INSTITUTE OF AERONAUTICAL ENGINEERING

(AUTONOMOUS)

Dundigal, Hyderabad - 500 043

LECTURE NOTES:

COMPILER DESIGN(ACSC40)

DRAFTED BY :

Dr.U Sivaji(IARE 10671)

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF AERONAUTICAL ENGINEERING

November 10, 2023

Contents

List of Figures

Chapter 1

INTRODUCTION TO COMPILERS

Course Outcomes

After successful completion of this module, students should be able to:

CO 1	**Describe** the components of a language processing system for the conversion of high level languages to machine level languages.	Apply
CO 2	**Classify** the importance of phases of a compiler for constructing a compiler.	Understand
CO 3	**Demonstrate** a lexical analyzer from a specification of a Language's lexical rules for dividing the programming statements into tokens.	Understand
CO 6	**Construct** LEX and YACC tools for developing a scanner and a parser.	Apply

1.1 INTRODUCTION TO COMPILERS

Introduction to compilers: Definition of compiler, interpreter and its differences, the phases of a compiler; Lexical Analysis: Role of lexical analyzer, input buffering, recognition of tokens, finite automata, and regular Expressions, from regular expressions to finite automata, pass and phases of translation, bootstrapping, LEX-lexical analyzer generator.
Overview Of Language Processing system

1.1.1 Preprocessor

A preprocessor produceinput to compilers. Theymayperform the followingfunctions.
1. Macroprocessing:Apreprocessormayallowausertodefinemacrosthatare shorthandsforlonger constructs.
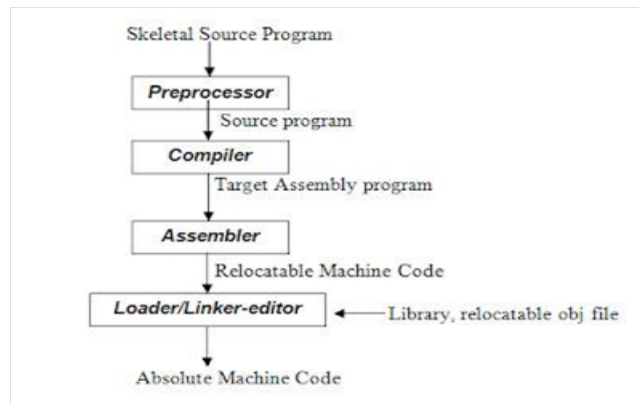
FIGURE 1.1: Language Processing system.

2.Fileinclusion:A preprocessor mayincludeheaderfiles into the program text.

3.Rationalpreprocessor:thesepreprocessorsaugmentolderlanguageswithmoremodern flow-of-control and data structuringfacilities.

4.LanguageExtensions:Thesepreprocessorattemptstoaddcapabilitiesto thelanguagebycertain amounts to build-in macro.

1.2 Compiler

Compiler is a translator program that translates a program written in (HLL) the source program and translates it in to an equivalent program in (MLL)the target program. As an important part of a compiler is error showing to the programmer. ExecutingaprogramwrittennHLLprogramminglan-
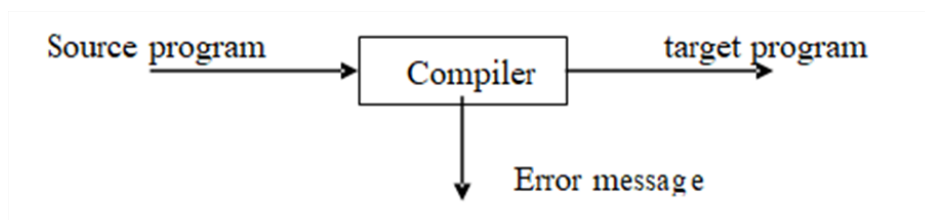
FIGURE 1.2: Executing a program

guageisbasicallyoftwoparts.the sourceprogrammustfirstbecompiledtranslatedintoaobjectprogram.Thentheresults object program is loadedinto a memory executed.

1.2.1 Assembler

programmersfounditdifficulttowriteorreadprogramsinmachinelanguage.Theybegintouseamnemonic(symbols)forea programiscalledsource program,theoutputisamachinelanguagetranslation (object program).

1.2.2 Interpreter

An interpreterisaprogramthatappearstoexecuteasourceprogram as if it were machine language Lan-
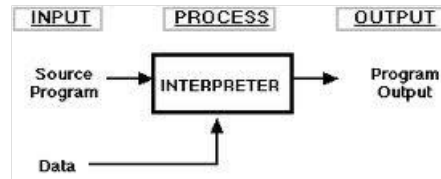
FIGURE 1.3: Executing a program

guages such as BASIC,SNOBOL,LISPcanbetranslatedusinginterpreters.JAVAalsouses interpreter. Theprocess of interpretation canbe carried out in followingphases.

1. Lexical analysis
2. Syntaxanalysis
3. Semantic analysis
4. Direct Execution

1.2.3 Advantages:

FIGURE 1.4: Advantages

1.2.4 Disadvantages:

1. The execution of the program is slower
2. Memory consumption is more.

1.2.5 Loader and Link-editor:

Oncetheassemblerproceduresanobjectprogram,thatprogrammustbeplacedintomemoryandexecuted.Theassemblerco inmemoryandtransfercontroltoit,therebycausingthemachinelanguageprogramtobeexecute.Thiswouldwastecoreby leavingtheassemblerinmemorywhiletheuser'sprogramwasbeing executed.Alsotheprogrammerwouldhave

toretranslatehisprogramwitheachexecution,thuswasting translationtime.Toovercomethisproblemsofwastedtranslatio
"Aloaderisaprogramthatplacesprogramsintomemoryandpreparesthemforexecution."Itwouldbemoreefficientifsubro
fourfunctions.

1.3 Translator

Atranslatorisaprogramthattakesasinputaprogramwritteninonelanguageand producesasoutputaprogram-
inanotherlanguage.Besideprogramtranslation,thetranslator performsanotherveryimportantrole,theerror-
detection.AnyviolationofdHLL specificationwould be detectedandreportedtotheprogrammers.Importantroleoftrans
are:

1. Translatingthe hll program input into an equivalent ml program.
2. Providingdiagnosticmessageswherevertheprogrammerviolatesspecificationof the hll.

Type Of Translators:-
1. Interpreter
2. Compiler
3. preprocessor

List of ompiler
1. ADA compliler
2. ALGOL compiler
3. BASIC compiler
4. C++ compiler
5. C *hash* compiler
6. C compiler
7. COBOL compiler
8. Java compiler

1.3.1 Phases Of A Compiler:

A compiler operates in phases.A phase isalogicallyinterrelatedoperationthattakessourceprograminonerepresentation
Thephases of a compilerare shown in below There aretwo phases of compilation.
a. Analysis(MachineIndependent/Language Dependent)
b.Synthesis(Machine Dependent/Languageindependent)

Compilation process is partitioned into no-of-subprocessescalled 'phases' **Lexical Analysis:-** Lex-
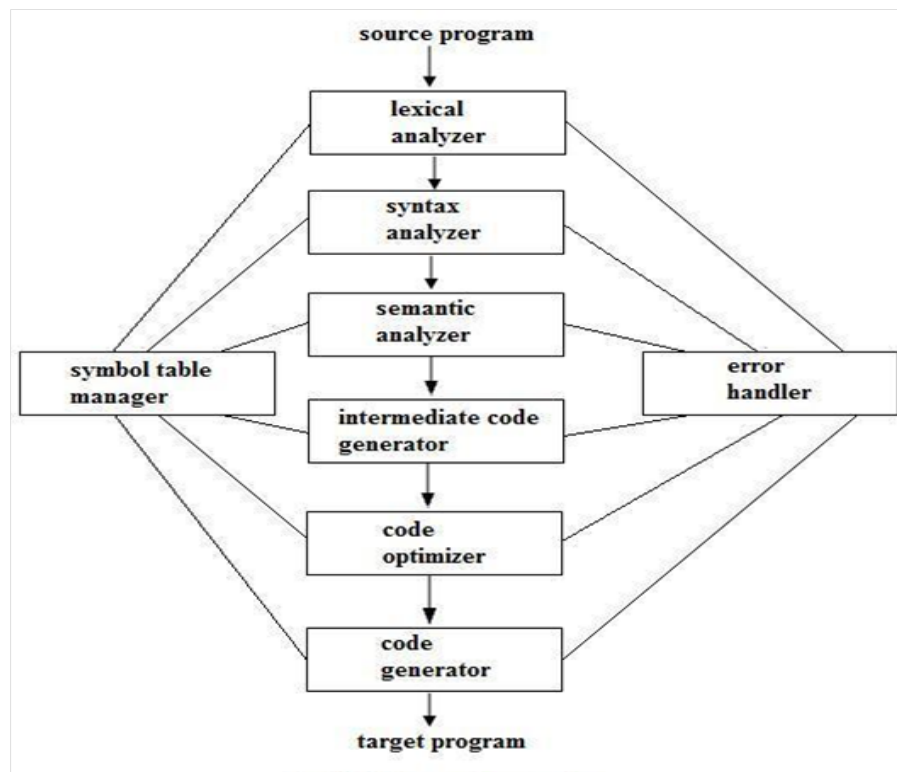
FIGURE 1.5: Phases of a compiler

ical AnalysisorScannersreadsthesourceprogramonecharacteratatime,carvingthe sourceprogram into a sequenceof automatic units called tokens.

Syntax Analysis:- Thesecondstageoftranslationiscalledsyntaxanalysisorparsing.Inthis phaseexpressions,statements,declarationsetc. . . areidentifiedby usingtheresultsoflexical analysis.Syntax analysis is aided by using techniques based on formal grammar of the programming language.

Intermediate Code Generations:- An intermediate representation of the finalmachinelanguagecodeisproduced. This phasebridges the analysis and synthesis phases of translation.

Code Optimization:- Thisisoptionalphasedescribedtoimprovetheintermediatecodesothatthe output runsfaster and takes lessspace.

Code Generation:- Thelastphaseoftranslationiscodegeneration.Anumberofoptimizationsto Reducethelengthofmachinelanguageprogramarecarriedoutduringthisphase.The output of the code generator is the machine languageprogram ofthe specified computer. TableManagement(or)Book-keeping:- Thisistheportiontokeepthenamesusedbytheprogramandrecords essentialinformationabouteach.Thedatastructureuse ‘Symbol Table’.

Error Handlers:- It is invoked when aflawerrorin the sourceprogramis detected. Theoutputof-LAisastreamoftokens,whichispassedtothenextphase,the syntaxanalyzerorparser.TheSAgroupsthetokenstogetherin

asexpression.Expressionmayfurtherbecombinedtoformstatements.Thesyntactic structurecan beregardedas a tree whoseleavesarethe token called as parsetrees.

The parser has two functions.It checks if the tokens from lexical analyzer, occurinpatternthatarepermittedbythespecificationforthesourcelanguage.Italso imposes on tokens a tree-like structurethat is usedbythe sub-sequent phases of the compiler.

Example,if a program contains the expression A+/B after lexical analysis this expression might appear to the syntax analyzer as the token sequence id+/id.On seeingthe/, the syntax analyzer should detect an error situation,because the presence of these two adjacent binary operatorsviolates the formulations rule of an expression.

Syntax analysis is to make explicit the hierarchical structure of the incoming token stream by identifying which parts of the tokens tream should be grouped.

1.3.2 IntermediateCodeGeneration:-

The intermediate code generation uses the structure produced by the syntax analyzer to create a stream of simple instructions.Many styles of intermediate code are possible.One common style uses instruction with one operatorand a small number of operands. The output of the syntax analyzer is some representation of aparsetree.The intermediate code generation phase transforms this parsetree into an intermediate language representation of the source program.

Code Optimization:- This is optional phasedescribed to improve the intermediate code so that the output runs faster and takes lesss pace.Its output is another intermediate code program that does thesame job as the original, but in a waythat saves time and / orspaces.

/* 1,Local Optimization:-

Therearelocaltransformationsthatcanbeappliedtoaprogramto make an improvement.Forexample,

IfA > Bgoto L2

Goto L3L2:

This can bereplaced bya single statement IfA < B go to L3

Another important local optimization is the elimination of common sub-expressions

A:=B+C+D

E:=B+C+F

Might be evaluated as

T1:=B+C

A:=T1+D

E:=T1+F

Takethis advantage of the common sub-expressionsB+C.

Loop Optimization:- Anotherimportantsourceofoptimizationconcernsaboutincreasingthespeedofloops.Atypicallo

computationthatproducesthesameresulteachtimearoundtheloop to a point, in the programjustbe-forethe loop is entered.*/

Codegenerator:- Cproducestheobjectcodebydecidingonthememorylocationsfordata, selectingcodetoac-cesseachdataandselectingtheregistersinwhicheachcomputationis tobedone.Many computershaveonly afewhighspeedregistersinwhichcomputationscan beperformedquickly.Agoodcodegeneratorwouldattempttoutilizere as possible.

ErrorHanding:- Oneofthemostimportantfunctionsofacompileristhedetectionand reporting oferrorsinthe-source program.Theerror messageshouldallowtheprogrammerto determineexactlywheretheerror-shaveoccurred.Errorsmayoccurinallorthephasesofa compiler.

Whenever a phase of the compilerd is covers an error,it must report the error to the error han-dler,which is sues anappropriated iagnosticmsg.Both of the table-management and error-Handling routines interact with all phases of the compiler.

1.3.3   1LexicalAnalyzer:

The LA is the first phase of a compiler. Lexical analysis is called as linear analysis or scanning. In this phase the stream of characters making up the source program is read from left-to-right and grouped into tokens that are sequences of characters having a collective meaning. Upon receiving
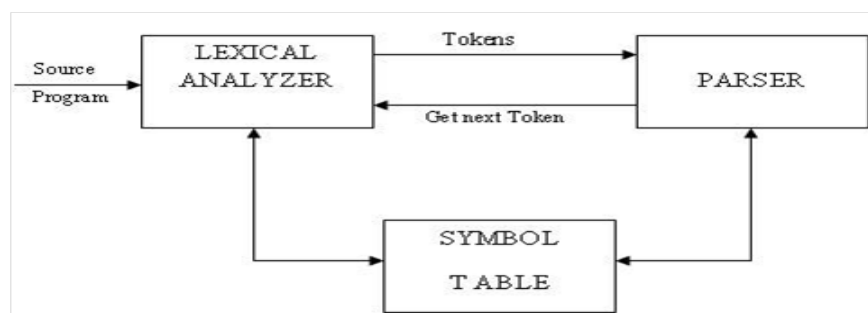
FIGURE 1.6: Lexical Analyzer

a'getnext token 'command form the parser, the lexical analyzer reads the input character until it can identify then exttoken.The LA return to the parse representation for the token it has found. There presentation will be an integer code,ifthe token is a simple construct such as parenthesis, comma or colon.

LA mayalsoper form certain secondary tasks as the user interface. One such task is striping out from the source program the commands and white spaces in the form of blank tab and new line

characters. Another is correlating error message from the compiler with the source program.

Lexical Analysis Vs Parsing

Lexical analysis	Parsing
A Scanner simply turns an input String (say a file) into a list of tokens. These tokens represent things like identifiers, parentheses, operators etc. The lexical analyzer (the"lexer") parses individual symbols from the source code file into tokens. From there, the"parser"proper turns those who le tokens into sentences of your grammar	A parser converts this list of tokens into a Tree-like object to represent how the tokens fit together to form a cohesive whole (sometimes referred toasa sentence). A parser does not give the no des Any meaning beyond structural cohesion. The next thing to do is extract meaning from this structure (sometimes called contextual analysis).

FIGURE 1.7: Lexical Analysis Vs Parsing

Token, Lexeme, Pattern:

Token: Token is a sequence of characters that canbe treated as a single logical entity.Typical tokens are,

1)Identifiers2) keywords 3) operators4)special symbols5) constants

Pattern: Aset ofstringsin the input forwhichthesame token is producedas output. This set of strings is described byarule called apattern associated with the token.

Lexeme: Alexemeisasequenceofcharactersinthesourceprogramthatismatchedbythepattern for a token. A pattern is a ruled scribing the set of lexemes that can represent a particular token in source

Token	lexeme	pattern
const	const	const
if	if	If
relation	<,<=,= ,<>,>=,>	< or<=or =or<>or>=or letter followed by letters & digit
i	pi	Any numeric constant
nun	3.14	Any character b/w" and "except"
literal	"core"	pattern

FIGURE 1.8: Description of token

program.

Lexical Errors: Lexical errors are the errors thrown by the laxer when unable to continue. Which means that there's no way to recognize a lexeme as a valid token for you lexer ? Syntax errors, on the other side, will be thrown by yours canner when a given set of already recognized valid tokensdon'tmatchanyoftherightsidesofyourgrammarrules.Simplepanic-modeerror handling system requires that were turn to a high-level parsing function when a parsingor lexical error is detected.

1.3.4   Difference Between Compiler And Interpreter:

1. Acompilerconvertsthehighlevelinstructionintomachinelanguagewhileaninterpreter converts the high level instruction into an intermediateform.

2. Beforeexecution,entireprogramisexecutedbythecompilerwhereasaftertranslatingthefirst line, an interpreter then executes it and so on.

3. Listoferrorsiscreatedbythecompilerafterthecompilationprocesswhileaninterpreterstops translatingafterthefirst error.

4. Anindependentexecutablefileiscreatedbythecompilerwhereasinterpreterisrequired by an interpreted program each time.

5. The compiler produceobject codewhereas interpreterdoes not produceobject code.In the process of compilation the program is analyzed on lyonce and then the code is generated where as source program is interpreted everytime it is to be executed and everytime the source program is analyzed.Hence interpreter is less efficient than -compiler.

6. Examples of interpreter:AUPS Debugger is basically a graphical source level debugger but itcontains built in C interpreter which can hand le multiple source files.

7. Example of compiler:Borland ccompileror Turbo Ccompiler compiles the programswritten in Cor C++.

1.4   Regular Expressions:

1.4.1   Specification Of Tokens

There are 3 specifications of tokens:
1) Strings
2) Language
3) Regular expression

Strings and Languages

An alphabet or character class is a finite set of symbols.

A string over an alphabet is a finite sequence of symbols drawn from that alphabet.

A language is any countable set of strings over some fixed alphabet.

In language theory, the terms "sentence" and word are often used as synonyms for string. The length of a string s, usually written , is the number of occurrences of symbols in s. For example, banana is a string of length six. The empty string, denoted σ, is the string of length zero.

Operations on strings

The following string-related terms are commonly used:

1. A prefix of string s is any string obtained by removing zero or more symbols from the end of strings. For example, ban is a prefix of banana.

2. A suffix of string s is any string obtained by removing zero or more symbols from the beginning of s. For example, nana is a suffix of banana.

3. A substring of s is obtained by deleting any prefix and any suffix from s. For example, nan is a substring of banana.

4. The proper prefixes, suffixes, and substrings of a string s are those prefixes, suffixes, and substrings, respectively of s that are not σ or not equal to s itself.

5. A subsequence of s is any string formed by deleting zero or more not necessarily consecutive positions of s.

For example, baan is a subsequence of banana.

Operations on languages:

The following are the operations that can be applied to languages:

1.Union

2.Concatenation

3. Kleene closure

4.Positive closure

The following example shows the operations on strings:

1.4.2 Regular Expressions:

Each regular expression r denotes a language L(r).

