SFU CMPT 379
Compilers
Summer 2023
Milestone 2

Assignment due Friday, July 7, by 11:59pm.

For this assignment, you are to expand your Tan-1 compiler from milestone 1 to handle Tan-2. Project setup and acceptable submissions are the same as for assignment 1 (for example, submit your java src directory).

Tan-2 is mainly backwards compatible with Tan-1. All restrictions and specifications from Tan-1 are still in force, unless specifically noted. The only part that is not backwards compatible with Tan-1 is typechecking; some operator signatures that were not accepted in Tan-1 are accepted in Tan-2 because of promotion.

Language Tan-2

Tokens:

integerLiteral → [**0..9**]$^{+}$ // has type "integer"

floatingLiteral → ([**0..9**]$^{+}$**.** [**0..9**]$^{+}$) ((**e** | **E**) (**+**|**-**) [**0..9**]$^{+}$)$^{?}$ // has type "floating"

booleanLiteral → **true** | **false** // has type "boolean"

stringLiteral → " [^"**\n**]* " // has type "string"

// In this *specification* only **\n** denotes the newline character.
// **(**tan-1 does not have character escapes in strings;
// **do not** interpret **$n**, **#n**, or **\n** in a string constant as a newline, for instance)
// In tan, only in print statements (beween arguments) is **\n** treated as newline.

characterLiteral → **'α'** |**%**[**0..7**] [**0..7**] [**0..7**] // has type "character"

// **α** is any printable ascii character (encoding is decimal 32 to 126)

identifier → [**a..zA..Z_@**][**a..zA..Z_@0..9**]*

punctuator → *operator* | *punctuation*
operator → *unaryOperator* | *arithmeticOperator* | *comparisonOperator* | *booleanOperator*
unaryOperator → **+** | **–** | **!** | **length**
arithmeticOperator → **+** | **–** | ***** | **/**
comparisonOperator → **<** | **<=** | **==** | **!=** | **>** | **>=**
booleanOperator → **&&** | **||**

punctuation → **;** | **{** | **}** | **(** | **)** | **[** | **]** | **%**| **:=** |**:**

comment → **#** [^**#\n**]* (**#** | **\n**) // starts at # ends at next # or newline.
// a comment cannot start inside a string or character.

Grammar:

S → **main** *blockStatement*
blockStatement → **{** *statement** **}**

statement → *declaration*
assignmentStatement
ifStatement
whileStatement
printStatement
blockStatement

declaration → **const** *identifier* **:=** *expression* **;** // immutable variable
var *identifier* **:=** *expression* **;** // mutable variable

assignmentStatement → *target* **:=** *expression* **;** // Reassignment. expr must be promotable to target type.
target → *expression* // must be a targetable expression.

printStatement → **print** *printExpressionList* **;** // print the expr values
printExpressionList → *printSeparator** (*expression printSeparator*+)* *expression*?
// a separated list of expressions (can be zero of them)
printSeparator → **\| \n | \s | \t**

ifStatement → **if (***expression***)** *blockStatement* (**else** *blockStatement*)? // expression must be Boolean
whileStatement → **while (***expression***)** *blockStatement* // expression must be Boolean

expression → *unaryOperator expression*
expression operator expression // all binary operations left-associative
(*expression* **)** // targetable iff the expression is.
< *type* **>(** *expression* **)** // casting
[*expression* **:** *expression* **]** // array indexing.
// first *expression* must be array type; second one integer
arrayExpression
literal

type → *primitiveType* | *arrayType*
primitiveType → **bool** | **char** | **string** | **int** | **float**
arrayType → **[** *type* **]** // an array of the given type

arrayExpression → **new** *arrayType* **(** *expression* **)**
[*expressionList* **]** // expressions must all be promotable to same type

literal → *integerLiteral* | *floatingLiteral* | *booleanLiteral* | *characterLiteral* | *stringLiteral*
identifier // identifier can be targetable.
expressionList → *expression* (**,** *expression*)*

1. Boolean expressions

The boolean-and operator **&&** has the signature (boolean, boolean) -> boolean.
The boolean-or operator **||** also has the signature (boolean, boolean) -> boolean.

Both of these operators perform short-circuit evaluation of their operands from left to right. This means that if the outcome of the operation is known after evaluating the left operand, then the right operand will not be evaluated.

The prefix unary boolean-not operator **!** has the signature (boolean) -> boolean.

2. Control flow

The control-flow statements (**if** and **while**) work as they do in most block-oriented languages (such as java and C++,) but note that they only take blockStatements as their clauses. In other words, one has to use the curly-braces **{** and **}** around the subordinate code, even if it is only one statement.

