the XMI (XML Metadata Interchange³) generated by Poseidon as the input. In order to experiment with large business processes, we constructed a random XMI file generator. We ran our algorithm on a laptop with 1.53 GHz AMD processor, 480MB of RAM, and Microsoft Windows XP operation system. Table 2 shows the performance of the algorithm and the output BPEL size over a set of hand-crafted and randomly generated processes.

Table 2. Experimental statistics for the REL algorithm

Number of nodes	Number of edges	Time (in milliseconds)	Output BPEL size (in lines)
3	9	80	99
5	8	60	46
6	11	100	102
14	16	120	62
40	106	770	1102
50	75	580	423
100	134	980	810

The largest benchmark in our system has 100 nodes. We do not anticipate any realistic business processes would exceed that limit. A business process that is too large should be modularized into sub-processes; each sub-process will be transformed into a separate BPEL process. Both the runtime duration of the algorithm and the output BPEL size depends on the number of nodes and the number of edges of the graph. The factor of the irreducible loop is hidden from this table. The time and output are linear to the number of nodes and edges if the graph does not contain irreducible loops, e.g., the graph of nodes 5, 14, and 50 are of this category. The more ILRs a graph contains, the longer the runtime duration is and the bigger the output size is. In the worst case (strongly connected component), the output size is exponential to the number of nodes. In our experiment, the first test case is a strongly connected component with 3 nodes. For the graphs that contain some ILRs but are not strongly connected components, the REL algorithm generates BPEL of an optimal size.

5. RELATED WORK

There are different ways an FA can be translated into a language that has structured control flow statements.

It has been proved by [Boh66] and also shown in [Hau04] that every state machine can be translated into a program with a single while and a single switch statement. Let us call this the state machine controller approach. Based on this approach, the FA in Figure 3 can be translated to the following code:

```

nextnode=A
while nextnode !=E
  switch
  case nextnode =A
    invoke A
    if a, nextnode =B
    if b, nextnode =C
  case nextnode =B
    invoke B
    if f, nextnode =D
  .....
```

The drawback of this approach is run time inefficiency for complex business processes. There will be an average test

² <http://www.gentleware.com/index.php>

³ <http://www.omg.org/technology/documents/formal/xmi.htm>

time of $O(n/2)$ for every state transition where n is the number of nodes in the process graph. It can be seen that the semantic process structure is not revealed from the code at all because of the uniform code representation, which gives poor code comprehensibility. Furthermore, this approach encodes the goto, a low-level language concept, in high-level language constructs.

Graph reduction [Koe05] has been used to translate a reducible flow graph to structured code. Code is output after each step of T1-T2 reduction until the graph has been reduced into a single node. However, this method leaves the problem of irreducible graphs open. The style of the code generated from this method is of the compact form below:

```
repeat
  invoke A
  if e, invoke D
  if ( eg or c ), invoke C
while ( a or ef or cd or egd )
if ( b or ei or ch or egh ) , invoke B
```

The code of this style, without additional augmentation such as concurrency in implementation, highly depends on the order of the statements. For example, in this code fragment

```
if ( eg or c ), invoke C
if e, invoke D
```

the first statement cannot be evaluated since the condition g is produced by activity D . We refer to the code of this style as the code that does not preserve the “look-ahead=1” semantics of FA, the semantics that ensures an FA makes one transition by reading in only one input symbol. Meanwhile, the long string of the conditions gives poor compensability of the code. Just for comparison, the code generated for this reducible graph using the method presented in this paper is shown in the appendix.

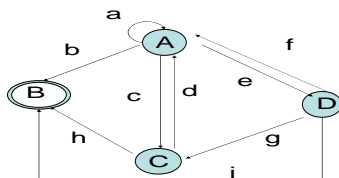


Figure 10 The FA for the product selection portion of Figure 2.

A: select; **B:** merge; **C:** put; **D:** configure.

The algorithms based on goto eliminations [Hau04] and continuation semantics [Amm92, Koe04] to untangle the unstructured loops have a very similar set of rules as the set equation algorithm. The main difference is that these approaches do not guarantee the optimality of the generated code for the irreducible process graph. These algorithms also use an intermediate representation similar to pseudo code. By using RE, we systematically factor out the structure information that is only relevant to the

algorithm. Another important difference is that these algorithms work on the states while the method in this paper work on the edges. Consequently, the code generated by [Amm92, Hau04, Koe04] is in a similar style of that resulted from graph reduction. On the other hand, our method introduces state repetitions in the code as the length of the generated code is bound to edges rather than states.