Here are the rules that define the regular expressions over some alphabet $\sum$ and the languages that those expressions denote:

Let L= {0, 1} and S= {a, b, c}

Union	: L U S={0,1,a,b,c}
Concatenation	: L.S={0a,1a,0b,1b,0c,1c}
Kleene closure	: L*= {ε, 0, 1, 00….}
Positive closure	: L+= {0, 1, 00….}

FIGURE 1.9: operations on strings

1. σ is a regular expression, and L(σ) is σ , that is, the language whose sole member is the empty string.

2. If'a'is a symbol in Σ, then 'a'is a regular expression, and L(a) = a, that is, the language with one string, of length one, with 'a'in its one position.

3. Suppose r and s are regular expressions denoting the languages L(r) and L(s). Then,

o (r)—(s) is a regular expression denoting the language L(r) U L(s). o (r)(s) is a regular expression denoting the language L(r)L(s). o (r)* is a regular expression denoting (L(r))*.

o (r) is a regular expression denoting L(r).

4. The unary operator * has highest precedence and is left associative.

5. Concatenation has second highest precedence and is left associative. has lowest precedence and is left associative.

Regular Definitions: For not ational convenience,we may wish to give names to regular expressions and to define regular expressions using these names as if they were symbols.Identifiers are the setor string of letters and digits beginning with a letter.The following regular definition provides aprecise specification for this class of string.

Shorthand's Certain constructs occur so frequently in regular expressions that it is convenient to introduce notational shorth ands for them.
1. One or more instances (+):

o The unary postfix operator + means " one or more instances of" .

o If r is a regular expression that denotes the language L(r), then (r)+ is a regular expression that denotes the language (L (r))+

o Thus the regular expression a+ denotes the set of all strings of one or more a s.

o The operator + has the same precedence and associatively as the operator *.

2. Zero or one instance (?):

- The unary postfix operator ? means "zero or one instance of".

- The notation r? is a shorthand for r — σ.

- If 'r' is a regular expression, then (r)? is a regular expression that denotes the language L(r) U σ .

3. Character Classes:

- The notation [abc] where a, b and c are alphabet symbols denotes the regular expression a — b — c.
- Character class such as [a – z] denotes the regular expression a — b — c — d — . . .—z. - We can describe identifiers as being strings generated by the regular expression, [A–Za–z][A–Za–z0–9]*
Non-regular Set A language which cannot be described by any regular expression is a non-regular set. Example: The set of all strings of balanced parentheses and repeating strings cannot be described by a regular expression. This set can be specified by a context-free grammar.

Recognition Of Tokens:
Consider the following grammar fragment:
stmt $\rightarrow$ if expr then stmt
—if expr then stmt else stmt —σ
expr $\rightarrow$ term relop term —term
term $\rightarrow$ id —num

where the terminals if , then, else, relop, id and num generate sets of strings given by the following regular definitions:

If	→	if
then	→	then
else	→	else
relo		
	→	<|<=|=|<>|>|>=
id	→	letter(letter | digit)*
num	→	Digit$^+$ (.digit$^+$)? (E(+|-)?digit$^+$)?

FIGURE 1.10: Regular definitions

For this language fragment the lexical analyzer will recognize the keywords if, then, else, as well as the lexemes denoted by relop, id, and num. To simplify matters, we assume keywords are reserved; that is, they cannot be used as identifiers.

Lexeme	Token Name	Attribute Value
Any ws	_	_
if	if	_
then	then	_
else	else	_
Any id	id	pointer to table entry
Any number	number	pointer to table entry
<	relop	LT
<=	relop	LE
=	relop	ET
<>	relop	NE

FIGURE 1.11: language fragment

1.5   Transition Diagram:

Transition Diagram has a collection of nodes or circles, called states. Each state represents a condition that could occur during the process of scanning the input looking for lexeme that matches one of several patterns .Edgesaredirectedfromonestateofthetransitiondiagramtoanother.eachedgeislabeledbyasymbol or set of symbols. If we are in one states, and the next input symbol is a, we look for an edge out of states labeled by a. If we find such an edge, we advance the forward pointer and enter the state of the transition diagram to which that edge leads.

Some important conventions about transition diagrams are

1. Certain states are said to be accepting or final. These states indicate that alexeme has been found; although the actual lexeme may not consist of all positions b/wthelexeme Begin and forward pointers we always indicate an accepting state by a double circle.

2.Inaddition,ifitisnecessarytoreturntheforwardpointeroneposition,thenweshalladditionallyplace a*
near that acceptingstate.

3. Onestate is designed thestate ,orinitialstate ., it isindicated byan edgelabeled "start"entering
from no where.the transition diagram always begins in the state before any input symbols have
been used.

As an intermediate step in the construction of aLA,we first produce a stylized flowchart,called a
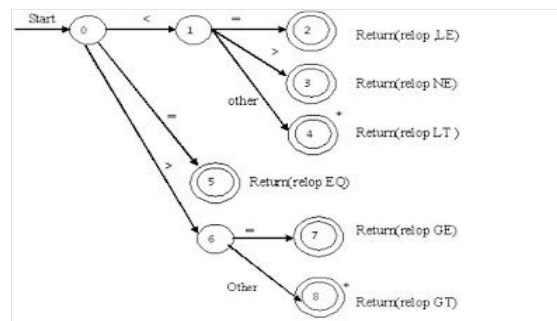
FIGURE 1.12: edge labeled

transition diagram.Positionin a transition diagram,a re drawn as circles and are called asstates.

Thea boveTDforanidentifier,definedtobealetterfollowedbyanynoofletters ordigits.A sequence of tran-
sition diagram can be converted in to program to look for the tokens specified by the diagrams.
Each stategets asegment ofcode.
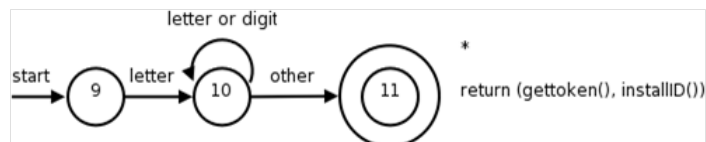
FIGURE 1.13: edge labeled

1.5.1   Automata:

Automationisdefinedasasystemwhereinformationistransmittedandusedfor
performingsomefunctionswithout direct participation of man.

1. Anautomationinwhichtheoutputdependsonlyontheinputiscalled automationwithoutmemory.

2. Anautomationinwhichtheoutputdependsontheinputandstatealsoiscalledas automationwithmem-
ory.

3. AnautomationinwhichtheoutputdependsonlyonthestateofthemachineiscalledaMooremachine.

4. Anautomationinwhichtheoutputdependsonthestateandinputatanyinstantof time iscalledamealy-
machine.

Description of Automata

1. An automata hasa mechanism to read input from input tape,

2. Anylanguageisrecognizedbysomeautomation,Hencetheseautomationare basicallylanguage 'acceptors' or'languagerecognizers'.

TypesofFiniteAutomata

Deterministic Automata

Non-Deterministic Automata.

DeterministicAutomata:

Adeterministicfiniteautomatahasatmostonetransitionfromeachstateonany input. ADFA is aspecialcaseofa NFA in which:-

1. it has no transitionson input ,

2. Each input symbol hasat most one transition from anystate.

Regular expression → NFA → DFA → Minimized DFA

The Finite Automata is called DFA if there is only one path for a specific input from current state to next state.From state So for input' a' there is only one path going toS2.similarly from so there is only one path for input going to S1.
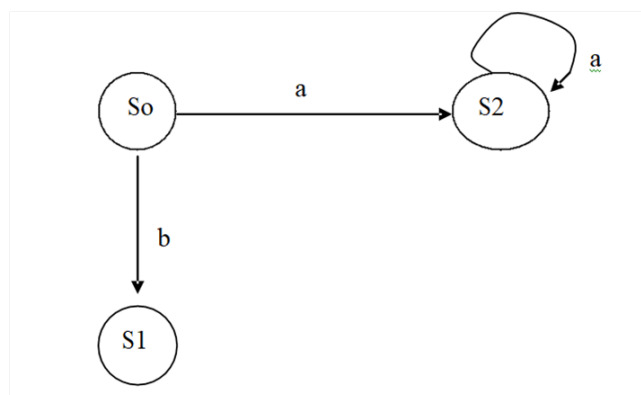
FIGURE 1.14: Minimized DFA

1.5.2 Non deterministic Automata:

1. A NFA ia A mathematical model consists of
2. Aset ofstatesS.
3. Aset of input symbols $\sum$.
4. Atransition is a movefrom one state to another.
5. Astate so that is distinguished as the start (orinitial)state
6. Aset ofstatesFdistinguished as accepting(orfinal)state.
7. A number of transition to a single symbol.

A NFA can be diagrammatically represented by a labeled directed graph,called a transition graph,in which the nodes are the statesand the labeled edges represent the transition function.

This graph looks like a transition diagram,but the same character can label two or more transitions out of one state and edges can be labeled by the special symbol € as well as input symbols.

The transition graph for an NFA that recognizes the language (a—b)*abb is shown
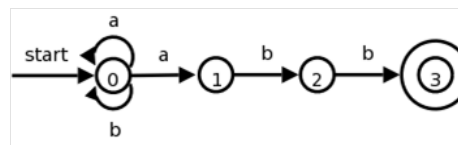
FIGURE 1.15: Transition graph

1.6 Boot strapping:

When a computer is first turned on or restarted, a special type of absolute loader, called as Boot strap loader is executed. This bootstrap loads the first program to be run by the computer usually an operating system. The bootstrap itself begins at address O in the memory of the machine. It loads the operating system (or some other program) starting at address 80. After all of the object code from device has been loaded, the bootstrap program jumps to address 80,which begins the execution of the program that was loaded.

Such loaders can be used to run stand-alone programs independent of the operating system or the system loader. They can also be used to load the operating system or the loader itself into memory.

Loaders are of two types:
• Linking loader.

• Linkage editor.

Linkage loaders, perform all linking and relocation at load time.

Linkage editors, perform linking prior to load time and dynamic linking, in which the linking function is performed at execution time.

A linkage editor performs linking and some relocation; however, the linkaged program is written to a file or library instead of being immediately loaded into memory. This approach reduces the overhead when the program is executed. All that is required at load time is a very simple form of relocation.
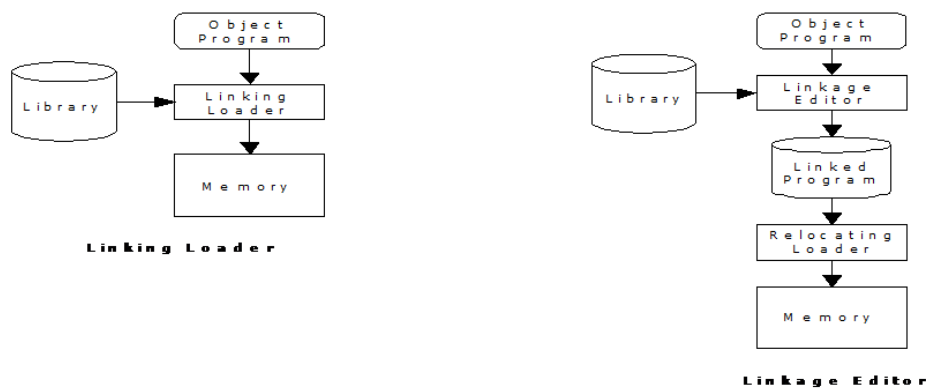
FIGURE 1.16: linkaged program

1.7   Pass and Phases Of Translation

Phases: (Phases are collected into a front end and back end)

Frontend: The front end consists of those phases, or parts of phase, that depends primarily on the source language and is largely independent of the target machine. These normally include lexical and syntactic analysis,the creation of the symbol table,semantic analysis, and the generation of intermediate code. A certain amount of code optimization can be done by front end as well. the front end also includes the error handling tha goes along with each of these phases.

Back end: The back end includes those portions of the compiler that depend on the target machine andgenerally, these portions do not depend on the source language .

1.8 Lexical Analyzer Generator:

Creating a lexical analyzer with Lex:

• First, a specification of a lexical analyzer is prepared by creating a program lex.l in the Lex language. Then, lex.l is run through the Lex compiler to produce a C program lex.yy.c.

• Finally, lex.yy.c is run through the C compiler to produce an object program a.out, which is the lexical analyzer that transforms an input stream into a sequence of tokens

Lex Specification

A Lex program consists of three parts:

{ definitions}
{ rules}
{ user subroutines }

1. Definitions include declarations of variables, constants, and regular definitions

2. Rules are statements of the formp1 {action1}p2 {action2} ... pn {action}

3. where pi is regular expression and actioni describes what action the lexical analyzer

should take when pattern pi matches a lexeme. Actions are written in C code.

4. User subroutines are auxiliary procedures needed by the actions. These can be

compiledseparately and loaded with the lexical analyzer.

1.9 Input Buffering

The LA scans the characters of the source program one at a time to discover tokens. Becauseoflargeamountoftimecanbeconsumedscanningcharacters,specializedbuffering techniqueshavebeendevelopedtoreducetheamountofoverheadrequiredtoprocessan input character.

Bufferingtechniques:

1.Buffer pairs

2. Sentinels

Thelexical analyzerscans the charactersof thesourceprogram onea t a timeto discover tokens. Often, however,many characters beyondthe next token manyhave to be examined beforethenexttokenitselfcanbedetermined.Forthisandotherreasons,itisdesirableforthe lexicalanalyzertoreaditsinputfromaninputbuffer.Figureshowsabufferdividedinto two halves of, say100characters each. One pointer marks thebeginningof the token being discovered. A lookaheadpointerscans ahead of the beginningpoint, until the token is discovered.weviewthepositionofeachpointerasbeingbetweenthecharacterlastreada

thecharacternexttoberead.Inpracticeeachbufferingschemeadoptsoneconventioneither apointer isat the symbollast read or thesymbol itisreadyto read.

Chapter 2

SYNTAX ANALYSIS

Course Outcomes

After successful completion of this module, students should be able to:

CO 4	**Construct** the derivations, FIRST set, FOLLOW set on the context free grammar for performing the top-down and bottom up parsing methods.	Apply
CO 5	**Distinguish** top down and bottom up parsing methods for developing parser with the parse tree representation of the input.	Analyze
CO 6	**Construct** LEX and YACC tools for developing a scanner and a parser.	Apply

2.1   SYNTAX ANALYSIS

Syntax Analysis: Parsing, role of parser, context free grammar, derivations, parse trees, ambiguity, elimination of left recursion, left factoring, eliminating ambiguity from dangling-else grammar; Types of parsing: Top-down parsing, backtracking, recursive-descent parsing, predictive parsers, LL (1) grammars. Bottom-up parsing: Definition of bottom-up parsing, handles, handle pruning, stack implementation of shift- reduce parsing, conflicts during shift-reduce parsing, LR grammars, LR parsers-simple LR, canonical LR and Look Ahead LR parsers, error recovery in parsing, parsing ambiguous grammars, YACC-automatic parser generator.

2.2   SYNTAX ANALYSIS :

Top-down Parsing
Context-free Grammars: Definition:

Formally, a context-free grammar G is a 4-tuple G = (V, T, P, S), where:

1. V is a finite set of variables (or non terminals). These describe sets of "related" strings.

2. T is a finite set of terminals (i.e., tokens).

3. P is a finite set of productions, each of the form A a where AÎ V is a variable, and aÎ (V È T)* is a sequence of terminals and nonterminals. S Î V is the start symbol.

Example of CFG: E == > EAE — (E) — -E — id

A== > + — - — * — / —

Where E,A are the non-terminals while id, +, *, -, /,(,) are the terminals.

Syntax analysis: In syntax analysis phase the source program is analyzed to check whether if conforms to the source language's syntax, and to determine its phase structure. This phase is often separated into two phases:

1. Lexical analysis: which produces a stream of tokens?

2. Parser: which determines the phrase structure of the program based on the context-free grammar for the language?

Parsing:

Parsing is the activity of checking whether a string of symbols is in the language of some grammar, where this string is usually the stream of tokens produced by the lexical analyzer. If the string is in the grammar, we want a parse tree, and if it is not, we hope for some kind of error message explaining why not.

There are two main kinds of parsers in use, named for the way they build the parse trees:

1. Top-down: A top-down parser attempts to construct a tree from the root, applying productions forward to expand non-terminals into strings of symbols.

2. Bottom-up: A Bottom-up parser builds the tree starting with the leaves, using productions in reverse to identify strings of symbols that can be grouped together.

In both cases the construction of derivation is directed by scanning the input sequence from left to right, one symbol at a time.

Parse Tree:

A parse tree is the graphical representation of the structure of a sentence according to its grammar.

 Example:

Let the production P is:

E → T | E+T

T → F | T*F
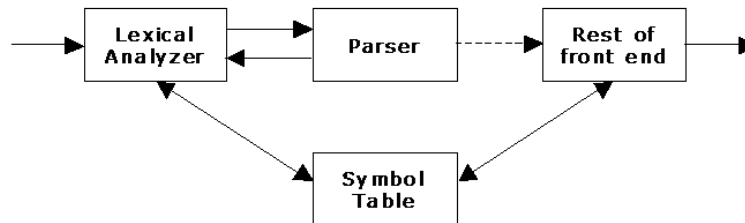
FIGURE 2.1: Parse Tree

F → V | (E)

V → a | b | c | d |

The parse tree may be viewed as a representation for a derivation that filters out the choice regarding the order of replacement.

Parse tree for a * b + c
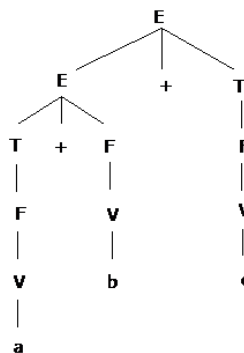
FIGURE 2.2: Parse Tree

Parse tree for a + b * c is:

Parse tree for (a * b) * (c + d)
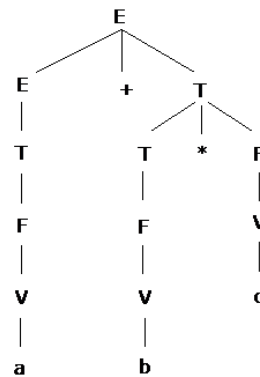
FIGURE 2.3: Parse Tree
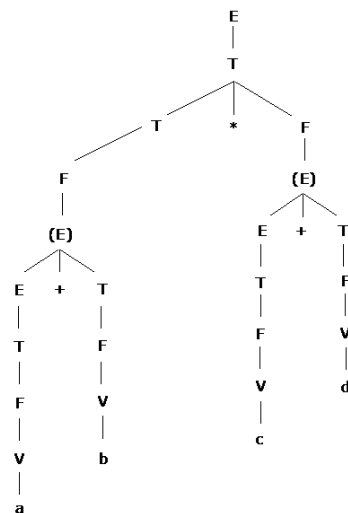
FIGURE 2.4: Parse Tree

2.3 Syntax Trees:

Parse tree can be presented in a simplified form with only the relevant structure information by:

1. Leaving out chains of derivations (whose sole purpose is to give operators difference precedence).

2. Labeling the nodes with the operators in question rather than a non-terminal. The simplified Parse tree is sometimes called as structural tree or syntax tree.

a * b + c   a + b * c   (a + b) * (c + d)
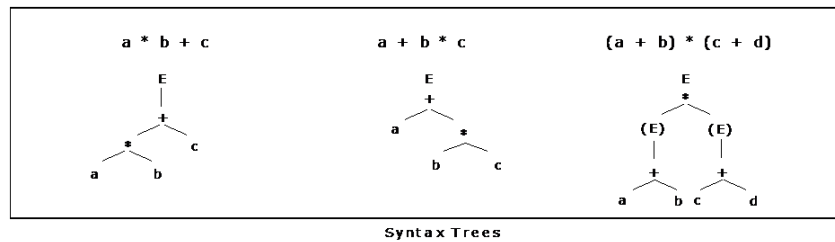
FIGURE 2.5: Syntax Tree

2.3.1 Syntax Error Handling:

If a compiler had to process only correct programs, its design and implementation would be greatly simplified. But programmers frequently write incorrect programs, and a good compiler should assist the programmer in identifying and locating errors.The programs contain errors at many different levels.