The **while (** *condition* **)** *body* loop checks the condition before entering the loop body, and may therefore execute the body zero times (if the condition is false upon statement entry).

3. Targetable expressions

An expression is called *targetable* if it may appear as the target of an assignment statement. Syntactically, an identifier is targetable and an array indexing expression ([*expression* : *expression*]) is targetable. Also, a parenthesized expression is targetable if the expression inside the parentheses is. Any other target expression generates an error.

Semantically, an identifier must be declared with **var** to be targetable. Using any other identifier as a target results in a semantic error. All array elements are targetable.

In the code generator, targetable expressions will conveniently make *address code* rather than *value code*. In C++, the concept of *lvalue* corresponds with Tan's *targetable*.

4. Type system

We now define a *type system* for Tan: a way to write down types and some rules about them. Type systems are a broad concept: they are used to write about, reason about, and specify features of languages and programs. They often contain features that are not simply the types used within a language. For instance, in Tan, the type system will include the *signatures* of the operators, which one cannot specify within the language Tan itself.

Be very careful with types. The notation described below **is not used <u>in</u> Tan programs.** It is used by people talking about Tan programs. Only where it refers to explicit type specification or Tan-syntax representation are we talking about the things a Tan programmer can use in a program.

4.1. Primitive types

Primitive types are the basic built-in types in a type system, which cannot be further decomposed.

There are five primitive types in Tan_2: the five types from Tan_1: boolean, character, floating, integer, and string. In our type system, we will denote these types using their corresponding keywords **bool**, **char**, **float**, **int**, and **string,** or, when brevity is desired, **b**, **c**, **f**, **i**, and **s**, respectively. The bold type on these notations

in the previous sentence is to make them evident in that sentence; it is not necessary to embolden them in practice.

4.2. Compound types

A compound type is a type somehow composed from another type or types. Records, arrays, and objects are examples of compound types.

Tan_2 introduces compound array types. Array is not itself a type, but **array [** *type* **]** is, for any valid type. (Here and henceforth we consider types and their denotation in the type system to be identical.) We may abbreviate array[*type*] as **a[***type***]** or simply **[***type***]**.

Thus, the following are all valid Tan-2 types:

array[bool]
array[int]
array[array[f]]
a[a[a[a[c]]]]

Although Array is not a type, we use the phrase "array type" to stand in for "some type **array[**T**]**" where **T** can be any type".

array[ ] is an example of a *type constructor*. It is an operator we can apply to types in our type system to create new types. **array[ ]** can apply to any type.

4.3. Value types and reference types

Variables in the source program are associated with memory locations in a running program. If the memory location holds a value, the variable is a *value variable*, and if it holds a reference (pointer) to where the value is, the variable is called a *reference variable*. Typically the choice between keeping a variable as a value variable or reference variable is made based on its type. Thus, we will classify types as *value types* or *reference types* if their variables are implemented as value variables or reference variables, respectively.

Primitive types are often value types, and we will mainly follow this convention in Tan. Strings are an exception, being a reference type; we have already noted this in Tan-1. The compound *array* types will be reference types. Reference variables (variables whose type is a reference type) in Tan will each consume 4 bytes, which is the size of a pointer on the ASM.

If *T* is a reference type, we will make a distinction between a *variable of type T*, which evaluates to a pointer, and a *record of type T*, which is a section of memory that (1) the pointer of a variable of type *T* points at, and (2) holds the values associated with the referenced structure. For instance, a variable of type array[float] holds a pointer, and that pointer should point at a record of type array[float], which is a hunk of memory holding a sequence of floating-point numbers (along with some extra array-control information; see below).

In java, primitive types are value types, and all objects are reference types. In C++, all types are value types, but there are compound types for pointers and references.

4.4. A signature of an operator

An n-ary operator is said to have a *signature* of

$$(type_1, type_2, \ldots type_n) \rightarrow type$$

if it accepts operands of types $type_1, type_2, \ldots type_n$ in order, and from them produces a result of type *type*. Sometimes the commas between the operand types are replaced with the Cartesian product symbol ×:

$(type_1 \times type_2 \times \ldots type_n) \rightarrow type$

in which case the parentheses are optional.

For instance, the binary operator ***** has a signature of (int, int) → int, as it accepts two integer operands and from them produces an integer result. It also has a signature of (float, float) → float.

As a further example, the unary operator **!** has a signature of (bool) → bool.

