Main topics of the week:

- Variables
- Type Checking
- Storage types
- Scope

Variables

Although imperative languages put the focus on actions as reflected in the control flow, programs would not be very interesting without data. The data in a program is accessed through symbols known as variables. Particularly in an imperative language where assignment allows data values to be changed, the term variable is appropriate. Variables can be characterized by the properties: name, address, type, and value.

The name of a variable is simply how we refer to it in the program source. Each language has rules as to the permissible names that may be used as identifiers of variables, e.g., letters and digits, can’t begin with a digit, etc. The syntax of the language specifies these rules.

The address of the variable is its location in memory. Some languages such as C/C++ allow the programmer to know the address of a variable, and some languages do not provide any mechanism to get at a variable’s address – the mapping of variables to addresses is kept completely hidden from the programmer. Whether the address can be seen or not does not affect the fact that the variable is associated with some storage in memory and it is often useful to think about that storage even if we cannot gain direct access to it. When two variables use the same address, we say that the variable names are aliases. In this case, since the same location can be referred to by different names, actions that change the value could be through either of the names. This certainly complicates program analysis, but can be useful for certain programming paradigms. Although Java does not allow addresses to be seen, all non-primitive objects in Java are references and so aliases can occur there. In C/C++, where there is an address operator and pointer types, aliases are directly available.

The type of a variable characterizes the use and interpretation of the data stored. Most languages define some basic data types such as integer, character, floating point, and maybe Boolean. A type specifies a data size, interpretation of the bits and value ranges, and valid operations that can be performed with the type. Some languages may also define a primitive type of string. These primitive types are built in to the language. In C and Java, we have these basic types and keywords to declare them.

In addition to primitive types, languages may provide derived types. In C, there are derived types of pointers (for every variable type, we can have a variable which is a pointer to that type). C also permits arrays of any type. Unfortunately, the treatment of arrays in C is fairly simplistic. Arrays are essentially used to refer to contiguous chunks of memory, with the language aiding in viewing the memory as a sequence of objects. However, C provides little support beyond this. An array has no inherent size – in a sense the compiler forgets about the size as soon as the space is allocated. This means there can be no built in bounds checking or understanding of whether two array objects are the same size. It also means arrays cannot be easily passed by value, although there are some...
tricks to get around this. In addition to derived types, languages often provide user defined types. This is basically a mechanism for grouping data together. We’ll see in object oriented programming that user defined data there gets all the focus. In imperative languages, the aggregate data types give structure to the data and allow the user to view the groups of data as a single record type.

Finally, a variable has a value. The value is just the contents of the memory associated with the variable, interpreted properly according to its type. The way that we usually think about a variable is in terms of its value, typically abstracting away the idea of the physical storage as a collection of bits and viewing the variable as an integer, or a floating point number, or as a record.

Lvalue versus Rvalue. We have seen that an important characteristic of imperative languages is assignment – the ability to change a value of a variable in memory. The assignment expression in C is a binary operator ‘=’, using infix notation, with the left hand side being the object to be changed and the right hand side being the new value. The semantics of an assignment construct thus make an important distinction between the operands. In particular, the left operand must be an “assignable” object – typically a variable name, but more generally a reference to a memory location that is to be changed. The right operand on the other hand is just a value, again commonly a variable, but more generally a value obtained as the evaluation of a complicated expression. This distinction is where we get the terms lvalue and rvalue. For example, the C compiler may produce an error message like “invalid lvalue in assignment” when it sees a statement like i+j = i; This is because syntactically, the object to be assigned to is i+j, but that expression does not refer to a changeable memory location. Certainly any variable (non-constant) could appear as an lvalue, but there are other less obvious lvalues. The contents of a pointer variable, an offset into an array, or even a function return value de-referenced to a member of a structure. That is, the left operand of an assignment could be any expression, as long as that expression evaluates to a memory location that can be assigned to.

Although these rules are consistent, they can make static analysis of an imperative program very hard or impossible if we want to determine everything that can change by a given assignment. Also, a simple variable may appear as either a left or right operand in an assignment; in this case, it is viewed in one context as an lvalue and another context as an rvalue. Actually, when an lvalue is required, we use the location of the variable, and when an rvalue is required, we use the value of the variable. This could be confusing since the variable as it appears in the statement looks the same on either side – its context is what determines the interpretation. (We’ll see in ML that explicit notation is required to make this distinction.)

Type Checking. As soon as we introduce types into a language we come to the issue of how much attention the compiler or interpreter should pay to types. Different languages have different degrees of type sensitivity: we refer to languages from “not typed”, to typed, to weakly typed, to strongly typed. What this means is how much the language enforces type agreement among operands in expressions. C is a typed language, but certainly not strongly typed. C often defaults a missing type to integer, and implements no type checking on function arguments (even worse, it doesn’t check function argument counts – a form of checking the type of the function). C++ and Java
are much more strict about function arguments and less tolerant of implicit typing. However, all of these languages allow some mixing of types, particularly for numerical types where the mixing does not result in any loss of information. For example, in expressions involving integers and floating point numbers, the integers are implicitly converted to floating point for the operation: this is called coercion. Compilers typically warn or prohibit assignment of float to int since this would result in loss of information. These conversions are a great convenience to the programmer in most cases, but can occasionally produce unexpected results or compiler/platform dependencies. (We'll see later in ML how annoyingly explicit we must be about types, but there are at least no surprises.)