For example, errors can be:

1) Lexical – such as misspelling an identifier, keyword or operator

2) Syntactic – such as an arithmetic expression with un-balanced parentheses.

3) Semantic – such as an operator applied to an incompatible operand.

4) Logical – such as an infinitely recursive call.

Much of error detection and recovery in a compiler is centered around the syntax analysis phase. The goals of error handler in a parser are:

1. It should report the presence of errors clearly and accurately.

2. It should recover from each error quickly enough to be able to detect subsequent errors.

3. It should not significantly slow down the processing of correct programs.

2.3.2 Ambiguity:

Several derivations will generate the same sentence, perhaps by applying the same productions in a different order. This alone is fine, but a problem arises if the same sentence has two distinct parse trees.A grammar is ambiguous if there is any sentence with more than one parse tree.

Any parses for an ambiguous grammar has to choose somehow which tree to return. There are a number of solutions to this; the parser could pick one arbitrarily, or we can provide some hints about which to choose. Best of all is to rewrite the grammar so that it is not ambiguous. There is no general method for removing ambiguity.Ambiguity is acceptable in spoken languages. Ambiguous programming languages are useless unless the ambiguity can be resolved.

Fixing some simple ambiguities in a grammar:Ambiguous language unambiguous

(i) A → B | AA Lists of one or more B's A → BC C → A | E

(ii) A → B | A;A Lists of one or more B's with punctuation A → BC C → ;A | E

(iii) A → B | AA | E lists of zero or more B's A → BA | E

Any sentence with more than two variables, such as (arg, arg, arg) will have multiple parse trees.

2.3.3 Left Recursion:

If there is any non terminal A, such that there is a derivation A $\Longrightarrow$ A α for some string α,

then the grammar is left recursive.

Algorithm for eliminating left Recursion:

1. Group all the A productions together like this:

A $\Longrightarrow$ A α 1 | A α 2 | - - - | A α m | β | β2 | - - - | β n

Where,

A is the left recursive non-terminal,

α is any string of terminals and

β is any string of terminals and non terminals that does not begin with A.

2. Replace the above A productions by the following:

A $\Longrightarrow$ β 1 AI | β2 AI | - - - | βn AI

AI $\Longrightarrow$ α 1 AI | α AI | - - - | α m AI | $\in$

Where, AI is a new non terminal.

Top down parsers cannot handle left recursive grammars.

If our expression grammar is left recursive:

1. This can lead to non termination in a top-down parser.

2. for a top-down parser, any recursion must be right recursion.

3. we would like to convert the left recursion to right recursion.

Example 2:

Remove the left recursionfrom the productions:

E → E + T | T

T → T * F | F

Applying the transformation yields:

E → T EI

T → F TI

EI →T EI | $\in$

TI → * F TI | $\in$

Example 3:

Remove the left recursion from the productions:

E → E + T | E – T | T

T → T * F | T/F | F

Applying the transformation yields:

E → T EI

T → F TI

E → + T EI | - T EI | ∈

TI → * F TI | /F TI | ∈

Example 4:

Remove the left recursion from the productions:

S → A a | b

A → A c | S d | ∈

1.The non terminal S is left recursive because S → A a → S d a

But it is not immediate left recursive.

2.Substitute S-productions in A → S d to obtain:

A → A c | A a d | b d | ∈

3.Eliminating the immediate left recursion:

S → A a | b

A → b d AI | AI

AI → c AI | a d AI | ∈

2.3.4 Left Factoring:

Left factoring is a grammar transformation that is useful for producing a grammar suitable for predictive parsing. When it is not clear which of two alternative productions to use to expand a non-terminal A, we may be able to rewrite the productions to defer the decision until we have some enough of the input to make the right choice.

Algorithm:

For all A ∈ non-terminal, find the longest prefix α that occurs in two or more right-hand sides of A.

If α

neq ∈then replace all of the A productions,

A → $\alpha \beta$I | $\alpha \beta$ 2 | - - - | $\alpha \beta$ n| r With

A → α AI | r

AI → βI | β 2 | - - - | β n | ∈ Where, AI is a new element of non-terminal. Repeat until no common prefixes remain. It is easy to remove common prefixes by left factoring, creating new non-terminal.

2.3.5 Top Down Parsing:

Top down parsing is the construction of a Parse tree by starting at start symbol and "guessing" each derivation until we reach a string that matches input. That is, construct tree from root to leaves.The

advantage of top down parsing in that a parser can directly be written as a program.Table-driven top-down parsers are of minor practical relevance. Since bottom-up parsers are more powerful than top-down parsers, bottom-up parsing is practically relevant. For example, let us consider the grammar to see how top-down parser works:

S → if E then S else S | while E do S | print

E → true | False | id

The input token string is: If id then while true does print else print.

1. Tree: S

Input: if id then while true do print else print.

Action: Guess for S.

2. Tree:
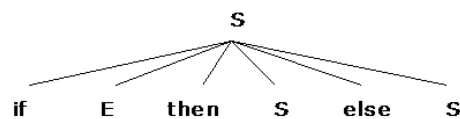
FIGURE 2.6: Tree

Input: if id then while true do print else print. Action: if matches; guess for E.

3.Tree:
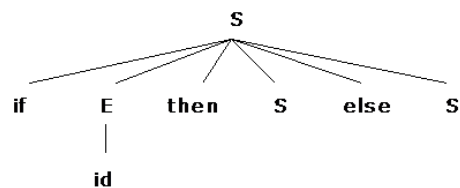
FIGURE 2.7: Tree

Input: id then while true do print else print. Action: id matches; then matches; guess for S.

4. Tree:
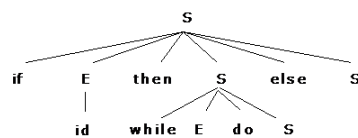
FIGURE 2.8: Tree

Input: while true do print else print. Action: while matches; guess for E.

5. Tree:
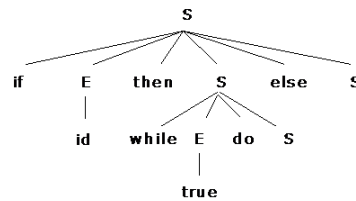
FIGURE 2.9: Tree

Input: true do print else print Action: true matches; do matches; guess S.

6. Tree:
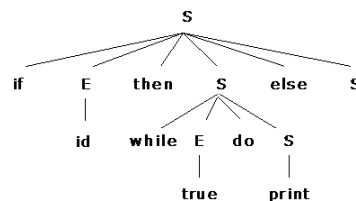
FIGURE 2.10: Tree

2.3.6 Recursive Descent Parsing:

Top-down parsing can be viewed as an attempt to find a left most derivation for an input string. Equivalently, it can be viewd as a attempt to construct a parse tree for the input starting from the root and creating the nodes of the parse tree in preorder.

The special case of recursive –decent parsing, called predictive parsing, where no backtracking is required. The general form of top-down parsing, called recursive descent, that may involve backtracking, thatis, makingrepeated scans of the input. Recursive descent or predictive parsing works only on grammars where the first terminal symbol of each sub expression provides enough information to choose which production to use.

Recursive descent parser is a top down parser involving backtracking. It makes a repeated scans of the input. Backtracking parsers are not seen frequently, as backtracking is very needed to parse programming language constructs. Example: consider the grammar

S → cAd

A → ab | a

And the input string w=cad. To construct a parse tree for this string top-down, weinitially create a tree consisting of a single node labeledscan input pointer points toc,the first symbol of w.we then

use the first production for S to expand tree and obtain the tree of Fig(a).
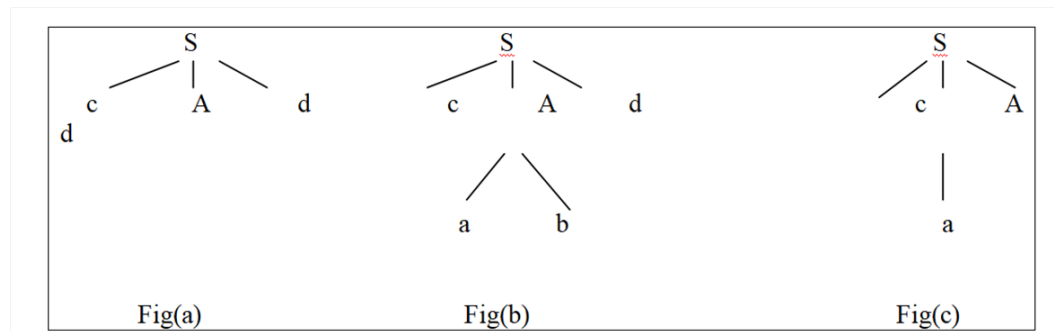
FIGURE 2.11: Tree

The left most leaf, labeled c,matches the first symbol of w, so we now advance the input pointer to a ,the second symbol of w, and consider the next leaf,labeled A.We can then expand A using the first alternative for A to obtain the tree in Fig (b). we now have a match for the second input symbol so we advance the input pointer to d,the third, input symbol,and compare d against the next leaf,labeled b.since b does not match the d ,we report failure and go back to A to see where there is any alternative for Ac that we have not tried but that might produce a match

In going back to A,we must reset the input pointer to position2,we now try second alternative for A to obtain the tree of Fig(c).The leaf matches second symbol of w and the leaf d matches the third symbol .

The left recursive grammar can cause a recursive- descent parser, even one with backtracking, to go into an infinite loop.That is ,when we try to expand A,we may eventually find ourselves again trying to ecpand A without Having consumed any input.

2.3.7 Predictive Parsing:

Predictive parsing is top-down parsing without backtracking or look ahead. For many languages, make perfect guesses (avoid backtracking) by using 1-symbol look-a-head. i.e., if:

$A \rightarrow \alpha 1 \mid \alpha 2 \mid - - - \mid \alpha n$.

Choose correct α i by looking at first symbol it derive. If $\in$ is an alternative, choose it last.

This approach is also called as predictive parsing. There must be at most one production in order to avoid backtracking. If there is no such production then no parse tree exists and an error is returned. The crucial property is that, the grammar must not be left-recursive. Predictive parsing works well on those fragments of programming languages in which keywords occurs frequently.

For example:

stmt $\rightarrow$ if exp then stmt else stmt

| while expr do stmt

| begin stmt-list end.

then the keywords if, while and begin tell, which alternative is the only one that could possibly succeed if we are to find a statement.

The model of predictive parseris as follows:
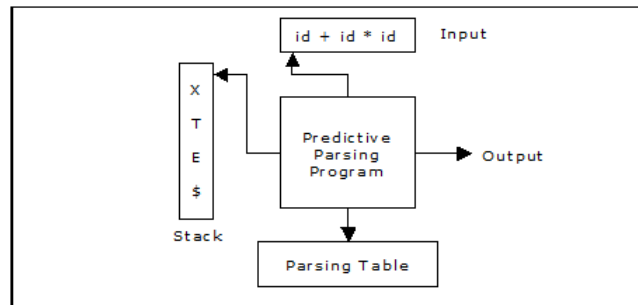
FIGURE 2.12: The model of predictive parseris

A predictive parser has:

1. Stack

2. Input

3. Parsing Table

4. Output

The input buffer consists the string to be parsed, followed by $, a symbol used as a right end marker to indicate the end of the input string. The stack consists of a sequence of grammar symbols with $ on the bottom, indicating the bottom of the stack. Initially the stack consists of the start symbol of the grammar on the top of $. Recursive descent and LL parsers are often called predictive parsers, because they operate by predicting the next step in a derivation.

The algorithm for the Predictive Parser Program is as follows:

Input:A string w and a parsing table M for grammar G

Output:if w is in L(g),a leftmost derivation of w;otherwise, an error indication.

Method: Initially, the parser has $ S on the stack with S, the start symbol of G on top, and w $ in the input buffer. The program that utilizes the predictive parsing table M to produce a parse for the input is: Set ip to point to the first symbol of w $; repeat

let x be the top stack symbol and a the symbol pointed to by ip;

if X is a terminal or $ then

if X = a then

pop X from the stack and advance ip

else error()

else /* X is a non-terminal */

if M[X, a] = X → Y1 Y2 . . . . Yk then begin

pop X from the stack;

push Yk, Yk-1, ...Y1 onto the stack, with Y1 on top;

output the production X → Y1 Y2 . . . . . Yk

end

else error()

until X = $ /*stack is empty*

FIRST and FOLLOW:

The construction of a predictive parser is aided by two functions with a grammar G. these functions,FIRST and FOLLOW, allow us to fill in the entries of a predictive parsing table for G, whenever possible.Sets of tokens yielded by the FOLLOW function can also be used as synchronizing tokens during pannic-mode error recovery. If α is any string of grammar symbols,

letFIRST (α) be the set of terminals that begin the strings derived from

α.If $\alpha => \xi$,then ξ is also in FIRST(α).

Define FOLLOW (A),for nonterminalsA,to be the set of terminals a that can appear immediately to the right of A in some sentential form, that is,the set of terminals a such that there exist a derivation of the form

S=> αAaβ for some α and β.

If A can be the rightmost symbol in some sentential form,then $ is in FOLLOW(A).

Computation of FIRST (): To compute FIRST(X) for all grammar symbols X,apply the following rules until no more terminals or ξ can be added to any FIRST set.

1. If X is terminal,then FIRST(X) is X.

2. If X $\rightarrow \xi$ is production, then add ξ to FIRST(X).

3. If X is nonterminal and X $\rightarrow$ Y1 Y2......Yk is a production,then place a in FIRST(X) if for some i,a is in FIRST(Yi),and ξ is in all of FIRST(Yi),and ξ is in all of FIRST(Y1),.....FIRST(Yi-1);that is Y1.........Yi-1==> ξ.if ξ is in FIRST(Yj), for all j=,2,3.......k,then add ξ to FIRST(X).for example, everything in FIRST(Y1) is surely in FIRST(X).if Y1 does not derive ξ,then we add nothing more to FIRST(X),but if Y1=> ξ,then we add FIRST(Y2) and so on.

FIRST (A) = FIRST (α1) U FIRST (α2) U - - - U FIRST (αn)

Where, A $\rightarrow \alpha$1 | α 2 | - - - | α n, are all the productions for A.

FIRST (A α) = if $\in$ $\in$ FIRST (A) then FIRST (A)

else (FIRST (A) - $\in$) U FIRST (α)

Computation of FOLLOW ():

To compute FOLLOW (A) for all nonterminals A, apply the following rules until nothing can be added to any FOLLOW set.

1 Place $ in FOLLOW(s), where S is the start symbol and $ is input right end marker .

2 If there is a production A$\rightarrow \rightarrow \alpha$B , then everything in FIRST(β) except for ξ is placed in FOLLOW(B).

3 If there is production A$\rightarrow \rightarrow \alpha$B, or a production A$\rightarrow \alpha$B β where FIRST (β) contains ξ(i.e.,β $\rightarrow \xi$),then everything in FOLLOW(A)is in FOLLOW(B).

2.4 Bottom Up Parsing

Bottom-up parser builds a derivation by working from the input sentence back towards the start symbol S. Right most derivation in reverse order is done in bottom-up parsing.

(The point of parsing is to construct a derivation. A derivation consists of a series of rewrite steps)

$S \Rightarrow r0 \Rightarrow r1 \Rightarrow r2 \Rightarrow - - - \Rightarrow rn\text{-}1 \Rightarrow rn \Rightarrow$ sentence

Bottom-up

Assuming the production $A \Rightarrow \beta$, to reduce ri ri-1 match some RHS β against ri then replace β with its corresponding LHS, A.

In terms of the parse tree, this is working from leaves to root.

Top down Vs Bottom-up parsing:

Top-down	Bottom-up
1. Construct tree from root to leaves	1. Construct tree from leaves to root
2. "Guers" which RHS to substitute for non terminal.	2. "Guers" which rule to "reduce" terminals
3. Produces left-most derivation	3. Produces reverse right-most derivation.
4. Recursive descent, LL parsers	4. Shift-reduce, LR, LALR, etc.
5. Recursive descent, LL parsers	5. "Harder" for humans.
6. Easy for humans	

FIGURE 2.13: Top down Vs Bottom-up parsing

$\rightarrow$ Bottom-up can parse a larger set of languages than top down.

$\rightarrow$ Both work for most (but not all) features of most computer languages

2.4.1 Handles:

Always making progress by replacing a substring with LHS of a matching production will not lead to the goal/start symbol.

For example:

abbcde

aAbcde A $\Rightarrow$ b

aAAcde A $\Rightarrow$ b

struck

Informally, A Handle of a string is a substring that matches the right side of a production, and whose reduction to the non-terminal on the left side of the production represents one step along the reverse of a right most derivation.

If the grammar is unambiguous, every right sentential form has exactly one handle.

More formally, A handle is a production A⇒B and a position in the current right-sentential form $\alpha\ \beta\ \omega$ such that:

$S \Rightarrow \alpha\ A\ \omega \Rightarrow \alpha\ /\ \beta\ \omega$ For example grammar, if current right-sentential form is

a/Abcde

Then the handle is A → Ab at the marked position. 'a' never contains non-terminals.

2.4.2 Handle Pruning:

Keep removing handles, replacing them with corresponding LHS of production, until we reach S. Example:

E→E+E/E*E/(E)/id

Right-sentential form	Handle	Reducing production
a+b*c	a	E→id
E+b*c	b	E→id
E+E*C	C	E→id
E+E*E	E*E	E→E*E
E+E	E+E	E→E+E
E		

FIGURE 2.14: Handle Pruning example

The grammar is ambiguous, so there are actually two handles at next-to-last step. We can use parser-generators that compute the handles for us.

2.4.3 Shift- Reduce Parsing:

Shift Reduce Parsing uses a stuck to hold grammar symbols and input buffer to hold string to be parsed, because handles always appear at the top of the stack i.e., there's no need to look deeper into the state.

A shift-reduce parser has just four actions:

1. Shift-next word is shifted onto the stack (input symbols) until a handle is formed.

2. Reduce – right end of handle is at top of stack, locate left end of handle within the stack. Pop handle off stack and push appropriate LHS.

3. Accept – stop parsing on successful completion of parse and report success.

4. Error – call an error reporting/recovery routine.

Possible Conflicts:

Ambiguous grammars lead to parsing conflicts.

1. Shift-reduce: Both a shift action and a reduce action are possible in the same state (should we shift or reduce)

Example: dangling-else problem

2. Reduce-reduce: Two or more distinct reduce actions are possible in the same state. (Which production should we reduce with 2)

2.4.4   Operator – Precedence Parsing:

Precedence/ Operator grammar: The grammars having the property:

1. No production right side is should contain $\in$.

2. No production sight side should contain two adjacent non-terminals. Is called an operator grammar.

Operator – precedence parsing has three disjoint precedence relations, $< .,=$ and $.>$ between certain pairs of terminals. These precedence relations guide the selection of handles and have the following meanings:

RELATION	MEANING
a <.b	'a' yields precedence to 'b'.
a=b	'a' has the same precedence 'b'
a.>b	'a' takes precedence over 'b'.