4.5. The type of operators

The *type* of an operator is the set of all of signatures that it accepts. For instance, ***** in Tan-1 has a type of

{ (int, int)→int, (float, float)→float }

We sometimes call the type of an operator the *signatures* of the operator, for obvious reasons. (Note the plural usage here is distinct from the singular usage of section 4.4).

If an operator has only one signature, we sometimes shorten the notation by omitting the set braces. For instance, the operator **!** has a type of

(bool) → bool.

Or, equivalently and more formally,

{ (bool) → bool }.

4.6. Operators with *type* operands

Tan_2 has two operators that take *type* (rather than expression) as one of their operands. These are the operators **<>()** (casting) and **new** (array creation). There are a couple ways we can notate and think about these operators.

4.6.1. Type operands as ordinary operands

The first way is to treat the operand as we do other operands, giving it an effective type of whatever type it represents. For instance, we could consider the cast **<bool>(68)** to have a boolean first operand and an integer second operand. (Note that for it to really have a boolean first operand, it would have to be something like **<true>(68)**). However, if we set the types of type expressions this way in the semantic analyzer, we can use signatures where (bool, int)->bool is a signature of casting. This is effective but not really clear.

4.6.2. Type operands as type literals

Another way is to treat the type operand as a *type literal* of its given type. This just means that the program text is literally a type. We'll use the notation **type T** for a type literal for the type **T**. Using this convention, we can speak of casting as having the signature

(type bool, int) → bool

With the example

<bool>(42)

matching that signature…this expression thus evalutates to a boolean value.

Note that we never have a type literal as the result of an operation; they can only be operands.

4.7. Type variables

Oftentimes it is more effective to write the signatures of an operator using *type variables*. Type variables are denoted using capital letters near T. A type variable is a stand-in for "any type" unless its range is restricted. For instance, the array indexing operator **[:]** has a signature of:

(array[T], int)→T

That is, it takes an array of any type as its first argument, and an integer as its second, and produces a result of the type that the array is made from. This is normal array indexing: if A is of type array[float], and you write something like "[A:4]", you get back a float.

Type variables can be used inside type literals. For instance, we may consider the casting operator **<>()** to have a signature of:

(type T, int)→T

This is because it will take an integer and an explicit type T as operands, and give a result of type T. For example,

```
<char>(68)
```

Gives the result

```
D
```

However, this is not entirely true; if T is an array type, then casting does not have that signature. That is,

```
<[bool]>(68)
```

is a semantic error. So we would have to restrict the range of the variable, and say something like:

casting has a signature of (type T, int)→T for primitive types T.

Ideally, we would like to give the type of casting as

{ (type T, S) → T }

That is, casting should take an explicitly specified type T and an expression of any type S, and convert the expression to the specified type. However, it does not do this for all pairs of types (S, T). One way to deal with this is to use the "such that" bar that is a standard part of set notation:

{ (type T, S) → T | (S, T) ∈{(int, float), (int, char), (float, int), (char, int)} }

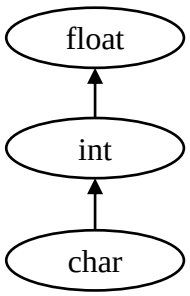
If the signatures are stated this way, then they are generally listed out (e.g. in FunctionSignatures.java). Type variables are more useful when dealing with an operator that takes any type rather than a specific subset of types.

5. Promotion: Implicit type conversions

Tan-2 introduces implicit (unstated) type conversions, called *promotions*. Promotions are performed when the source has an operator whose operands do not match any signature of the operator.

A single operand may be promoted:

(1) from **char** to **int**,
(2) from **char** to **float**,
(3) from **int** to **float**,



This gives us a *promotion digraph* with three vertices (char, int, and float) and the two directed edges (char, int) and (int, float). We don't have a (char, float) edge because we can get from char to float with the two edges. A promotion must follow a path in this digraph. With more complex languages, and perhaps in the 3rd milestone, we get a more complicated promotion digraph. Most languages have a DAG (directed acyclic graph) for the promotion digraph. If it is not a DAG, then special rules must be used for types that can cyclically promote.

Let us call a promotion *matching* if applying it gives operand types that match a signature of the operator.

Promotion is different when applied to populated array creation, which can have many arguments, than other operators, which have one or two. See the section on populated array creation for how promotion is handled there.

We consider three different promotion levels when dealing with the other operators. For unary operators, only the first two levels are used.

Promotion	**Promotion(s)**
1	None
2	One operand
3	Two operands

We proceed to check each promotion level from 1 to 3 in turn. At any level we encounter,

1) If there are two or more matches to signatures, we issue an error.
2) If there is exactly one match, then we stop checking and use that promotion (or those promotions).
3) If there are no matches, we go on to the next level; if we're on level three, there's no match overall.