RE has been used as an intermediate step through which a program with goto statements can be transformed into programs without gotos [Mor99]. However, the method uses a uniform RE “ $E_1 * E_2$ ” to represent the program, where E_1 is the elementary nontrivial paths back to the start node and E_2 is the paths to the stop node that do not return to the start node. The uniform RE hurts the natural structure of the process graph. More important, the method did not pay attention to the optimizations. The resulting RE is usually unnecessarily complex. Comparing to other algorithms (e.g. [Hop79]), converting an FA to an RE, the set equations algorithm gives us the opportunity for optimizations and is quite straightforward and easy to implement.

Although model transformation is quite a popular topic today, there is much work on translating UML activity diagrams or similar behavior models [Pat04, Sko04] to structured languages that does not consider the transformation for loops at all.

6. CONCLUSIONS AND OUTLOOK

In this paper, a novel algorithm is presented for compiling a business process model with irreducibility to a structured flow language. The complexity of the code is controlled by using different levels of optimizations. Strategies are used during equation solving to ensure the optimal solution. The generated RE can also be optimized using general RE optimization rules. The benefits of our approach is that the logic related optimizations are at the RE layer, which makes the optimizations more efficient, easier to understand and to implement. Generated target code can also be optimized for platform specific concerns although we do not discuss this aspect in the paper.

Taking the RE as an intermediate step enables the reuse of core implementation of the algorithm to different source models and target languages. This supports the vision of mapping a PIM to multiple PSMs. The compilation framework discussed in this paper by taking an intermediate step could be a good practice in implementing a compiler from a graphic model to a textual language in general. Having an intermediate representation is a general solution in traditional compiler construction as well.

There are some limitations in our approach. The algorithm assumes the transition semantics in the model are XOR and AND (in the case of concurrency). The algorithm fails for transitions with OR semantics. Fortunately, OR semantics is not a standard that is supported by all the business modeling tools⁴. According to [Woh02], OR transitions in

⁴ <http://tmitwww.tm.tue.nl/research/patterns/standards.htm>

BPEL can be represented by links in the style of Web Service Flow Language (WSFL). Such links are not structured constructs but straightforward encodings of the model. Thus translating the model to this representation is outside our problem domain. Patterns from workflowpatterns.com, which are realized in BPEL by “link”, “pick”, and “serializable scope” but are not considered in our algorithm, are: Multiple Choice, Implicit Termination, Interleaved Parallel Routing, and Milestone.

Under the current implementation of BPEL, although a BPEL program invokes services distributed over several servers, the orchestration of these services is centralized on one server. The orchestration is written in structured languages. On the other hand, the execution control of an FA-based model is distributed in nature considering that each state stands for a server hosting a web service. We encountered the problem of unstructured loop because we translate the distributed to centralized execution control. Efforts are taking place at IBM to decentralize BPEL [Nan04]. This work partitions a centralized BPEL program into decentralized processes. If in the future BPEL becomes decentralized, there are two implications: 1) the unstructured loop problem will go away in the translation from UML activity diagrams to BPEL; and 2) the work of [Nan04] is not really needed because the centralized BPEL might go away as well and decentralized BPEL comes directly from the process model. Nevertheless, the algorithm presented in this paper still applies to translation from a business process model to any other structured language if it is not BPEL.

Independent of the model transformation domain, this paper also contributes to the set equation algorithm by introducing equation solving strategies that gives the optimal solution. Meanwhile, it is new that the irreducibility is solved by using set equation algorithm and RE as traditionally the problem of irreducibility is solved by node-splitting techniques.

To fully justify the usefulness of this work, we also need to answer these questions: why bother to translate unstructured flows into structured representations? and why not design the modeling language to have the structured loops? In fact, UML2 [UML03 pp. 341-342] does provide a looping construct although the specification is incomplete. Based on our customer engagement, we found out it is difficult for the business level user to comfortably use such a loop construct. It is rather easier for them to understand the concept of “goto” and self-loop (a single node loop). Therefore, frequently business level users use the tools such as UML to construct unstructured “goto” flows. It is important we have an algorithm to translate such unstructured flows to structured representations if we choose a structured representation as the executable PSM.

7. ACKNOWLEDGMENTS

The authors thank Pranam Kolari of the University of Maryland, Baltimore County, for his participation in the

early development of this work. We thank Brent Hailpern and Tim Klinger of IBM T. J. Watson Research, and Jana Koehler and Jochen Kuster of IBM Zurich Research for their review and editing of the paper. Thanks are also extended to Kumar Bhaskaran of IBM T. J. Watson Research for his valuable comments and support.