FIGURE 2.15: meanings

Operator precedence parsing has a number of **disadvantages:**

1. It is hard to handle tokens like the minus sign, which has two different precedences.

2. Only a small class of grammars can be parsed.

3. The relationship between a grammar for the language being parsed and the operator-precedence parser itself is tenuous, one cannot always be sure the parser accepts exactly the desired language.

Disadvantages:

1. L(G) $\neq$ L(parser)

2. error detection

3. usage is limited

4. They are easy to analyse manually

Solution: This is not operator grammar, so first reduce it to operator grammar form, by eliminating adjacent non-terminals.

Example:
Grammar: E→EAE|(E)|-E/id

A→+|-|*|/|↑

Input string: id+id*id

The operator – precedence relations are:

	Id	+	*	$
Id		.>	.>	.>
+	<.	.>	<.	.>
*	<.	.>	.>	.>
$	<.	<.	<.	

FIGURE 2.16: example

2.5   LR Parsing Introduction:

The"L"is for left-to-right scanning of the input and the "R" is for constructing a right most derivation in reverse.
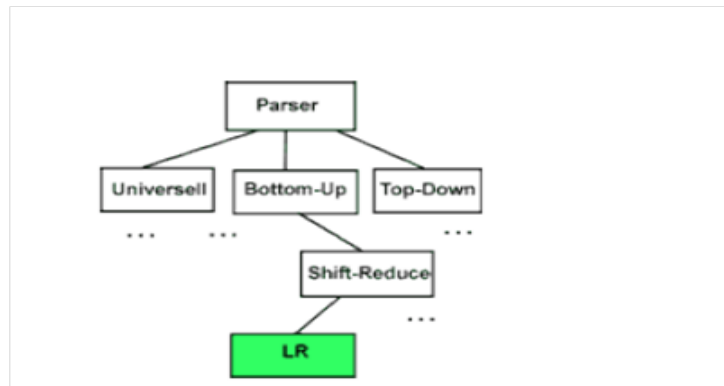
FIGURE 2.17: Parsing

Why LR Parsing:
1. LRparserscanbeconstructedtorecognizevirtuallyallprogramming-language constructs for which context-free grammars can be written.
2. TheLRparsingmethodisthemostgeneralnon-backtrackingshift-reduceparsing method known, yet it can be implement edas efficiently as other shift-reduce methods.
3. TheclassofgrammarsthatcanbeparsedusingLRmethodsisapropersubsetofthe class of grammars that can be parsed with predictive parsers.
4. An LR parser can detect a syntactic error as soon as it is possible to do soon a left-to-right scan of the input. The disadvantage is that it takes too much work to construct an LR parser by hand for a typical programming-language grammar

LR Parsers:

LR(k) parsers are most general non-backtracking shift-reduce parsers. Two cases of interest are k=0 and k=1. LR(1) is of practical relevance.

'L' stands for "Left-to-right" scan of input.

'R' stands for "Rightmost derivation (in reverse)".

'K' stands for number of input symbols of look-a-head that are used in making parsing decisions. When (K) is omitted, 'K' is assumed to be 1.

LR(1) parsers are table-driven, shift-reduce parsers that use a limited right context (1 token) for handle recognition.

LR(1) parsers recognize languages that have an LR(1) grammar. A grammar is LR(1) if, given a right-most derivation

$S \Rightarrow r0 \Rightarrow r1 \Rightarrow r2 - - - rn-1 \Rightarrow rn \Rightarrow$ sentence.

We can isolate the handle of each right-sentential form ri and determine the production by which to reduce, by scanning ri from left-to-right, going atmost 1 symbol beyond the right end of the handle of ri.

Parser accepts input when stack contains only the start symbol and no remaining input symbol are left.

LR(0) item: (no lookahead)

Grammar rule combined with a dot that indicates a position in its RHS.
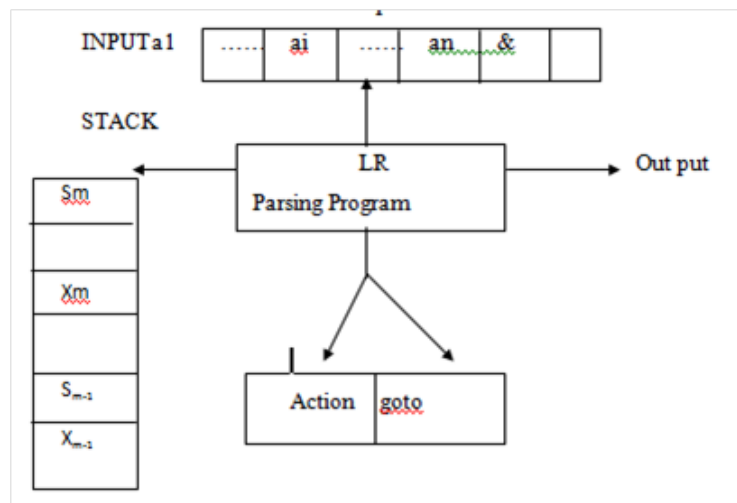
FIGURE 2.18: The schematic form of an LR parser is shown below:

It consists of an input, an output, a stack, a driver program, and a parsing table that has two parts: action and goto.The LR parser program determines Sm, the current state on the top of the stack, and ai, the current input symbol. It then consults action [Sm, ai], which can have one of four values:

1. Shift S, where S is a state.

2. reduce by a grammar production A $\rightarrow \beta$

3. accept and

4. error

The function goes to takes a state and grammar symbol as arguments and produces a state. The go to function of a parsing table constructed from a grammar G using the SLR, canonical LR or LALR method is the transition function of DFA that recognizes the viable prefixes of G. (Viable prefixes of G are those prefixes of right-sentential forms that can appear on the stack of a shift-reduce parser, because they do not extend past the right-most handle).

2.5.1   Augmented Grammar:

If G is a grammar with start symbol S, then GI, the augmented grammar for G with a new start symbol SI and production SI → S.

The purpose of this new start stating production is to indicate to the parser when it should stop parsing and announce acceptance of the input i.e., acceptance occurs when and only when the parser is about to reduce by SI → S.

Construction of SLR Parsing Table:

Example:

The given grammar is:

1. E→E+T

2. E→ T

3. T →T*F

4. T→F

5. F→(E)

6. F→id

Step I: The Augmented grammar is:

EI→E

E→E+T

E→T

T→T*F

T→F

F→(E)

F→id

Step II: The collection of LR (0) items are:

Io: EI→.E

E→.E+T

E→.T

T→.T*F

T→.F

F→.(E)

F→.id

Start with start symbol after since ( ) there is E, start writing all productions of E.

Start writing 'T' productions Start writing F productions

Go to (Io,E): States have successor states formed by advancing the marker over the symbol it pre-ceeds. For state 1 there are successor states reached by advancing the masks over the symbols E,T,F,C or id. Consider, first, the

I1: E I→E. - reduced Item (RI)

E→E.+T

Goto (I0,T):

I2: E→T. - reduced Item (RI)

T→T.*F

Go to (I0,F):

I2: E→T. - reduced item (RI)

T→T.*F

Go to (I0,C):

I4: F→(.E)

E→.E+T

E→.T

T→.T*F

T→.F F→.(E) F→.id If '.' Precedes non-terminal start writing its corresponding production. Here first E then T after that F.

Start writing F productions.

Goto (Io,id):

I5: F →id. - reduced item.

E successor (I, state), it contains two items derived from state 1 and the closure operation adds no more (since neither marker precedes a non-terminal). The state I2 is thus:

Goto (I1,+):

I6: E→E+.T start writing T productions

T→.T*F

T→.F start writing F productions

F→.(E)

F→.id

Goto (I2,*):

I7: T→T*.F start writing F productions

F→.(E)

F→.id

Goto (I4,E):

I8: F→(E.)

E→E.+T

Goto (I4,T):

I2: E→T. these are same as I2.

T→T.*F

Goto (I4,C):

I4: F→(.E)

E→.E+T

E→.T

T→.T*F

T→.F

F→.(E)

F→.id

goto (I4,id):

I5: F→id. - reduced item

Goto (I6,T): I9: E→E+T. - reduced item

T→T.*F Goto (I6,F):

I3: T→F. - reduced item

Goto (I6,C):

I4: F→(.E)

E→.E+T

E→.T

T→.T*F

T→.F

F→.(E)

F→.id

Goto (I6,id):

I5: F→id. - reduced item.

Goto (I7,F):

I10: T→T*F - reduced item

Goto (I7,C):

I4: F→(.E)

E→.E+T

E→.T

T→.T*F

T→.F

F→.(E)

F→.id

Goto (I7,id):

I5: F→id. - reduced item

Action									Goto
State	Id	+	*	(	)	$	E	T	F
I0	S5			S4			1	2	3
1		S6				Accept			
2		r2	S7		R2	R2			
3		R 4	R 4		R4	R4			
4	S5			S4			8	2	3
5		R 6	R 6		R6	R6			
6	S5			S4				9	3
7	S5			S4					10
8		S 6			S11				
9		R1	S7		r1	r1			
10		R3	R3		R3	R3			
11		R5	R5		R5	R5			

Goto (I8,)):

I11: F→(E). - reduced item

Goto (I8,+):

I11: F→(E). - reduced item

Goto (I8,+):

I6: E→E+.T

T→.T*F

T→.F

F→.(E)

F→.id

Goto (I9,+):

I7: T→T*.f

F→.(E)

F→.id

Step IV: Construction of Parse table:

Construction must proceed according to the algorithm 4.8

S→shift items

R→reduce items

Initially EI→E. is in I1 so, I = 1.

Set action [I, $] to accept i.e., action [1, $] to Acc

Chapter 3

SYNTAX DIRECTED TRANSLATION AND INTERMEDIATECODE GENERATION

Course Outcomes

After successful completion of this module, students should be able to:

CO 6	**Describe** syntax directed definitions and translations for performing Semantic Analysis.	Apply
CO 7	**Classify** the different intermediate forms for conversion of syntax translations into Intermediate Code.	Understand

3.1   Introduction

Syntax-Directed Translation: Syntax directed definitions, construction of syntax trees, S-attributed and L- attributed definitions; Syntax Directed Translation schemes. Intermediate code generation: Intermediate forms of source programs– abstract syntax tree, polish notation and three address code, types of three address statements and its implementation, syntax directed translation into three-address code, translation of simple statements, Boolean expressions and flow-of- Control statements.

3.1.1 Syntax Directed Translation:

1. A formalist called as syntax directed definition is used fort specifying translations for programming language constructs.

2. A syntax directed definition is a generalization of a context free grammar in which each grammar symbol has associated set of attributes and each and each productions is associated with a set of semantic rules.

Definition of (syntax Directed definition) SDD:

SDD is a generalization of CFG in which each grammar productions X-> α isassociated with it a set of semantic rules of the form

a: = f(b1,b2…..bk)

Where a is an attributes obtained from the function f.

1. A syntax-directed definition is a generalization of a context-free grammar inwhich: Each grammar symbol is associated with a set of attributes.

This set of attributes for a grammar symbol is partitioned into two subsetscalled synthesized and inherited attributes of that grammar symbol.

Each production rule is associated with a set of semantic rules. 2. Semantic rules set up dependencies between attributes which can be representedby a dependency graph.

3. This dependency graph determines the evaluation order of these semantic rules.

4. Evaluation of a semantic rule defines the value of an attribute. But a semantic rulemay also have some side effects such as printing a value.

The two attributes for non terminal are :

1) Synthesized attribute (S-attribute) : ($\uparrow$)

An attribute is said to be synthesized attribute if its value at a parse tree node isdetermined from attribute values at the children of the node 2) Inherited attribute: ($\uparrow,\rightarrow$)

An inherited attribute is one whose value at parse tree node is determined interms of attributes at the parent and — or siblings of that node.

1. The attribute can be string, a number, a type, a, memory location oranything else.

2. The parse tree showing the value of attributes at each node is called anannotated parse tree.

The process of computing the attribute values at the node is called annotatingor decorating the parse tree.Terminals can have synthesized attributes, but not inherited attributes.

Annotated Parse Tree

1. A parse tree showing the values of attributes at each node is called anAnnotated parse tree.

2. The process of computing the attributes values at the nodes is called annotating(or decorating) of the parse tree.

3. Of course, the order of these computations depends on the dependency graphinduced by the semantic rules.

Ex1:1) Synthesized Attributes :

Ex: Consider the CFG :

S$\rightarrow$ EN

E→ E+T

E→E-T

E→ T

T→ T*F

T→T/F

T→F

F→ (E)

F→digit

N→;

Solution:The syntax directed definition can be written for the above grammar by usingsemantic actions for each production.

Production rule	Semantic actions
S →EN	S.val=E.val
E →E1+T	E.val =E1.val + T.val
E →E1-T	E.val = E1.val – T.val
E →T	E.val =T.val
T →T*F	T.val = T.val * F.val
T →T|F	T.val =T.val | F.val
F → (E)	F.val =E.val
T →F	T.val =F.val
F →digit	F.val =digit.lexval
N →;	can be ignored by lexical Analyzer as; I is
terminating symbol	

FIGURE 3.1: syntax directed definition :

For the Non-terminals E,T and F the values can be obtained using the attribute "Val". The taken digit has synthesized attribute "lexval".

In S→EN, symbol S is the start symbol. This rule is to print the final answer of expressed. Following steps are followed to Compute S attributed definition

1. Write the SDD using the appropriate semantic actions for correspondingproduction rule of the given Grammar.

2. The annotated parse tree is generated and attribute values are computed. TheComputation is done in bottom up manner.

3. The value obtained at the node is supposed to be final output.

PROBLEM 1:

Consider the string 5*6+7; Construct Syntax tree, parse tree and annotated tree

Solution:

The corresponding annotated parse tree is shown below for the string 5*6+7;

Annotated parse tree
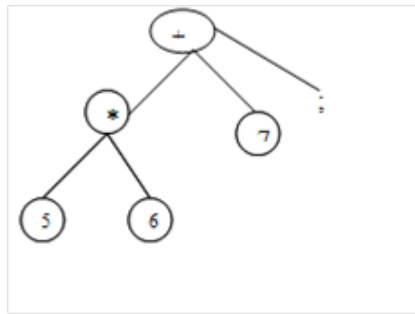
FIGURE 3.2: Syntax tree:
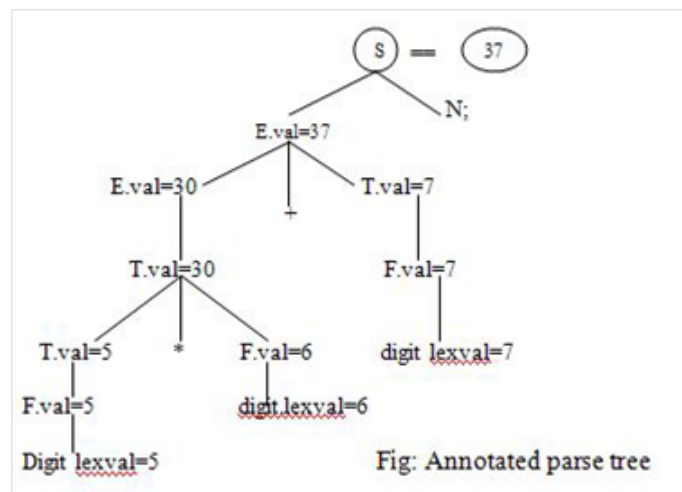
FIGURE 3.3: Annotated parse tree

Advantages: SDDs are more readable and hence useful for specifications

Disadvantages: not very efficient.

PROBLEM 2:

Consider the grammar that is used for Simple desk calculator. Obtain the Semantic action and also the annotated parse tree for the string

3*5+4n.

L→En

E→E1+T

E→T

T→T1*F

T→F

F→(E)

F→ digit

Solution:

The corresponding annotated parse tree U shown below, for the string 3*5+4n

Production rule	Semantic actions
L→ En	L.val=E.val
E→E1+T	E.val=E1.val + T.val
E→T	E.val=T.val
T→T1*F	T.val=T1.val*F. Val
T→F	T.val=F.val
F→ (E)	F.val=E.val
F→ digit	F.val=digit.lexval

FIGURE 3.4: solution
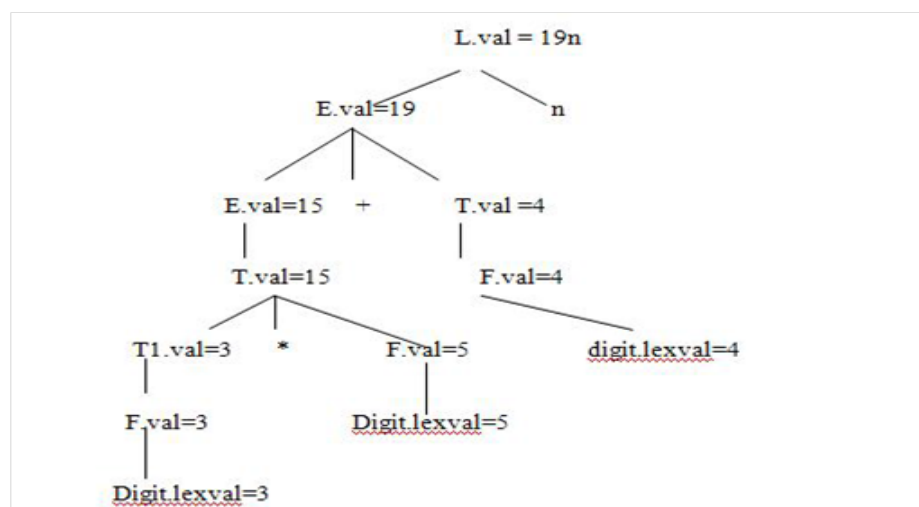
FIGURE 3.5: solution

Dependency Graphs:

3.1.2   dependency graph and topological sort:

1. For each parse tree node, say a node labeled by grammar symbol x, the dependency graph has a node for each attribute associated with x

2. If a semantic rule associated with a production p defines the value ofsynthesized attribute A.b in terms of the value of X.c. Then thedependency graph has an edge from X.c to A.b
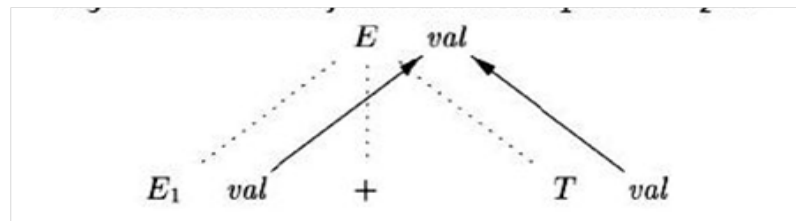
FIGURE 3.6: solution

3. If a semantic rule associated with a production p defines the value ofinherited attribute B.c in terms of the value X.a. Then , the dependencygraph has an edge from X.a to B.c.

Applications of Syntax-Directed Translation

1. Construction of syntax Trees The nodes of the syntax tree are represented by objects with a suit-ablenumber of fields. Each object will have an op field that is the label of the node. The objects will have additional fields as follows

2. If the node is a leaf, an additional field holds the lexical value forthe leaf. A constructor function Leaf (op, val) creates a leaf object.

3. If nodes are viewed as records, the Leaf returns a pointer to a newrecord for a leaf.

4. If the node is an interior node, there are as many additional fieldsas the node has children in the syntax tree. A constructor function

Node takes two or more arguments:

Node (op , c1,c2,…..ck) creates an object with first field op and k additional fields forthe k children c1,c2,…..ck

Syntax-Directed Translation Schemes A SDT scheme is a context-free grammar with program fragments embedded withinproduction bodies .The program fragments are called semantic actions and can appear atany position within the production body.