As an example of 1), consider a situation where we have actuals (operands) of types **c** and **i**, and an operator with signatures **{(i, i) → i, (f, i) → f}**. Promotion of argument 1 can yield matches to these two signatures.

For the purposes of promotion, assignment (in declarations and assignment statements) is considered an operator. Only expressions are promoted. Type literals are not promoted.

To implement promotions, I recommend having a promotion **enum** that can store which of the promotions is called for (and maybe the "null promotion" which doesn't do anything). Then have either FunctionSignature or a wrapper of FunctionSignature store one of these promotions for each argument. One must be careful with the former approach to make a copy of a FunctionSignature when there is a match to it, so that the promotions it contains aren't installed in every usage of that FunctionSignature throughout the program. This is why a wrapper object *might* be a better choice; you wouldn't have to copy FunctionSignatures, but you'd have to deal with handing out and using these wrappers. Then you'd need to write code that does matching with promotions (as part of the matching code in FunctionSignatures) and generating the code for the promotions when you get an operatorNode in the code generator. I will elaborate on this in class.

6. Arrays

Arrays are a new compound type in Tan-2. The term *length,* when referring to an array, means the number of elements in the array.

6.1. Array expressions

There are two ways to create array records in Tan; these are the options for *arrayExpression*.

6.1.1. Populated array creation

The first way to create an array record is to list the members of an array between square brackets:

[*expressionList* **]**

All expressions in the *expressionList* must be promotable to the same type. If there is more than one type that they are all promotable to, then promote the least amount possible to get to a common type. Currently, this means that if they are all promotable to int and float, then promote them all to int.

If there are *n* expressions in the list, each promoted to type *T,* then this syntax creates an array record with length *n* with elements of type *T*. The values of the expressions are assigned to the elements of the array, in order.

For example, the expression

```
['s', 'a', 'i', 'd']
```

creates a 0-based array record with four characters, with the first character (index 0) being **s**, the second character (index 1) being **a**, etc.

One may not create a zero-length array with populated array creation:

```
var word = [];
```

is **not** legal Tan. (This is because we would not know the type of the array at this definition of it.)

Array elements are always mutable.

6.1.2. Empty array creation

The second way to create an array record is to give its type and length:

new *arrayType* **(** *expression* **)**

Here the expression must be of type int. This form creates an array record of the given type and length. The expression gives the length of the array, and indexing is 0-based.

For example,

```
new [char](14)
```

creates a character array record of length 14 with lower index 0 (cf. java "new char[14]").

If the length given to an empty array creation expression is negative, then a runtime error is issued. Zero-length arrays are permitted with empty array creation:

```
new [char](0)
```

Creates a zero-length array of characters with lower index 0.

Arrays created with empty array creation have the bytes of their data initialized to all zeroes.

6.2. Array variables

Arrays are reference types, so a variable of type array[T] (remember, this is type-system notation, not Tan syntax notation) is a pointer to an array record, and thus the variable occupies 4 bytes. It does ***not*** occupy 16 + length*size(T) bytes.

Array variables, and any future reference-type variables, obey semantics much like java objects (which are themselves reference types):

Array variables declared with **const** do not change their pointer; however, the elements in the record they point at may change.

```
const R := [7, 5];
[R:0] := 4;
```

is valid Tan-2 and results in R pointing at an array record that contains the elements 4 and 5.
Assignment or initialization of arrays with other arrays is simply a pointer copy.

```
const R := [7, 5];
const S := R;
```

is valid Tan-2 and results in R and S being two pointers to the same record. If this is followed by

```
[R:0] := 4;
```

then not only will [R:0] be 4, but [S:0] will, as well.

Array variables declared with **var** may change their pointer as well as the elements in the record.

```
var R := [7, 5];
[R:0] := 2;
…
R := new [int](5);
```

is valid Tan-2.

However, array variables may not be assigned an array of a different type.

```
const intArray := [7, 5];
const charArray := ['a', 'z'];

var R := intArray;
R := charArray;
```

is ***not*** valid Tan-2. The **var** declaration sets R's type as array[int], and the assignment tries to update it with an array[char], so this should generate a typechecking error.

6.3. Array indexing

The *array indexing expression* is of the form:

[*expression*$_1$ **:** *expression*$_2$ **]**

Here, *expression*$_1$ must have type array[T] for some T, and *expression*$_2$ must be of integer type. The result of this expression is the *n*-th element of the array *expression*$_1$, where *expression*$_2$ evaluates to *n*. The type of the result is T. Thus **[:]** (array indexing) has the signature **(array(T), int) → T** for any type **T**.