8. REFERENCES

- [Aho86] Aho, A., Sethi, R., and Ullman, J., *Compilers-Principles, Techniques, and Tools*, Addison-Wesley, 1986.
- [Amm92] Ammarguella, Z., “A Control-Flow Normalization Algorithm and Its Complexity”, *IEEE Transactions on Software Engineering*, Vol. 18, No. 3, 1992.
- [Boh66] Bohm, C., and Jacopini, G., “Flow Diagrams, Turing Machines and Languages with Only Two Formation Rules”, *Communications of the ACM*, Vol. 9, No. 5, pp. 366-371, 1966.
- [BPEL03] Business Process Execution Language for Web Services, Version 1.1, 05 May 2003, <http://www-106.ibm.com/developerworks/library/ws-bpel/>
- [Car03] Carter, L., Ferrante, J., and Thomborson, C., “Folklore Confirmed: Reducible Flow Graphs are Exponentially Larger”, *Proc. of the 30th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pp. 106-114, 2003.
- [Coc69] Cocke, J., and Miller, R. E., “Some Analysis Techniques for Optimizing Computer Programs”, *Proc. of 2nd Hawaii International Conference on Systems Sciences (HICSS)*, pp. 143-146, 1969.
- [Coo01] Cooper, K., Harvey, T., Kennedy, K., “A Simple, Fast Dominance Algorithm”, *Software—Practice and Experience*, Vol. 31, No. 4, pp. 1-10, 2001.
- [Den78] Denning, P. J., Dennis, J. B., and Qualitz, J. E., *Machines, Languages, and Computation*, Prentice-Hall, Inc., 1978.
- [Fra03] Frankel, D. S., *Model Driven Architecture: Applying MDA to Enterprise Computing*. Wiley Publishing, Inc., 2003.
- [Fra04] Frank, J. H., Gardner, T. A., Johnston, S. K., White, S. A., and Iyengar, S., “Business Processes Definition Metamodel Concepts and Overview”, IBM’s proposal in response to the OMG’s RFP for a Business Process Definition Metamodel, www.bpmn.org/Documents/BPDM/BPDM%20Whitepaper%202004-05-03.pdf
- [Gam95] Gamma, E., Helm, R., Johnson, R., Vlissides, J., *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1995.
- [Hau04] Hauser, R., and Koehler, J., “Compiling Process Graphs into Executable Code”, *Proc. of the 3rd International Conference on Generative Programming and Component Engineering (GPCE’04)*, pp. 317-336, 2004.

[Hop79] Hopcroft, J., and Ullman, J., *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley Publishing Company, Inc. 1979.

[Kam76] Kam, J., Ullman, J., “Global data flow analysis and iterative algorithms”, *Journal of the ACM*, Vol. 23, No. 1, pp. 158-171, 1976.

[Koe04] Koehler, J., and Hauser, R., “Untangling Unstructured Cyclic Flows - A Solution based on Continuations”, *Proc. of the International Conference on Cooperative Information Systems*, pp. 121-138, 2004.

[Koe05] Koehler, J., Hauser, R., Sendall, S., and Wahler, M., “Declarative Techniques for Model-Driven Business Process Integration”, *IBM Systems Journal*, Vol. 44, No. 1, pp. 47-65, 2005.

[Knu71] Knuth, E., “An empirical study of FORTRAN programs,” *Software – Practice and Experience*, Vol. 1, No. 2, pp. 105-133, 1971.

[Mor99] Morris, P. H., Gray, R. A., and Filman, R. E., “GOTO Removal Based on Regular Expressions”, *Journal of Software Maintenance: Research and Practice*, Vol. 9, No. 1, pp. 47-66, 1999.

[Nan04] Nanda, M. G., Chandra, S., Sarkar, V., “Decentralized Execution of Composite Web Services”, *Proc. of the 19th Annual ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2004.

[Pat04] Patrascoiu, O., “Mapping EDOC to Web Services Using YATL”, *Proc. of the 8th International IEEE Enterprise Distributed Object Computing Conference*, 2004.

[Sko04] Skogan, D., Gronmo, R., and Solheim, I., “Web Service Composition in UML”, *Proc. of the 8th International IEEE Enterprise Distributed Object Computing Conference*, 2004.

[UML03] UML 2.0 Superstructure Final Adopted Specification, ptc/03-08-02, <http://www.omg.org/cgi-bin/doc?ptc/2003-08-02>

[Woh02] Wohed, P., van der Aalst, W. M. P., Dumas, M., and ter Hofstede, A. H. M., “Pattern Based Analysis of BPEL4WS”, QUT Technical report, FIT-TR-2002-04, Queensland University of Technology, 2002.

APPENDIX

Based on the method presented in this paper, the RE generated for Figure 10 is:

```
(a+c[h]d +e[i](f+g [h] d))* (b +h +i)
```

The pseudo BPEL code is:

```
thisExit=false
while (a or c or e and not thisExit)
  switch
    case a, invoke A
    case c, invoke C
      if h, thisExit =true
      if d, invoke A
    case e, invoke D
      if i, thisExit=true
      switch
        case f, invoke A
        case g, invoke C
          if h, thisExit=true
          if d, invoke A
```