Any SDT can be implemented by first building a parse tree and then pre-forming theactions in a left-to-right depth first order. i.e during preorder traversal.

The use of SDT's to implement two important classes of SDD's

1. If the grammar is LR parsable, then SDD is S-attributed.

2. If the grammar is LL parsable, then SDD is L-attributed.

3.2 Intermediate code forms:

An intermediate code form of source program is an internal form of a program created by the compiler while translating the program created by the compiler while translating the program from a high –level language to assembly code(or)object code(machine code).an intermediate source form represents a more attractive form of target code than does assembly.An optimizing Compiler performs optimizations on the intermediate source form and produces an object module.
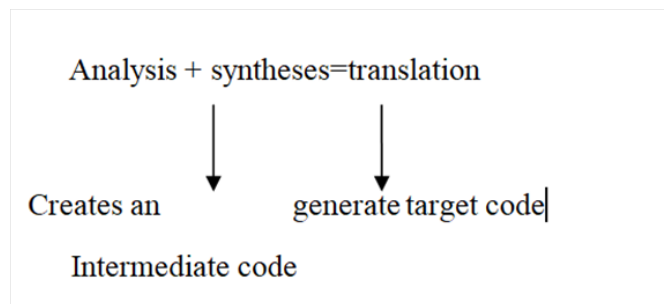
FIGURE 3.7: Intermediate code forms:

In the analysis –synthesis model of a compiler, the front-end translates a source program into an intermediate representation from which the back-end generates target code, in many compilers the source code is translated into a language which is intermediate in complexity between a HLL and machine code. the usual intermediate code introduces symbols to stand for various temporary quantities. We assume that the source program has already been parsed and statically checked..
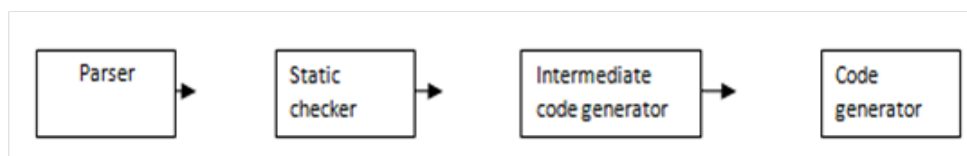
FIGURE 3.8: position of intermediate code generator:

the various intermediate code forms are:

The ordinary (infix) way of writing the sum of a and b is with the operator in the middle: a+b. the postfix (or postfix polish) notation for the same expression places the operator at the right end, as ab+.

In general, if e1 and e2 are any postfix expressions, and Ø to the values denoted by e1 and e2 is indicated in postfix notation nby e1e2Ø.no parentheses are needed in postfix notation because the position and priority (number of arguments) of the operators permits only one way to decode a postfix expression.

3.2.1   syntax tree:

The parse tree itself is a useful intermediate-language representation for a source program, especially in optimizing compilers where the intermediate code needs to extensively restructure.

A parse tree, however, often contains redundant information which can be eliminated, Thus producing a more economical representation of the source program. One such variant of a parse tree is what is called an (abstract)syntax tree,a tree in which each leaf represents an operand and each interior node an operator.

Exmples

1) Syntax tree for the expression a*(b+c)/d
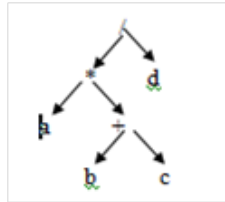
FIGURE 3.9: a*(b+c)/d

2) syntax tree for if a=b then a:=c+d else b:=c-d
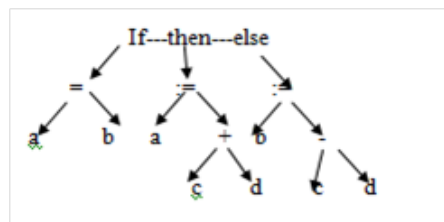
FIGURE 3.10: if a=b then a:=c+d else b:=c-d

Address and Instructions

Example Three-address code is built from two concepts: addresses and instructions.

An address can be one of the following: A name: A source name is replaced by a pointer to its symbol table entry. A name: For convenience, allow source-program names to Appear as addresses in three-address code.

In an Implementation, a source name is replaced by a pointer to its symbol-table entry, where all information about thename is kept. A constant

A constant: In practice, a compiler must deal with many differenttypes of constants and variables

A compiler-generated temporary

A compiler-generated temporary. It is useful, especially inoptimizing compilers, to create a distinct name each time atemporary is needed. These temporaries can be combined, ifpossible, when registers are allocated to variables.

A list of common three-address instruction forms:

Assignment statements

– x= y op z, where op is a binary operation

– x= op y, where op is a unary operation

– Copy statement: x=y

– Indexed assignments: x=y[i] and x[i]=y

– Pointer assignments: x= & y, *x=y and x=*y

Control flow statements

– Unconditional jump: goto L

– Conditional jump: if x relop y goto L ; if x goto L; if False x goto L

– Procedure calls: call procedure p with n parameters and return y, is

Optional

param x1

param x2

...

param xn

call p, n

C. quadruples:

1 Three-address instructions can be implemented as objects or as record with fieldsfor the operator and operands.

2 Three such representations Quadruple, triples, and indirect triples

3 A quadruple (or quad) has four fields: op, arg1, arg2, and result.

D. Triples

1 A triple has only three fields: op, arg1, and arg2

2 Using triples, we refer to the result of an operation x op y by its position, rather by an explicit temporary name.

d. Triples:

1 A triple has only three fields: op, arg1, and arg2

2 Using triples, we refer to the result of an operation x op y by its position, rather byan explicit temporary name.
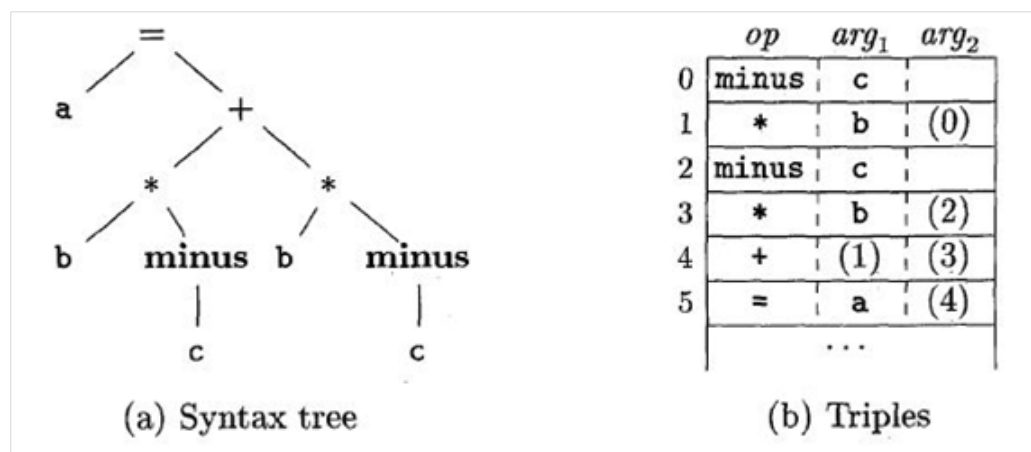
FIGURE 3.11: Example

Single Assignment Static Form

Static single assignment form (SSA) is an intermediate representation that facilitates certain code optimization,two distinct aspects distinguish SSA from three –address code, all assignments in SSA are to variables with distinct names; hence theterm static single-assignment.

instruction			op	arg_1	arg_2
35	(0)	0	minus	c	
36	(1)	1	*	b	(0)
37	(2)	2	minus	c	
38	(3)	3	*	b	(2)
39	(4)	4	+	(1)	(3)
40	(5)	5	=	a	(4)
	...			...	

FIGURE 3.12: Representations of a = b * - c + b * - c

Chapter 4

TYPE CHECKING AND RUN TIME ENVIRONMENT

Course Outcomes

After successful completion of this module, students should be able to:

CO 9	**Demonstrate** type systems for performing the static and dynamic type checking.	Understand
CO 10	**Describe** the run-time memory elements for storage allocation strategies which include procedure calls, local variable allocation, and dynamic memory allocation.	Apply

4.1 Introduction

Type checking: Definition of type checking, type expressions, type systems, static and dynamic checking of types, specification of a simple type checker. Run time environments: Source language issues, Storage organization, storage-allocation strategies, access to nonlocal data on the stack, garbage collection, symbol tables.

Type Checking:

1 A compiler has to do semantic checks in addition to syntactic checks.

2 Semantic Checks

–Static –done during compilation

–Dynamic –done during run-time

Type checking is one of these static checking operations.

–we may not do all type checking at compile-time.

–Some systems also use dynamic type checking too.

1 A type system is a collection of rules for assigning type expressions to the parts of a program.

A type checker implements a type system.

2 A sound type system eliminates run-time type checking for type errors.

3 A programming language is strongly-typed, if every program its compiler accepts will execute without type errors.

In practice, some of type checking operations is done at run-time (so, most of the programming languages are not strongly yped).

Ex: int x[100]; ... x[i] most of the compilers cannot guarantee that i will be between 0 and 99

Type Expression:

The type of a language construct is denoted by a type expression.

A type expression can be:

A basic type

a primitive data type such as integer, real, char, Boolean, ...

type-error to signal a type error

void: no type

A type name

a name can be used to denote a type expression.

A type constructor applies to other type expressions.

arrays: If T is a type expression, then array (I,T)is a type expression where I denotes index range. Ex: array (0..99,int)

products: If T1and T2 are type expressions, then their Cartesian product T1 x T2 is a type expression. Ex: int x int

pointers: If T is a type expression, then pointer (T) is a type expression. Ex: pointer (int)

functions: We may treat functions in a programming language as mapping from a domain type D to a range type R. So, the type of a function can be denoted by the type expression D→R where D are R type expressions. Ex: int→int represents the type of a function which takes an int value as parameter, and its return type is also int.

4.2 Type Checking of Statements:

S -> d= E if (id.type=E.type then S.type=void else S.type=type-error

S ->if E then S1 if (E.type=boolean then S.type=S1.type else S.type=type-error

S->while E do S1 if (E.type=boolean then S.type=S1.type else S.type=type-error

E->E1(E2) if (E2.type=s and E1.type=s>t) then E.type=t

else E.type=type-error

Ex: int f(double x, char y) ...

4.2.1   Structural Equivalence of Type Expressions:

1. How do we know that two type expressions are equal?

2. As long as type expressions are built from basic types (no type names), we may use structural equivalence between two type expressions

4.2.2   (

Structural Equivalence Algorithm (sequin):)

if (s and t are same basic types) then return true

else if (s=array(s1,s2) and t=array(t1,t2)) then return (sequiv(s1,t1) and sequiv(s2,t2))

else if (s = s1 x s2and t = t1 x t2) then return (sequiv(s1,t1) and sequiv(s2,t2))

else if (s=pointer(s1) and t=pointer(t1)) then return (sequiv(s1,t1))

else if (s = s1,s2and t = t1 ,t2) then return (sequiv(s1,t1) and sequiv(s2,t2))

else return false

4.2.3   Names for Type Expressions:

In some programming languages, we give a name to a type expression, and we use that name as a type expression afterwards.

type link = ↑cell; ? p,q,r,s have same types ?

var p,q : link;

var r,s : ↑cell

1. How do we treat type names?

–Get equivalent type expression for a type name (then use structural equivalence), or –Treat a type name as a basic type

4.3   Symbol Tables

A symbol tableis a major data structure used in a compiler:

1 Associatesattributeswithidentifiersused in a program.

2 For instance, atype attributeisusually associatedwitheach identifier.

3 A symbol tableis anecessarycomponent.

4 Definition(declaration)ofidentifiersappearsonceina program

5 Useofidentifiersmayappearinmanyplacesoftheprogramtext

6 Identifiers andattributes are entered bytheanalysisphases

7 Whenprocessingadefinition(declaration)of anidentifier

8 Insimplelanguageswithonlyglobalvariablesandimplicitdeclarations:

9 Thescannercanenteranidentifierintoasymboltableifit isnotalreadythere

10 Inblock-structuredlanguageswithscopesandexplicitdeclarations:

11 Theparserand/orsemanticanalyzerenteridentifiersandcorrespondingattributes

12 Symbol table information is used by the analysis and synthesis phases

13 Toverifythatusedidentifiershavebeendefined(declared)

14 Toverifythatexpressionsandassignmentsaresemanticallycorrect–typechecking

15 Togenerateintermediateortargetcode

4.3.1   Symbol Table Interface:

The basic operations defined on a symbol table include:

1 allocate– to allocateanew empty symbol table

2 free–to remove allentries and freethe storageof asymbol table

3 insert–toinsert aname in asymbol tableandreturn apointerto its entry

4 lookup– tosearch foraname and return apointer to its entry

5 set-attribute–toassociateanattribute with a given entry

6 get-attribute–to get anattribute associatedwithagiven entry

7 ther operations can be added depending on requirement

For example, adeleteoperation removesa name previously inserted

Someidentifiersbecomeinvisible(outofscope)afterexitingablock 8 This interface provides an abstract view of a symbol table.

9 Supports the simultaneous existence of multiple tables

10 Implementation can vary without modifying the interface

Basic Implementation Techniques:

First consideration is how to insert and lookup names

Variety of implementation techniques

Unordered List

Simplest toimplement

Implementedas an arrayor alinked list

Linked listcan grow dynamically –alleviates problemof afixed size array

Insertion isfast O(1), butlookup isslow for largetables– O(n) onaverage

Ordered List

Ifanarray is sorted,itcan besearched using binarysearch– O(log2n)

Insertion into asortedarray isexpensive–O(n) onaverage

Usefulwhen set of namesisknowninadvance –table of reservedwords

Binary Search Tree

Cangrow dynamically

Insertion and lookup are O(log2n) onaverage

4.3.2   Hash Tables and Hash Functions:

A hash table isanarraywithindexrange:0toTableSize – 1

Most commonly used data structure to implement symbol tables

Insertion and lookup can be made very fast – O(1)

A hash function mapsan identifier name into a table index

A hash function, h(name),should depend solely on name

h(name) should becomputed quickly

hshould beuniformand randomizingin distributing names

All tableindicesshould bemapped withequal probability

Similar names should not clusterto the same tableindex.

Storage Allocation:

Compiler must do the storage allocation and provide access to variables and data

Memory management

Stack allocation

Heap management

Garbage collection
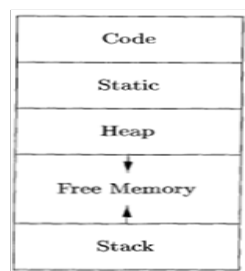
FIGURE 4.1: Storage Organization:

Assumes a logical address space

Operating system will later map it to physical addresses; decide how touse cache memory, etc.

Memory typically divided into areas for

Program code

Other static data storage, including global constants and compilergenerated data

Stack to support call/return policy for procedures

Heap to store data that can outlive a call to a procedure

4.3.3   Static vs Dynamic Allocation

1. Static: Compile time, Dynamic Runtime allocation

2. Many compilers use some combination of following

3. Stack storage: for local variables, parameters and so on

4. Heap storage: Data that may outlive the call to the procedure that created it

5. Stack allocation is a valid allocation for procedures since procedure calls are nest

Example:

Consider the quick sort program

```
int a[11];
void readArray() { /* Reads 9 integers into a[1], ..., a[9]. */
    int i;
    ...
}
int partition(int m, int n) {
    /* Picks a separator value v, and partitions a[m .. n] so that
        a[m .. p − 1] are less than v, a[p] = v, and a[p + 1 .. n] are
        equal to or greater than v. Returns p. */
    ...
}
void quicksort(int m, int n) {
    int i;
    if (n > m) {
        i = partition(m, n);
        quicksort(m, i-1);
        quicksort(i+1, n);
    }
}
main() {
    readArray();
    a[0] = -9999;
    a[10] = 9999;
    quicksort(1,9);
}
```

FIGURE 4.2: quick sort program

Activation for Quicksort:

Activation tree representing calls during an execution of quicksort:

Activation records

Procedure calls and returns are usually managed by a run-time stack called thecontrol stack.

1 Each live activation has an activation record (sometimes called a frame)

2 The root of activation tree is at the bottom of the stack

3 The current execution path specifies the content of the stack with the last

4 Activation has record in the top of the stack.

Activation Record

1 Temporary values

```
enter main()
    enter readArray()
    leave readArray()
    enter quicksort(1,9)
        enter partition(1,9)
        leave partition(1,9)
        enter quicksort(1,3)
            ...
        leave quicksort(1,3)
        enter quicksort(5,9)
            ...
        leave quicksort(5,9)
    leave quicksort(1,9)
leave main()
```

FIGURE 4.3: Activation for Quicksort:
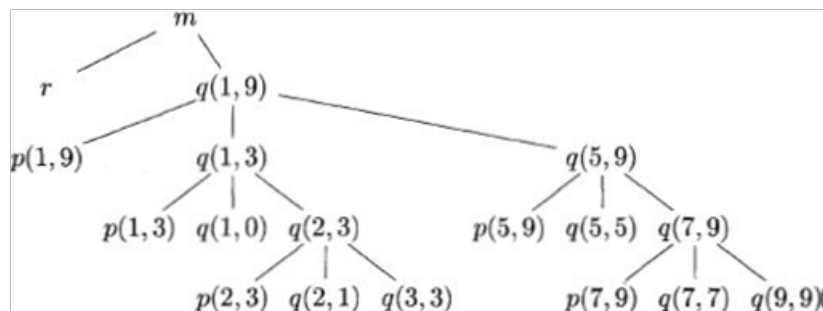
FIGURE 4.4: A General Activation Record

2 Local data

3 A saved machine status

4 An "access link"

5 A control link

6 Space for the return value of the called function

7 The actual parameters used by the calling procedure

8 Elements in the activation record:

9 Temporary values that could not fit into registers.

10 Local variables of the procedure.

11 Saved machine status for point at which this procedure called. Includes return address and contents of registers to be restored.