Whenever an array is indexed, the index must be checked for validity (i.e. it must be between 0 and the highest index in the array, inclusive). If the index is not valid, a runtime error is issued.

Array indexing expressions are targetable. In the code generator, generate *address code* for them.

6.4. Array length

The *length expression* is of the form:

length *expression*

The *expression* must have array type, and the result is the integer length (number of elements) of the array. Length thus has the signature **array(T) → int** for any type T.

6.5. Array printing

If an expression of array type is in the expressionList of a printing statement, then the array is printed as the following sequence:

[
A print of the first element
,

a space
a print of the next element
,
a space
…

A print of the last element
]

This is a quite natural way to print an array. For example, the array [1.23, 2.79, 5.41] is printed as

```
[1.23, 2.79, 5.41]
```

Note that the only spaces printed are single spaces after after each comma.
The "print of the *n*th element" parts are done recursively. It is your problem to figure out when that recursion is done—at compile time or at run time.

6.6. Multidimensional arrays

Tan does not have true multidimensional arrays. Instead (like java), one must use arrays of arrays. An array with array elements must have all of those array elements of the same type, but they need not be of the same size. For instance, the following is legal:

```
const numSets := [ [1, 2, 3], [4, 5], [6, 7, 8], new [int](0)];
```

This makes numSets have type array[array[int]], with four elements. For empty array creation, the following is the proper idiom for a rectangular array:

```
const width := 4;
const height := 7;
const matrix := new [[float]](width);

var x := 0;
while(x < width) {
      [matrix:x] := new [float](height);
      x := x + 1;
}
```

6.7. Array casting

Array types may not be cast to other types, including other array types. No other type may be cast to an array type, although array types may be cast to themselves.

7. Records

In a running Tan program, we will often have many pieces of heap memory that we need to process and keep track of. In order to do this, we will enforce a simple record format, as follows:

Type identifier (4 bytes)	Status (4 bytes)	Rest of record (?? bytes)

The type identifier is an integer describing what type is being stored in this record. Currently there are only two "types" with records: **string** and **array( )**. (Remember that **array( )** is not by itself a type.) String is given type identifier 3, and array is given type identifier 5.

The status field currently holds only four bits of data, in the lowest bits.

- The datum in bit 0 is the *immutability status* of the elements of the record (1 is immutable, 0 is mutable).
- The datum in bit 1 is the *subtype-is-reference status* of the array; this indicates whether the subtype T of the array is a reference type (i.e. if the subtype is itself an array or string type)
- The datum in bit 2 is the *is-deleted status* of the record, which indicates that this record has been given to the memory deallocator.
- The datum in bit 3 is the *permanent status* of the record. Records with this bit set won't be deallocated. If an attempt is made to deallocate a permanent record, it is silently ignored. (It doesn't issue a runtime error).

7.1. Array records

An *arrayExpression* results in the allocation of a new block of memory—a record—from the heap at runtime. The record for the type array[T] (informally, an *array record*) has the following format:

Type identifier (4 bytes)	Status (4 bytes)	Subtype size (4 bytes)	Length (4 bytes)	Elements ((subtypeSize * length) bytes)

The *type identifier* for an array is the integer 5. When creating an array, set

- The *immutability status* to 0. (Array elements are mutable.)
- The *subtype-is-reference status* according to the subtype of the array.
- The *is-deleted status* to 0.
- The *permanent status* to 0. Array records can be deleted.

The subtype size is the number of bytes consumed by a variable of type T; we will refer to this as size(T). The *length* is the number of elements in the array. (Note that the highest index in the array is *length*-1).

Over the lifetime of this record, only the elements and possibly the status flags may change. The other values will not.

To create an Array record, you will need to call the memory manager. This means you must enable the memory manager. This involves uncommenting one line (in makeASM( ) in the code generator), and adding another line somewhere. Look at the public methods in MemoryManager.java to figure out what the added line should be (and then you must decide exactly where it should go). You do not need to understand the MemoryManager in detail; you only need to figure out its API.

8. Operator precedence

The precedence of operators is

Highest precedence	parentheses, populated array creation empty array creation casting array indexing	**( ) [ ] new[]() <>[] [:]**

(prefix unary operators are right-associative)	prefix unary operators	**+ - ! length**
	multiplicative operators	***** /
	additive operators	**+ –**
	comparisons	**< > <= >= == !=**
	and	**&&**
Lowest precedence	or	||

These are all left-associative operators, except as noted.