12 Access link to activation record of previous block or procedure in lexical scope chain.

13 Control link pointing to the activation record of the caller.

14 Space for the return value of the function, if any.

15 actual parameters (or they may be placed in registers, if possible)

Downward-growing stack of activation records:
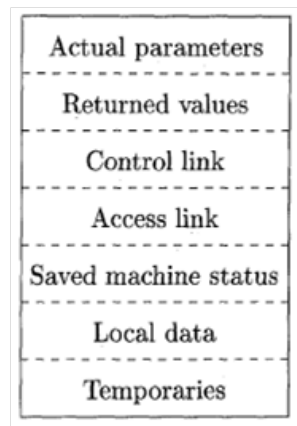
FIGURE 4.5: A General Activation Record
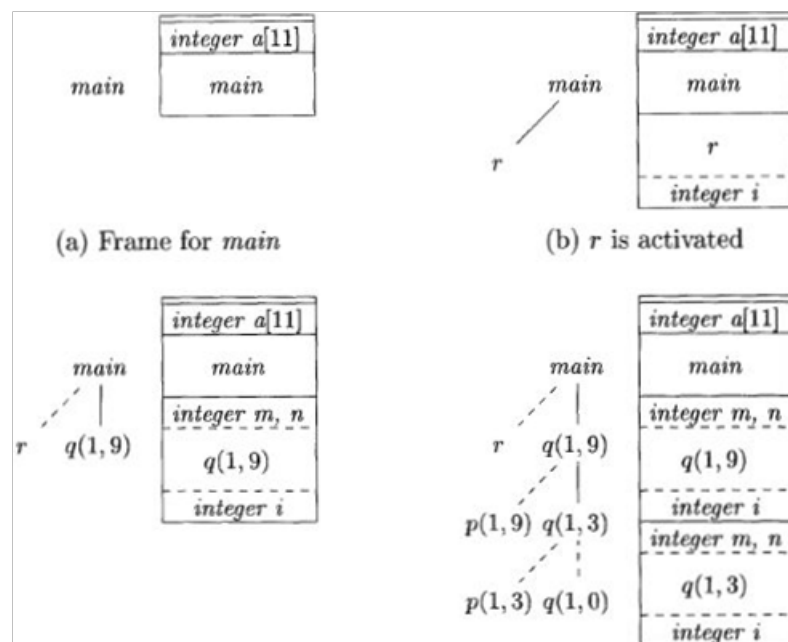
FIGURE 4.6: Downward-growing stack of activation records:

4.3.4   Designing Calling Sequences:

1 Values communicated between caller and callee are generally placed at thebeginning of callee's activation record

2 Fixed-length items: are generally placed at the middle

3 Items whose size may not be known early enough: are placed at the end ofactivation record

4 We must locate the top-of-stack pointer judiciously: a common approach is tohave it point to the end of fixed length fields

Access to dynamically allocated arrays:

ML:

1. ML is a functional language
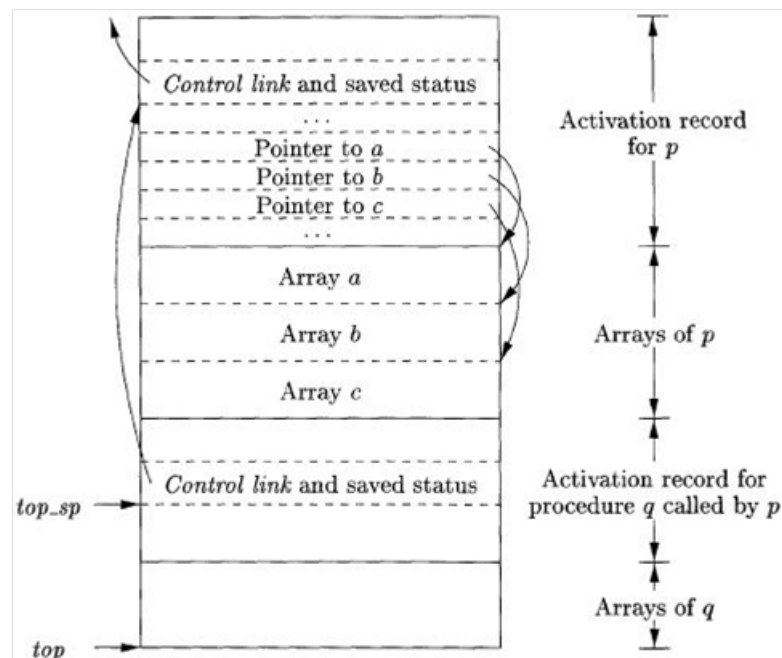
FIGURE 4.7: Downward-growing stack of activation records:

2. Variables are defined, and have their unchangeable values initialized, by a statementof the form:

val (name) = (expression)

3. Functions are defined using the syntax:

fun (name) ((arguments)) = (body)

4. For function bodies we shall use let-statements of the form let (list of definitions) in (statements) end

A version of quicksort, in ML style, using nested functions:

```
 1) fun sort(inputFile, outputFile) =
        let
 2)         val a = array(11,0);
 3)         fun readArray(inputFile) = ··· ;
 4)              ··· a ··· ;
 5)         fun exchange(i,j) =
 6)              ··· a ··· ;
 7)         fun quicksort(m,n) =
                let
 8)                 val v = ··· ;
 9)                 fun partition(y,z) =
10)                      ··· a ··· v ··· exchange ···
                in
11)                  ··· a ··· v ··· partition ··· quicksort
                end
        in
12)          ··· a ··· readArray ··· quicksort ···
        end;
```

FIGURE 4.8: ML style, using nested functions

Access links for finding nonlocal data:
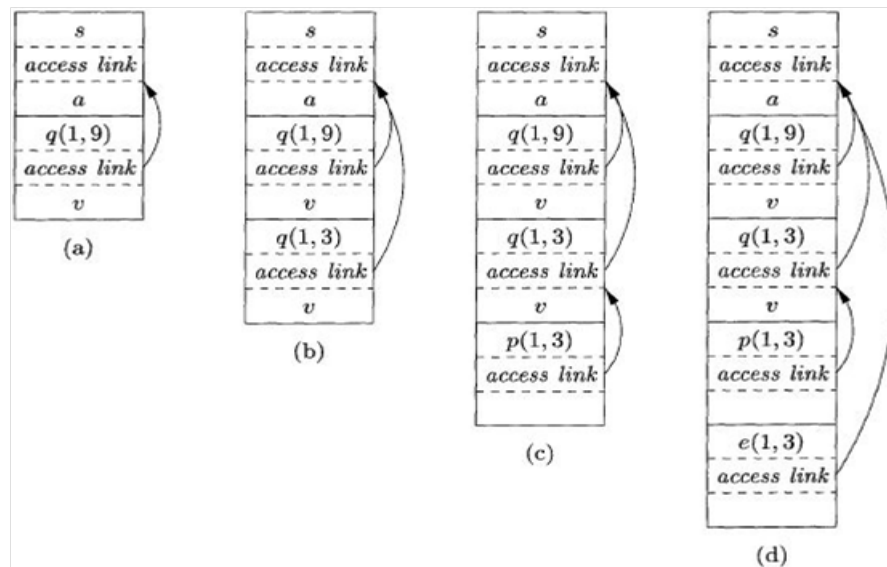
FIGURE 4.9: Access links for finding nonlocal data

Sketch of ML program that uses function-parameters:

```
fun a(x) =
    let
        fun b(f) =
            ... f ... ;
        fun c(y) =
            let
                    fun d(z) = ...
            in
                    ... b(d) ...
            end
    in
            ... c(1) ...
    end;
```

FIGURE 4.10: ML program that uses function-parameters:

Actual parameters carry their access link with them:
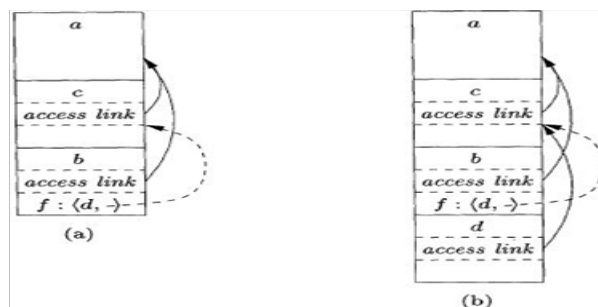
FIGURE 4.11: Actual parameters carry their access link with them

Maintaining the Display:

Memory Manager:

1. Two basic functions:
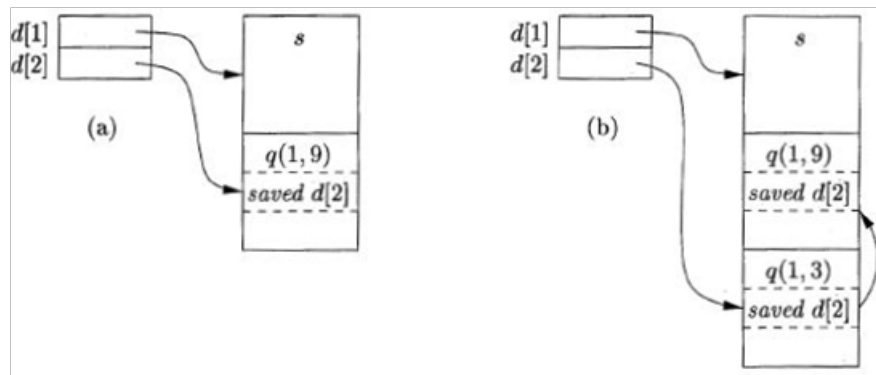
FIGURE 4.12: Maintaining the Display

2. Allocation

3. De allocation

4. Properties of memory managers:

5. Space efficiency

6. Program efficiency

7. Low overhead

Typical Memory Hierarchy Configurations:
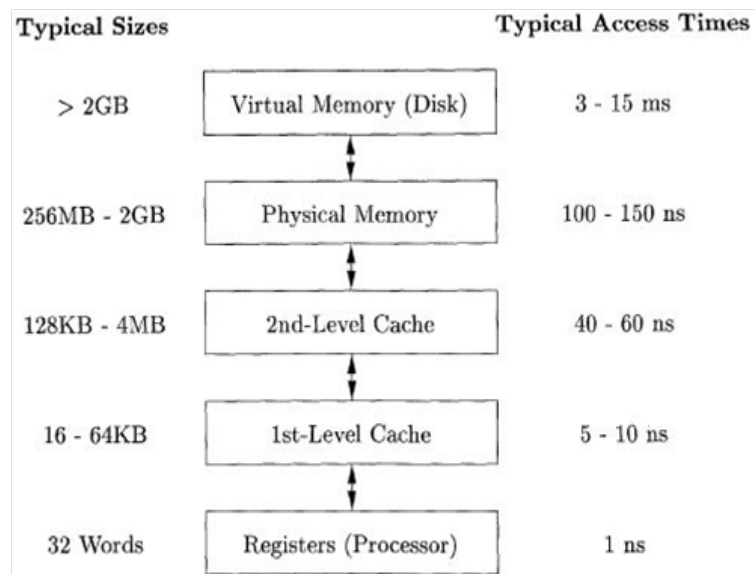
FIGURE 4.13: Typical Memory Hierarchy Configurations:

4.3.5 Locality in Programs:

The conventional wisdom is that programs spend 90% of their time executing 10% of the code:

1. Programs often contain many instructions that are never executed.

2. Only a small fraction of the code that could be invoked is actually executed in atypical run of the program.

3. The typical program spends most of its time executing innermost loops and tightrecursive cycles in a program.

Chapter 5

CODE OPTIMIZATION AND CODE GENERATOR

Course Outcomes

After successful completion of this module, students should be able to:

CO 11	**Apply** the code optimization techniques on intermediate code form for improving the performance of a program.	Apply
CO 12	**Make use of** optimization techniques on basic blocks for reducing utilization of registers in generating the target code.	Apply

5.1 INTRODUCTION

Code optimization: The principle sources of optimization, optimization of basic blocks, loops in flow graphs, peephole optimization. Code Generation: Issues in the Design of a Code Generator, The Target Language, addresses in the Target Code, Basic Blocks and Flow Graphs, Optimization of Basic Blocks, A Simple Code Generator, register allocation and assignment, DAG representation of basic blocks.

* The code produced by the straight forward compiling algorithms can often be made to run faster or take less space, or both. This improvement is achieved by program transformations that are traditionally called optimizations. Compilers that apply code-improving transformations are called optimizing compilers.

optimizations are classified into two categories. They are

1. Machine independent optimizations:

2. Machine dependant optimizations:

63

Machine independent optimizations:

Machine independent optimizations are program transformations that improve the target code without taking into consideration any properties of the target machine.

Machine dependant optimizations:

Machine dependant optimizations are based on register allocation and utilization of special machine-instruction sequences.

5.1.1   The criteria for code improvement transformations:

• Simply stated, the best program transformations are those that yield the most benefit for the least effort.

• The transformation must preserve the meaning of programs. That is, the optimization must not change the output produced by a program for a given input, or cause an error such as division by zero, that was not present in the original source program. At all times we take the "safe" approach of missing an opportunity to apply a transformation rather than risk changing what the program does.

• A transformation must, on the average, speed up programs by a measurable amount. We are also interested in reducing the size of the compiled code although the size of the code has less importance than it once had. Not every transformation succeeds in improving every program, occasionally an "optimization" may slow down a program slightly.

• The transformation must be worth the effort. It does not make sense for a compiler writer to expend the intellectual effort to implement a code improving transformation and to have the compiler expend the additional time compiling source programs if this effort is not repaid when the target programs are executed. "Peephole" transformations of this kind are simple enough and beneficial enough to be included in any compiler.

• Flow analysis is a fundamental prerequisite for many important types of code improvement.

• Generally control flow analysis precedes data flow analysis.

• Control flow analysis (CFA) represents flow of control usually in form of graphs, CFA constructs such as

• control flow graph

• Call graph

• Data flow analysis (DFA) is the process of ascerting and collecting information prior to program execution about the possible modification, preservation, and use of certain entities (such as values or attributes of variables) in a computer program.

5.1.2 Principal Sources of Optimization

• A transformation of a program is called local if it can be performed by looking only at the statements in a basic block; otherwise, it is called global.

• Many transformations can be performed at both the local and global levels. Local transformations are usually performed first.

5.1.3 Function-Preserving Transformations

• There are a number of ways in which a compiler can improve a program without changing the function it computes.

• The transformations

• Common sub expression elimination,

• Copy propagation,

• Dead-code elimination, and

• Constant folding, are common examples of such function-preserving transformations. The other transformations come up primarily when global optimizations are performed.

• Frequently, a program will include several calculations of the same value, such as an offset in an array. Some of the duplicate calculations cannot be avoided by the programmer because they lie below the level of detail accessible within the source language.

5.1.4 Common Sub expressions elimination:

• An occurrence of an expression E is called a common sub-expression if E was previously computed, and the values of variables in E have not changed since the previous computation. We can avoid recomputing the expression if we can use the previously computed value.

• For example

t1: =4*i t2: =a [t1]

t3: =4*j t4:=4*i

t5: =n

t6: =b [t4] +t5

The above code can be optimized using the common sub-expression elimination as

t1: =4*i t2: =a [t1] t3: =4*j t5: =n

t6: =b [t1] +t5

The common sub expression t 4: =4*i is eliminated as its computation is already in t1. And value of i is not been changed from definition to use

5.1.5   Copy Propagation:

Assignments of the form f : = g called copy statements, or copies for short. The idea behind the copy-propagation transformation is to use g for f, whenever possible after the copy statement f: = g. Copy propagation means use of one variable instead of another. This may not appear to be an improvement, but as we shall see it gives us an opportunity to eliminate x.

For example:

x=Pi;

. . .

A=x*r*r;

The optimization using copy propagation can be done as follows:

A=Pi*r*r;

Here the variable x is eliminated

5.1.6   Dead-Code Eliminations:

A variable is live at a point in a program if its value can be used subsequently; otherwise, it is dead at that point. A related idea is dead or useless code, statements that compute values that never get used. While the programmer is unlikely to introduce any dead code intentionally, it may appear as the result of previous transformations. An optimization can be done byeliminating dead code.

Example:

i=0;

if(i=1)

{

a=b+5;

}

Here, 'if'statement is dead code because this condition will never get satisfied.

5.1.7   Constant folding:

• We can eliminate both the test and printing from the object code. More generally, deducing at compile time that the value of an expression is a constant and using the constant instead is known as constant folding.

• One advantage of copy propagation is that it often turns the copy statement into dead code.

For example,

a=3.14157/2 can be replaced by

a=1.570 thereby eliminating a division operation.

5.1.8   Loop Optimizations:

•o We now give a brief introduction to a very important place for optimizations, namely loops, especially the inner loops where programs tend to spend the bulk of their time. The running time of a program may be improved if we decrease the number of instructions in an inner loop, even if we increase the amount of code outside that loop.

• Three techniques are important for loop optimization:

• code motion, which moves code outside a loop;

• Induction -variable elimination, which we apply to replace variables from inner loop.

• Reduction in strength, which replaces and expensive operation by a cheaper one, such as a multiplication by an addition.

5.1.9   Code Motion:

• An important modification that decreases the amount of code in a loop is code motion. This transformation takes an expression that yields the same result independent of the number of times a loop is executed (a loop-invariant computation) and places the expression before the loop. Note that the notion "before the loop" assumes the existence of an entry for the loop. For example, evaluation of limit-2 is a loop-invariant computation in the following while-statement:

while (i <= limit-2) /* statement does not change Limit*/

Code motion will result in the equivalent of

t= limit-2;

while (i<=t) /* statement does not change limit or t */

5.1.10   Induction Variables:

• Loops are usually processed inside out. For example consider the loop around B3.

• Note that the values of j and t4 remain in lock-step; every time the value of j decreases by 1, that of t4 decreases by 4 because 4*j is assigned to t4. Such identifiers are called induction variables.

• When there are two or more induction variables in a loop, it may be possible to get rid of all but one, by the process of induction-variable elimination. For the inner loop around B3 in Fig. we cannot get rid of either j or t4 completely; t4 is used in B3 and j in B4.

• However, we can illustrate reduction in strength and illustrate a part of the process of induction-variable elimination. Eventually j will be eliminated when the outer loop of B2 - B5 is considered.

Example:

As the relationship t4:=4*j surely holds after such an assignment to t 4 in Fig. and t4 is not changed elsewhere in the inner loop around B3, it follows that just after the statement j:=j -1 the relationship t4:= 4*j-4 must hold. We may therefore replace the assignment t 4:= 4*j by t4:= t4-4. The only problem is that t 4 does not have a value when we enter block B3 for the first time. Since we must maintain the relationship t4=4*j on entry to the block B3, we place an initializations of t4 at the end of the block where j itself is initialized, shown by the dashed addition to block B1 in second Fig.

The replacement of a multiplication by a subtraction will speed up the object code if multiplication takes more time than addition or subtraction, as is the case on many machines.

5.1.11   Reduction in Strength:

• Reduction in strength replaces expensive operations by equivalent cheaper ones on the target machine. Certain machine instructions are considerably cheaper than others and can often be used as special cases of more expensive operators.

• For example, x^2 is invariably cheaper to implement as x*x than as a call to an exponentiation routine. Fixed-point multiplication or division by a power of two is cheaper to implement as a shift. Floating-point division by a constant can be implemented as multiplication by a constant, which may be cheaper.

 There are two types of basic block optimizations. They are :

Structure -Preserving Transformations

Algebraic Transformations

5.1.12   Structure- Preserving Transformations:

The primary Structure-Preserving Transformation on basic blocks are:

• Common sub-expressionelimination

• Dead code elimination

• Renaming of temporary variables

• Interchange of two independent adjacent statements.

Common sub-expression elimination: Common sub expressions need not be computed over and over again. Instead they can be computed once and kept in store from where it's referenced when encountered again – of course providing the variable values in the expression still remain constant.

Example:

a: =b+c

b: =a-d
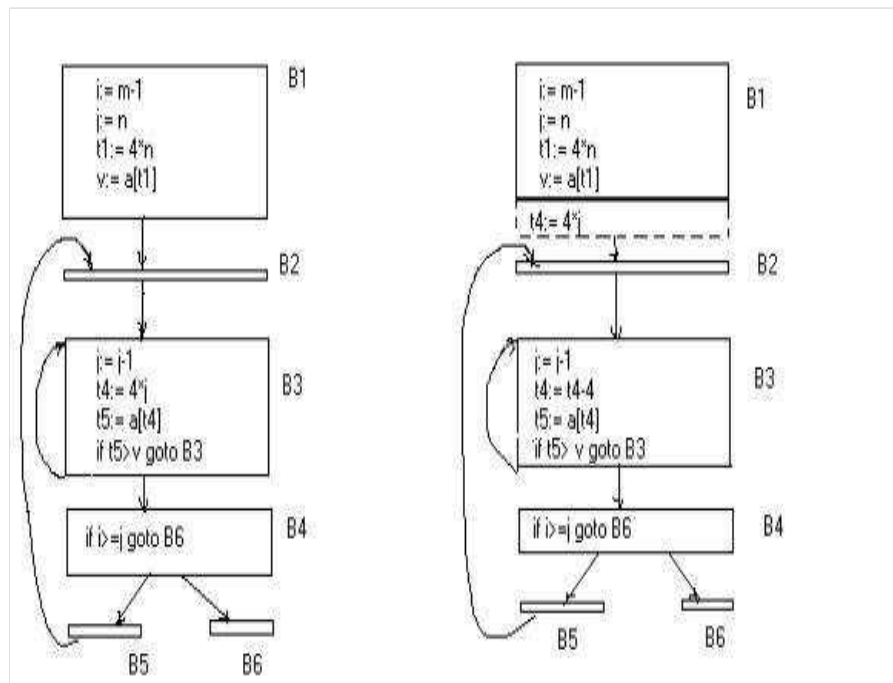
FIGURE 5.1: Optimization of Basic Blocks

c: =b+c

d: =a-d

The 2nd and 4th statements compute the same expression: b+c and a-d

Basic block can be transformed to

a: =b+c

b: =a-d

c: =a

d: =b

Dead code elimination:

It's possible that a large amount of dead (useless) code may exist in the program. This might be especially caused when introducing variables and procedures as part of construction or error -correction of a program – once declared and defined, one forgets to remove them in case they serve no purpose. Eliminating these will definitely optimize the code.

Renaming of temporary variables:

• A statement t:=b+c where t is a temporary name can be changed to u:=b+c where u is another temporary name, and change all uses of t to u.

• In this we can transform a basic block to its equivalent block called normal-form block.

Interchange of two independent adjacent statements:

Two statements

t1:=b+c

t2:=x+y

can be interchanged or reordered in its computation in the basic block when value of t1 does not affect the value of t2.

Algebraic Transformations:

• Algebraic identities represent another important class of optimizations on basic blocks. This includes simplifying expressions or replacing expensive operation by cheaper ones i.e. reduction in strength.

• Another class of related optimizations is constant folding. Here we evaluate constant expressions at compile time and replace the constant expressions by their values. Thus the expression 2*3.14 would be replaced by 6.28.

• The relational operators $< =, > =, <, >, +$ and $=$ sometimes generate unexpected common sub expressions.

• Associative laws may also be applied to expose common sub expressions. For example, if the source code has the assignments

a :=b+c e := c+d+b

the following intermediate code may be generated:

a :=b+c t :=c+d

e :=t+b

Example: x:=x+0 can be removed

x:=y**2 can be replaced by a cheaper statement x:=y*y

• The compiler writer should examine the language carefully to determine what rearrangements of computations are permitted; since computer arithmetic does not always obey the algebraic identities of mathematics. Thus, a compiler may evaluate x*y-x*z as x*(y-z) but it may not evaluate a+(b-c) as (a+b)-c.

5.2   Loops in Flow Graph

A graph representation of three-address statements, called a flow graph, is useful for understanding code-generation algorithms, even if the graph is not explicitly constructed by a code-generation algorithm. Nodes in the flow graph represent computations, and the edges represent the flow of control.

Dominators:

In a flow graph, a node d dominates node n, if every path from initial node of the flow graph to n goes through d. This will be denoted by d dom n. Every initial node dominates all the remaining nodes in the flow graph and the entry of a loop dominates all nodes in the loop. Similarlyeveryn-ode dominates itself.

Example:

*In the flow graph below,

*Initial node,node1 dominates every node. *node 2 dominates itself

*node 3 dominates all but 1 and 2. *node 4 dominates all but 1,2 and 3.

*node 5 and 6 dominates only themselves,since flow of control can skip around either by goin through the other.

*node 7 dominates 7,8 ,9 and 10. *node 8 dominates 8,9 and 10.

*node 9 and 10 dominates only themselves.
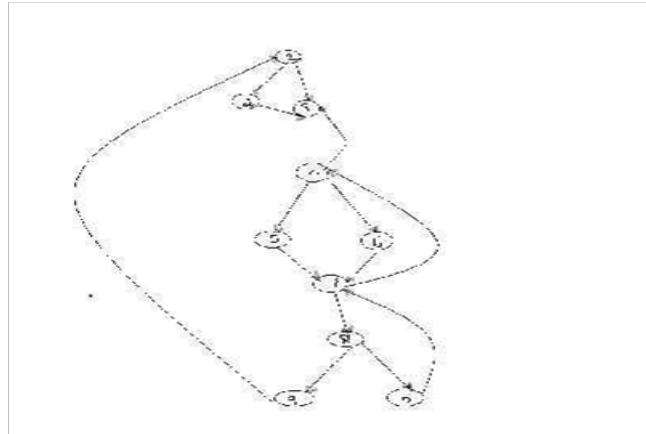
FIGURE 5.2: Dominators

• The way of presenting dominator information is in a tree, called the dominator tree in which the initial node is the root.

• The parent of each other node is its immediate dominator.

• Each node d dominates only its descendents in the tree.

• The existence of dominator tree follows from a property of dominators; each node has a unique immediate dominator in that is the last dominator of n on any path from the initial node to n.

• In terms of the dom relation, the immediate dominator m has the property is d=!n and d dom n, then d dom m.

D(1)={1}

D(2)={1,2}

D(3)={1,3}

D(4)={1,3,4}
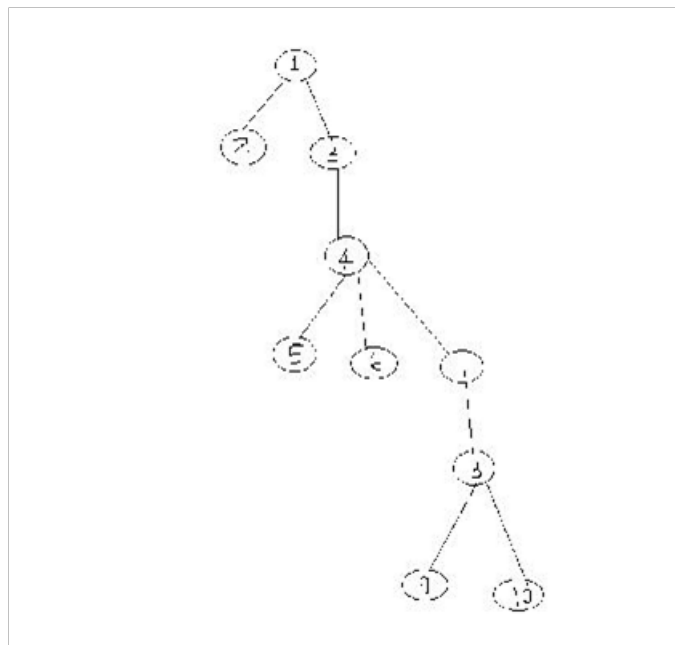
FIGURE 5.3: Dominators

D(5)={1,3,4,5}

D(6)={1,3,4,6}

D(7)={1,3,4,7}

D(8)={1,3,4,7,8}

D(9)={1,3,4,7,8,9}

D(10)={1,3,4,7,8,10}

5.2.1   Natural Loop:

• One application of dominator information is in determining the loops of a flow graph suitable for improvement.

The properties of loops are

• A loop must have a single entry point, called the header. This entry point-dominates all nodes in

the loop, or it would not be the sole entry to the loop.

• There must be at least one wayto iterate the loop(i.e.)at least one path back to the header.

• One way to find all the loops in a flow graph is to search for edges in the flow graph whose heads dominate their tails. If a→ b is an edge, b is the head and a is the tail. These types of edges are called as back edges.

7 → 4	4 DOM 7
0 →7	7 DOM 10
4 → 3	
8 → 3	
9 →1	
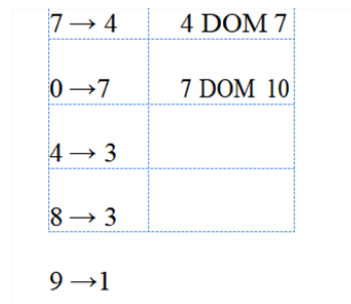
FIGURE 5.4: Natural Loop:

• The above edges will form loop in flow graph.

• Given a back edge n → d, we define the natural loop of the edge to be d plus the set of nodes that can reach n without going through d. Node d is the header of the loop.

Algorithm: Constructing the natural loop of a back edge.

Input: A flow graph G and a back edge n→d

Output: The set loop consisting of all nodes in the natural loop n→d.

Method: Beginning with node n, we consider each node m*d that we know is in loop, to makesure that m's predecessors are also placed in loop. Each node in loop, except for d, is placed once on stack, so its predecessors will be examined. Note that because d is put in the loop initially, we never examine its predecessors, and thus find only those nodes that reach n without going through d.

Procedure insert(m);

if m is not inloop then begin loop := loop U m;push m onto stack

end;

stack : =empty;

loop : =d; insert(n);

while stackis not empty do begin

pop m, the first element of stack, off stack; for eachpredecessorpofm do insert(p)

end;

InnerLoop:

• If we use the natural loops as "the loops", then we have the useful property that unless two loops have the same header, they are either disjointed or one is entirely contained in the other. Thus, neglecting loops with the same header for the moment, we have a natural notion of inner loop: one

that contains no other loop.

• When two natural loops have the same header, but neither is nested within the other, they are combined and treated as a single loop.

Pre-Headers:

• Several transformations require us to move statements "before the header". Therefore begin treatment of a loop L by creating a new block, called the preheater.

• The pre -header has only the header as successor, and all edges which formerly entered the header of Lfrom outside L instead enter the pre-header.

• Edges from inside loop L to the header are not changed.

• Initially the pre-header is empty, but transformations on L may place statements in it.
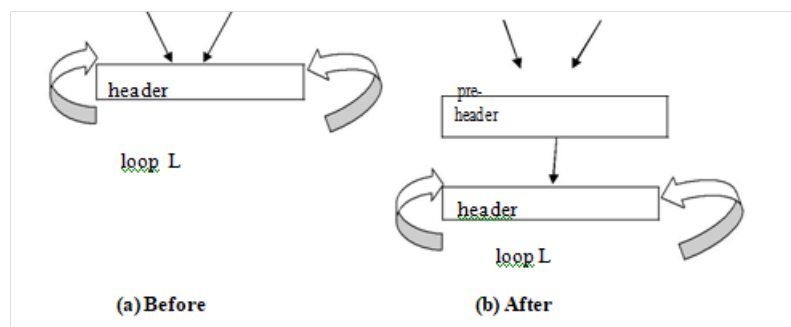
FIGURE 5.5: Pre-Headers:

5.3 Reducible flow graphs:

• Reducible flow graphs are special flow graphs, for which several code optimization transformations are especially easy to perform, loops are unambiguously defined, dominators can be easily calculated, data flow analysis problems can also be solved efficiently.

• Exclusive use of structured flow-of-control statements such as if-then-else, while-do, continue, and break statements produces programs whose flow graphs are always reducible. The most important properties of reducible flow graphs are that there are no jumps into the middle of loops from outside; the only entry to a loop is through its header. **Definition:**

• A flow graph G is reducible if and only if we can partition the edges into two disjoint groups, forward edges and back edges, with the following properties.

• The forward edges from an acyclic graph in which every node can be reached from initial node of G.

• The back edges consist only of edges where heads dominate theirs tails.

• Example: The above flow graph is reducible.

• If we know the relation DOM for a flow graph, we can find and remove all the back edges. • The

remaining edges are forward edges.

• If the forward edges form an acyclic graph, then we can say the flow graph reducible.

• In the above example remove the five back edges $4 \rightarrow 3$, $7 \rightarrow 4$, $8 \rightarrow 3$, $9 \rightarrow 1$ and $10 \rightarrow 7$ whose heads dominate their tails, the remaining graph is acyclic.

• The key property of reducible flow graphs for loop analysis is that in such flow graphs every set of nodes that we would informally regard as a loop must contain a back edge.

5.3.1 Peephole Optimization

• A statement-by-statement code-generations strategy often produce target code that contains redundant instructions and suboptimal constructs .The quality of such target code can be improved by applying "optimizing" transformations to the target program.

• A simple but effective technique for improving the target code is peephole optimization, a method for trying to improving the performance of the target program by examining a short sequence of target instructions (called the peephole) and replacing these instructions by a shorter or faster sequence, whenever possible.

• The peephole is a small, moving window on the target program. The code in the peephole need not contiguous, although some implementations do require this.it is characteristic of peephole optimization that each improvement may spawn opportunities for additional improvements.

• We shall give the following examples of program transformations that are characteristic of peephole optimizations:

1. Redundant-instructions elimination

2. Flow-of-control optimizations

3. Algebraic simplifications

4. Use of machine idioms

5. Unreachable Code

Redundant Loads And Stores:

If we see the instructions sequence

(1) MOV R0,a

(2) MOV a,R0

we can delete instructions (2) because whenever (2) is executed. (1) will ensure that the value of a is already in register R0.If (2) had a label we could not be sure that (1) was always executedimmediately before (2) and so we could not remove (2).

Unreachable Code: • Another opportunity for peephole optimizations is the removal of unreachable instructions. An unlabeled instruction immediately following an unconditional jump may be removed. This operation can be repeated to eliminate a sequence of instructions. For example, for debugging purposes, a large program may have within it certain segments that are executed only if a variable debug is 1. In C, the source code might look like:

define debug 0 . . . .

If (debug) { Print debugging information

}

In the intermediate representations the if-statement may be translated as:

debug =1 goto L2

goto L2

L1: print debugging information

L2:. . . . . (a)

• One obvious peephole optimization is to eliminate jumps over jumps .Thus no matter what the value of debug; (a) can be replaced by:

If debug $\neq$ 1 goto L2

Print debugging information

L2:. . . . . (b)

• As the argument of the statement of (b) evaluates to a constant true it can be replaced by If debug $\neq$0 goto L2

Print debugging information

L2: . . . . (c)

• As the argument of the first statement of (c) evaluates to a constant true, it can be replaced by goto L2. Then all the statement that print debugging aids are manifestly unreachable and can be eliminated one at a time.

5.4   Flows-Of-Control Optimizations:

• The unnecessary jumps can be eliminated in either the intermediate code or the target code by the following types of peephole optimizations. We can replace the jump sequence

goto L1

. . . .

L1: gotoL2

by the sequence

goto L2

. . . .

L1: goto L2

• If there are now no jumps to L1, then it may be possible to eliminate the statement L1:goto L2 provided it is preceded by an unconditional jump .Similarly, the sequence

if a > b goto L1

. . . .

L1: goto L2

can be replaced by

Ifa < b goto L2

. . . .

L1: goto L2

• Finally, suppose there is only one jump to L1 and L1 is preceded by an unconditional goto. Then the sequence

goto L1

……..

L1: if a <b goto L2

L3:…………………………………..(1)

• Maybe replaced by

Ifa< b goto L2

goto L3

…….

L3:…………………………………(2)

• While the number of instructions in (1) and (2) is the same, we sometimes skip the unconditional jump in (2), but never in (1).Thus (2) is superior to (1) in execution time

Algebraic Simplification:

• There is no end to the amount of algebraic simplification that can be attempted through peephole optimization. Only a few algebraic identities occur frequently enough that it is worth considering implementing them .For example, statements such as

x := x+0

Or

x := x * 1

• Areoften produced by straightforward intermediate code-generation algorithms, and they can be eliminated easily through peephole optimization.

Reduction in Strength:

• Reduction in strength replaces expensive operations by equivalent cheaper ones on the target machine. Certain machine instructions are considerably cheaper than others and can often be used as special cases of more expensive operators.

• For example, x^2 is invariably cheaper to implement as x*x than as a call to an exponentiation routine. Fixed-point multiplication or division by a power of two is cheaper to implement as a shift. Floating-point division by a constant can be implemented as multiplication by a constant, which may be cheaper.

X2 → X*X

5.5   Useof Machine Idioms:

• The target machine may have hardware instructions to implement certain specific operations effi-ciently. For example, some machines have auto-increment and auto-decrement addressing modes. These add or subtract one from an operand before or after using its value.

• The use of these modes greatly improves the quality of code when pushing or popping a stack, as in parameter passing. These modes can also be used in code for statements like i : =i+1.

i:=i+1 → i++

i:=i-1 → i–

Code Improvig Transformations

• Algorithms for performing the code improving transformations rely on data-flow information. Here we consider common sub-expression elimination, copy propagation and transformations for moving loop invariant computations out of loops and for eliminating induction variables.

• Global transformations are not substitute for local transformations; both must be performed.

Elimination of global common sub expressions:

• The available expressions data-flow problem discussed in the last section allows us to determine if an expression at point p in a flow graph is a common sub-expression. The following algorithm formalizes the intuitive ideas presented for eliminating common sub-expressions.

5.5.1   Algorithm: Global common sub expression elimination.

Input : A flow graph with available expression information.

Output: A revised flow graph.

Method: For every statement s of the form x := y+z6 such that y+z is available at the beginning of block and neither y nor r z is defined prior to statement s in that block, do the following.

• To discover the evaluations of y+z that reach s's block, we follow flow graph edges, searching backward from s's block. However, we do not go through any block that evaluates y+z. Thelast evaluation of y+z in each block encountered is an evaluation of y+z that reaches s.

• Create new variable u.

• Replace each statement w: =y+z found in (1) by

• u : = y + z w : =u

• Replace statement s by x:=u.

Some remarks about this algorithm are in order.

• The search in step(1) of the algorithm for the evaluations of y+z that reach statement s can also be formulated as a data-flow analysis problem. However, it does not make sense to solve it for all expressions y+z and all statements or blocks because too much irrelevant information is gathered.

• Not all changes made by algorithm are improvements. We might wish to limit the number of different evaluations reaching s found in step (1), probably to one.

• Algorithm will miss the fact that a*z and c*z must have the same value in

a :=x+y c :=x+y

vs

b :=a*z d :=c*z

• Because this simple approach to common sub expressions considers only the literal expressions themselves, rather than the values computed by expressions

Copy propagation: • Various algorithms introduce copy statements such as x :=copies may also be generated directly by the intermediate code generator, although most of these involve temporaries local to one block and can be removed by the dag construction. We may substitute y for x in all these places, provided the following conditions are met every such use u of x.

• Statement s must be the only definition of x reaching u.

• On every path from s to including paths that go through u several times, there are no assignments to y.

• Condition (1) can be checked using ud-changing information. We shall set up a new data-flow analysis problem in which in[B] is the set of copies s: x:=y such that every path from initial node to the beginning of B contains the statement s, and subsequent to the last occurrence of s, there are no assignments to y.

Algorithm: Copy propagation.

Input: A flow graph G, with ud-chains giving the definitions reaching block B, and with c-in[B] representing the solution to equations that is the set of copies x:=y that reach block B along every path, with no assignment to x or y following the last occurrence of x:=y on the path. We also need ud-chains giving the uses of each definition.

Output: A revised flow graph.

Method: For each copy s : x:=y do the following:

• Determine those uses of x that are reached by this definition of namely, s: x: =y.

• Determine whether for every use of x found in (1) , s is in c-in[B], where B is the block of this particular use, and moreover, no definitions of x or y occur prior to this use of x within B. Recall that if s is in c in[B]then s is the only definition of x that reaches B.

• If s meets the conditions of (2), then remove s and replace all uses of x found in (1) by y.

5.5.2   Detection of loop-invariant computations:

• Ud-chains can be used to detect those computations in a loop that are loop-invariant, that is, whose value does not change as long as control stays within the loop. Loop is a region consisting of set of blocks with a header that dominates all the other blocks, so the only way to enter the loop is through the header.

• If an assignment x := y+z is at a position in the loop where all possible definitions of y and z are

outside the loop, then y+z is loop-invariant because its value will be the same each time x:=y+z is encountered. Having recognized that value of x will not change, consider v := x+w, where w could only have been defined outside the loop, then x+w is also loop-invariant.

Algorithm: Detection of loop-invariant computations.

Input: A loop L consisting of a set of basic blocks, each block containing sequence of three - address statements. We assume ud-chains are available for the individual statements.

Output: the set of three-address statements that compute the same value each time executed, from the time control enters the loop L until control next leaves L.

Method: we shall give a rather informal specification of the algorithm, trusting that the principles will be clear.

• Mark "invariant" those statements whose operands are all either constant or have all their reaching definitions outside L.

• Repeat step (3) until at some repetition no new statements are marked "invariant".

• Mark "invariant" all those statements not previously so marked all of whose operands either are constant, have all their reaching definitions outside L, or have exactly one reaching definition, and that definition is a statement in L marked invariant.

5.5.3 Performing code motion:

• Having found the invariant statements within a loop, we can apply to some of them an optimization known as code motion, in which the statements are moved to pre-header of the loop. The following three conditions ensure that code motion does not change what the program computes. Consider s: x: =y+z.

• The block containing s dominates all exit nodes of the loop, where an exit of a loop is a node with a successor not in the loop.

• There is no other statement in the loop that assigns to x. Again, if x is a temporary assigned only once, this condition is surely satisfied and need not be changed.

• No use of x in the loop is reached by any definition of x other than s. This condition too will be satisfied, normally, if x is temporary.

Alternative code motion strategies:

• The condition (1) can be relaxed if we are willing to take the risk that we may actually increase the running time of the program a bit; of course, we never change what the program computes. The relaxed version of code motion condition (1) is that we may move a statement s assigning x only if:

1'. The block containing s either dominates all exists of the loop, or x is not used outside the loop. For example, if x is a temporary variable, we can be sure that the value will be used only in its own block.

If code motion algorithm is modified to use condition (1'), occasionally the running time will increase, but we can expect to do reasonably well on the average. The modified algorithm may move to pre-header certain computations that may not be executed in theloop. Not only does this risk slowing down the program significantly, it may also cause an error in certain circumstances.

• Even if none of the conditions of (2i), (2ii), (2iii) of code motion algorithm are met by an assignment x: =y+z, we can still take the computation y+z outside a loop. Create a new temporary t, and set t: =y+z in the pre-header. Then replace x: =y+z by x: =t in the loop. In many cases we can propagate out the copy statement x: = t.

Maintaining data-flow information after code motion: • The transformations of code motion algorithm do not change ud-chaining information, since by condition (2i), (2ii), and (2iii), all uses of the variable assigned by a moved statement s that were reached by s are still reached by s from its new position.

• Definitions of variables used by s are either outside L, in which case they reach the pre-header, or they are inside L, in which case by step (3) they were moved to pre-header ahead of s.

• If the ud-chains are represented by lists of pointers to pointers to statements, we can maintain ud-chains when we move statement s by simply changing the pointer to s when we move it. That is, we create for each statement s pointer ps, which always points to s.

• We put the pointer on each ud-chain containing s. Then, no matter where we move s, we have only to change ps , regardless of how many ud-chains s is on.

• The dominator information is changed slightly by code motion. The pre-header is now the immediate dominator of the header, and the immediate dominator of the pre-header is the node that formerly was the immediate dominator of the header. That is, the pre-header is inserted into the dominator tree as the parent of the header.

Elimination of induction variable: • A variable x is called an induction variable of a loop L if every time the variable x changes values, it is incremented or decremented by some constant. Often, an induction variable is incremented by the same constant each time around the loop, as in a loop headed by for i := 1 to 10.

• However, our methods deal with variables that are incremented or decremented zero, one, two, or more times as we go around a loop. The number of changes to an induction variable may even differ at different iterations.

• A common situation is one in which an induction variable, say i, indexes an array, and some other induction variable, say t, whose value is a linear function of i, is the actual offset used to access the array. Often, the only use made of i is in the test for loop termination. We can then get rid of i by replacing its test by one on t.

• We shall look for basic induction variables, which are those variables i whose only assignments within loop L are of the form i := i+c or i-c, where c is a constant.

5.6 OBJECT CODE GENERATION:

The final phase in our compiler model is the code generator. It takes as input an intermediate representation of the source program and produces as output an equivalent target program.
The requirements traditionally imposed on a code generator are severe. The output code must be correct and of high quality, meaning that it should make effective use of the resources of the target machine. Moreover, the code generator itself should run efficiently.
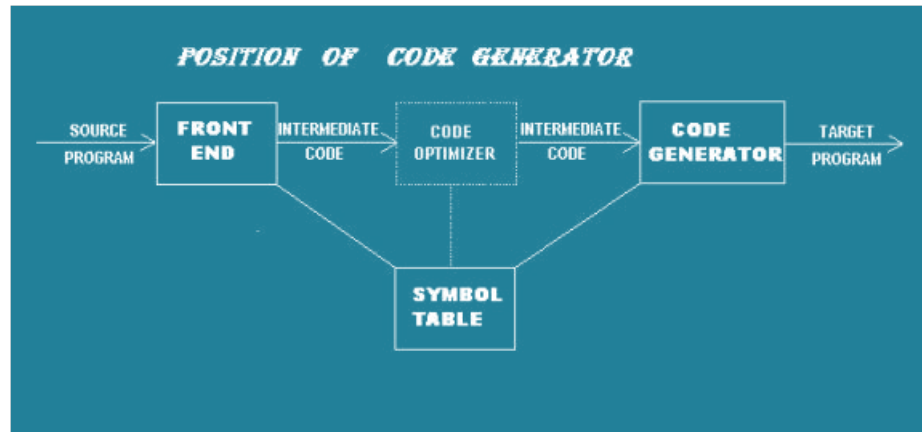
FIGURE 5.6: code generator

Issues in the design of a code generator While the details are dependent on the target language and the operating system, issues such as memory management, instruction selection, register allocation, and evaluation order are Inherent In Almost All Code Generation Problems. **Input to the code generator** The input to the code generator consists of the intermediate representation of the source program produced by the front end, together with information in the symbol table that is used to determine the run time addresses of the data objects denoted by the names in the intermediate representation.

There are several choices for the intermediate language, including: linear representations such as postfix notation, three address representations such as quadruples, virtual machine representations such as syntax trees and dags.

We assume that prior to code generation the front end has scanned, parsed, and translated the source program into a reasonably detailed intermediate representation, so the values of names appearing in the intermediate language can be represented by quantities that the target machine can directly manipulate (bits, integers, reals, pointers, etc.). We also assume that the necessary type checking has take place, so type conversion operators have been inserted wherever necessary and obvious semantic errors (e.g., attempting to index an array by a floating point number) have already been detected. The code generation phase can therefore proceed on the assumption that its input is free of errors. In some compilers, this kind of semantic checking is done together with code generation.

5.6.1 Target Programs

The output of the code generator is the target program. The output may take on a variety of forms: absolute machine language, relocatable machine language, or assembly language. Producing an absolute machine language program as output has the advantage that it can be placed in a location in memory and immediately executed. A small program can be compiled and executed quickly. A number of "student-job" compilers, such as WATFIV and PL/C, produce absolute code.

Producing a relocatable machine language program as output allows subprograms to be compiled separately. A set of relocatable object modules can be linked together and loaded for execution by a linking loader. Although we must pay the added expense of linking and loading if we produce relocatable object modules, we gain a great deal of flexibility in being able to compile subroutines separately and to call other previously compiled programs from an object module. If the target machine does not handle relocation automatically, the compiler must provide explicit relocation information to the loader to link the separately compiled program segments.

Producing an assembly language program as output makes the process of code generation some-what easier .We can generate symbolic instructions and use the macro facilities of the assembler to help generate code .The price paid is the assembly step after code generation. Because producing assembly code does not duplicate the entire task of the assembler, this choice is another reasonable alternative, especially for a machine with a small memory, where a compiler must uses several passes.

5.6.2 Memory Management

Mapping names in the source program to addresses of data objects in run time memory is done cooperatively by the front end and the code generator. We assume that a name in a three-address statement refers to a symbol table entry for the name.

If machine code is being generated, labels in three address statements have to be converted to addresses of instructions. This process is analogous to the "back patching". Suppose that labels refer to quadruple numbers in a quadruple array. As we scan each quadruple in turn we can deduce the location of the first machine instruction generated for that quadruple, simply by maintaining a count of the number of words used for the instructions generated so far.

This count can be kept in the quadruple array (in an extra field), so if a reference such as j: goto i is encountered, and i is less than j, the current quadruple number, we may simply generate a jump instruction with the target address equal to the machine location of the first instruction in the code for quadruple i. If, however, the jump is forward, so i exceeds j, we must store on a list for quadruple i the location of the first machine instruction generated for quadruple j. Then we process quadruple i, we fill in the proper machine location for all instructions that are forward jumps to i.

5.6.3 Instruction Selection

The nature of the instruction set of the target machine determines the difficulty of instruction selection. The uniformity and completeness of the instruction set are important factors. If the target machine does not support each data type in a uniform manner, then each exception to the general rule requires special handling.

Instruction speeds and machine idioms are other important factors. If we do not care about the efficiency of the target program, instruction selection is straightforward. For each type of three-address statement we can design a code skeleton that outlines the target code to be generated for that construct.

For example, every three address statement of the form x := y + z, where x, y, and z are statically allocated, can be translated into the code sequence

MOV y, R0 /* load y into register R0 */

ADD z, R0 /* add z to R0 */

MOV R0, x /* store R0 into x */

Unfortunately, this kind of statement – by - statement code generation often produces poor code. For example, the sequence of statements

a := b + c

d := a + e

Would be translated into

MOV b, R0

ADD c, R0

MOV R0, a

MOV a, R0

ADD e, R0

MOV R0, d

Here the fourth statement is redundant, and so is the third if 'a' is not subsequently used.

The quality of the generated code is determined by its speed and size.

A target machine with a rich instruction set may provide several ways of implementing a given operation. Since the cost differences between different implementations may be significant, a naive translation of the intermediate code may lead to correct, but unacceptably inefficient target code. For example if the target machine has an "increment" instruction (INC), then the three address statement a := a+1 may be implemented more efficiently by the single instruction INC a, rather than by a more obvious sequence that loads a into a register, add one to the register, and then stores the result back into a.

MOV a, R0

ADD #1,R0

MOV R0, a

Instruction speeds are needed to design good code sequence but unfortunately, accurate timing information is often difficult to obtain. Deciding which machine code sequence is best for a given

three address construct may also require knowledge about the context in which that construct appears.

5.6.4 Register Allocation

Instructions involving register operands are usually shorter and faster than those involving operands in memory. Therefore, efficient utilization of register is particularly important in generating good code. The use of registers is often subdivided into two subproblems:

1. During register allocation, we select the set of variables that will reside in registers at a point in the program.

2. During a subsequent register assignment phase, we pick the specific register that a variable will reside in.

Finding an optimal assignment of registers to variables is difficult, even with single register values. Mathematically, the problem is NP-complete. The problem is further complicated because the hardware and/or the operating system of the target machine may require that certain

Certain machines require register pairs (an even and next odd numbered register) for some operands and results. For example, in the IBM System/370 machines integer multiplication and integer division involve register pairs. The multiplication instruction is of the form

M x, y

where x, is the multiplicand, is the even register of an even/odd register pair. The multiplicand value is taken from the odd register pair. The multiplier y is a single register. The product occupies the entire even/odd register pair. The division instruction is of the form

D x, y

where the 64-bit dividend occupies an even/odd register pair whose even register is x; y represents the divisor. After division, the even register holds the remainder and the odd register the quotient.

Now consider the two three address code sequences (a) and (b) in which the only difference is the operator in the second statement. The shortest assembly sequence for (a) and (b) are given in(c). Ri stands for register i. L, ST and A stand for load, store and add respectively. The optimal choice for the register into which 'a' is to be loaded depends on what will ultimately happen to e.

t := a + b t := a + b

t := t * c t := t + c

t := t / d t := t / d

(a) (b)fig. 2 Two

three address code sequences

L R1, a L R0,

a

A R1, b A R0, b

M R0, c A R0,

c

D R0, d SRDA R0, 32

ST R1, t D R0, d

ST R1, t

(a) (b)

fig.3 Optimal machine code sequence

Choice of evaluation order The order in which computations are performed can affect the efficiency of the target code. Some computation orders require fewer registers to hold intermediate results than others. Picking a best order is another difficult, NP-complete problem. Initially, we shall avoid the problem by generating code for the three -address statements in the order in which they have been produced by the intermediate code generator.

Approaches to code generation The most important criterion for a code generator is that it produce correct code. Correctness takes on special significance because of the number of special cases that code generator must face. Given the premium on correctness, designing a code generator so it can be easily implemented, tested, and maintained is an important design goal.

Basic Blocks And Flow Graphs A graph representation of three-address statements, called a flow graph, is useful for understanding code-generation algorithms, even if the graph is not explicitly constructed by a code-generation algorithm. Nodes in the flow graph represent computations, and the edges represent the flow of control. Flow graph of a program can be used as a vehicle to collect information about the intermediate program. Some register-assignment algorithms use flow graphs to find the inner loops where a program is expected to spend most of its time.

5.6.5 Basic Blocks

A basic block is a sequence of consecutive statements in which flow of control enters at the beginning and leaves at the end without halt or possibility of branching except at the end. The following sequence of three-address statements forms a basic block:

t1 := a*a

t2 := a*b

t3 := 2*t2

t4 := t1+t3

t5 := b*b

t6 := t4+t5

A three-address statement x := y+z is said to define x and to use y or z. A name in a basic block is said to live at a given point if its value is used after that point in the program, perhaps in another basic block.

The following algorithm can be used to partition a sequence of three-address statements into basic blocks.

Algorithm 1: Partition into basic blocks.

Input: A sequence of three-address statements.

Output: A list of basic blocks with each three-address statement in exactly one block. Method:

1. We first determine the set of leaders, the first statements of basic blocks. The rules we use are the following:

I) The first statement is a leader.

II) Any statement that is the target of a conditional or unconditional goto is a leader.

III) Any statement that immediately follows a goto or conditional goto statement is a leader.

2. For each leader, its basic block consists of the leader and all statements up to but not including the next leader or the end of the program.

Example: Consider the fragment of source code shown in fig. 7; it computes the dot product of two vectors a and b of length 20. A list of three-address statements performing this computation on our target machine is shown in fig. 8.

begin

prod := 0;

i := 1;

do begin

prod := prod + a[i] * b[i];

i := i+1;

end

while i$<=$ 20

end

fig 7: program to compute dot product

Let us apply Algorithm 1 to the three-address code in fig 8 to determine its basic blocks. Statement (1) is a leader by rule (I) and statement (3) is a leader by rule (II), since the last statement can jump to it. By rule (III) the statement following (12) is a leader. Therefore, statements (1) and (2) form a basic block. The remainder of the program beginning with statement (3) forms a second basic block.

(1) prod := 0

(2) i := 1

(3) t1 := 4*i

(4) t2 := a [t1]

(5) t3 := 4*i

(6) t4 :=b [t3]

(7) t5 := t2*t4

(8) t6 := prod +t5

(9) prod := t6

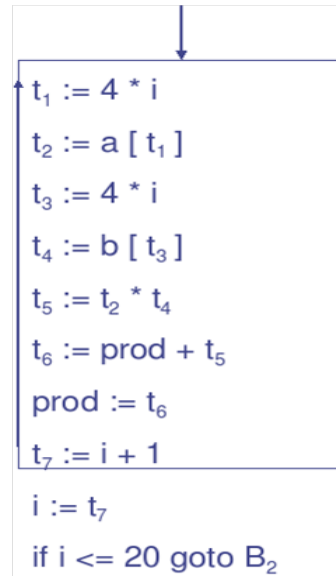(10) t7 := i+1

(11) i := t7

(12) if i¡=20 goto (3)

FIGURE 5.7: program to compute dot product

5.6.6 Transformations on basic blocks

A basic block computes a set of expressions. These expressions are the values of the names live on exit from block. Two basic blocks are said to be equivalent if they compute the same set of expressions.

A number of transformations can be applied to a basic block without changing the set of expressions computed by the block. Many of these transformations are useful for improving the quality of code that will be ultimately generated from a basic block. There are two important classes of local transformations that can be applied to basic blocks; these are the structure-preserving transformations and the algebraic transformations.

Structure-preserving transformations

The primary structure-preserving transformations on basic blocks are:

1. Common sub-expression elimination

2. dead-code elimination

3. Renaming of temporary variables

4. Interchange of two independent adjacent statements

We assume basic blocks have no arrays, pointers, or procedure calls.

5.6.6.1 Common sub-expression elimination

Consider the basic block

a:= b+c

b:= a-d

c:= b+c

d:= a-d

The second and fourth statements compute the same expression,

namely b+c-d, and hence this basic block may be transformed into the equivalent block

a:= b+c

b:= a-d

c:= b+c

d:= b

Although the 1st and 3rd statements in both cases appear to have the same expression on the right, the second statement redefines b. Therefore, the value of b in the 3rd statement is different from the value of b in the 1st, and the 1st and 3rd statements do not compute the same expression.

5.6.6.2 Dead-code elimination

Suppose x is dead, that is, never subsequently used, at the point where the statement x:= y+z appears in a basic block. Then this statement may be safely removed without changing the value of the basic block.

5.6.6.3 Renaming temporary variables

Suppose we have a statement t:= b+c, where t is a temporary. If we change this statement to u:= b+c, where u is a new temporary variable, and change all uses of this instance of t to u, then the value of the basic block is not changed. In fact, we can always transform a basic block into an equivalent block in which each statement that defines a temporary defines a new temporary. We call such a basic block a normal-form block.

5.6.6.4 Interchange of statements

Suppose we have a block with the two adjacent statements

t1:= b+c

t2:= x+y

Then we can interchange the two statements without affecting the value of the block if and only

if neither x nor y is t1 and neither b nor c is t2. A normal-form basic block permits all statement interchanges that are possible.

The target machine characteristics are

1. Byte-addressable, 4 bytes/word, n registers
2. Two operand instructions of the form
3. Op source, destination
4. Example opcodes: MOV, ADD, SUB, MULT
5. Several addressing modes
6. An instruction has an associated cost
7. Cost corresponds to length of instruction
8. Addressing Modes and Extra Costs

Mode	Form	Address	Extra cost
Absolute	M	M	1
Register	R	R	0
Indexed	k(R)	$k + contents$(R)	1
Indirect register	*R	$contents$(R)	0
Indirect indexed	*k(R)	$contents(k + contents$(R))	1

FIGURE 5.8: Addressing Modes and Extra Costs

1) Generate target code for the source language statement "(a-b) + (a-c) + (a-c);"

The 3AC for this can be written as

t := a − b

u := a − c

v := t + u

d := v + u

//d live at the end Show the code sequence generated by the simple code generation algorithm

What is its cost? Can it be improved?

Statements	Generated Code	RegDes	AdDes
		Registers empty	
t := a - b	MOV a, R0 SUB b, R0	R0 contains t	t in R0
u := a - c	MOV a, R1 SUB c, R1	R0 contains t R1 contains u	t in R0 u in R1
v := t + u	ADD R1, R0	R0 contains v R1 contains u	u in R1 v in R0
d := v + u	ADD R1, R0 MOV R0, d	R0 contains d	d in R0 d in R0 and memory

FIGURE 5.9: 1) Generate target code for the source language statement