In a typed language, one of the most important job for the compiler/interpreter is the type checking. The compiler/interpreter “thinks” of all of the objects in the program in terms of their type. The language will specify what types are allowed to be combined in expressions and whether any type coercion will take place. Type checking in a compiled language is a static analysis that is carried out at compile time. It is based on information about types in the program source (e.g., declarations). Compiled languages like C can only implement static type checking, so it may be that not all type errors are caught if the program uses dynamic typing. Formally, there is no dynamic typing in C, but programs may use type casts at run time. What is really going on with a type cast in this setting is that the type is actually unknown or known to be incorrect from the compiler’s static analysis point of view, and the programmer is getting the compiler to accept on faith the programmer’s knowledge about the types and to possibly do some conversion of bits to the indicated type. By the very nature of the type cast, errors cannot be caught by the compiler.

When dealing with types, the issue arises as to what are equivalent types. Certainly two variables declared together should be of the same type. What if the declarations are separate, but use the same syntax and type names? In this case, we would say they are equivalent. We could formally define structural equivalence as types being structurally equivalent to themselves (the trivial case), or formed by applying the same type constructor, or formed by a type definition of one to the other (e.g., a typedef in C). Thus, we may have different type names, but if they are essentially synonyms, they are the same type. Equivalence could be more restrictive, where the names of the types must be identical. In C and C++, structural equivalence is used, except for user defined structures, which must be name equivalent.

Some languages (like Pascal and C/C++) allow pointer types. This is a way of allowing direct access to memory contents. Operators are provided to allow the language to specify that the memory address of a variable be calculated and used in an expression. Similarly, another operator allows the value stored at a location to be obtained. These operations mean that a memory location’s contents can be manipulated in an expression without ever having seeing the name of the variable that “owns” the memory. The power of pointers is that the level of indirection allows more dynamic handling of memory contents. Also, as we’ll see in procedure parameter passing, pointers allow a call by value language to implement call by reference. Pointers can also allow efficient handling of dynamically sized arrays of objects. The downside is that the compiler cannot statically analyze what will happen at runtime.
C++ allows a `const` qualifier to be used in any variable declaration (and Java permits the `final` qualifier). The use of such a qualifier extends the type information available to the compiler. For a simple variable declaration that is marked constant, the compiler now knows that it is impermissible to modify the value of the variable. This could occur through direct assignment of course, but also less obviously in an array or more complicated expression. In C++, where pointers are possible, the compiler has to also be on the lookout for indirect assignments that could happen through the use of pointers. Although this might seem an impossible task for the compiler, it actually is possible for the compiler to keep track of when the value is modified as long as it includes the constant aspect as part of the type and carefully retains that constantness when it sees expressions using the address of the variable. For this reason, an array can be declared constant, and a pointer to an element of the array can be incremented or decremented and the constantness retained. However, if the pointer is incremented beyond the bounds of the array, then even though the constantness is retained, it is meaningless, as the pointer now refers to memory that may or may not be of the original array element’s type. The appropriate use of `const` in C++ allows the programmer to get the advantages of pointers in terms of performance and syntax efficiency, but avoid many of the common problems of unexpected side effects. This construct permits the expression of more design intent in the language itself.

**Scope.** The variables in a language are the identifiers we use to refer to stored data values. We have seen that blocks are used for control flow in an imperative language, but blocks also serve another very important purpose. They define a locale – an area of the code that is like a small environment within the larger program. Thus, variable declarations may be placed within a block, and the block determines the scope of the variable. The scope of the variable refers to where in the code the variable name may be used. It is a property of static analysis of the program. The basic scope rule in C/C++/Java is that a variable may be used any place in the code from the point where it is declared to the end of the block in which it is declared. C and C++ allow variables to have global scope (Java does not). This means that the variable can be referred to at any place in the code after its point of declaration. (Most languages’ scope rules have a sense of direction – forward toward the end of the file from the point of declaration.) A variable has local scope, also said to be a local variable, if it is declared within a block. We don’t want to confuse the scope of a variable with the storage for the variable. The scope simply refers to where the symbol can legally appear – it is part of the semantic analysis of the code since knowing if a variable is legal means finding its declaration in an appropriate place, and this is beyond the ability of a grammar to express. When scope is determined in this way by statically analyzing the source code without any notion of what the thread of control will look like for a particular execution, it is called static scope (or lexical scope). It is possible for a language to implement dynamic scope (Lisp does this), but we’ll see later that it can be unintuitive, and is not typically done. However, the dynamic binding that is done with virtual functions in C++ and interfaces in Java does capture some of the idea of dynamic scope in a predictable and useful way.

One issue that comes up with the notion of local variables is what the language should specify if a variable is declared a second time in the same scope. In C/C++/Java, this is considered to be an error. However, if a variable is re-declared in a nested scope,
that is acceptable in these languages. But the next question is about the semantics of such a re-declaration. First, it must be of the same type, and second, the compiler arranges for a new instance of a variable, and all subsequent references will refer to the variable instance in the closest block. In effect, this will hide the earlier declarations.

The following C++ code fragment shows these properties of variables.

```
int i;

main() {
    int i;
    ...
    if (i > 0) {
        int i;
        for (int i = 0; i < 10; ++i) {
            int j = i;
            ...
        }
        int j = i;
        ...
    }
    int j = i;
    ...
}
int j = i;

while (j-- > 0) {
    int i;
    ...
}
  ...
  j = i;
}
```

Scope often seems related to, but should not be confused with storage. The storage determines the lifetime of the variable, i.e., when during the program execution the storage exists and is valid. Remember that the scope only refers to where it is legal in the code to mention the variable’s name. There are three main types of storage that may be used for a variable:

- **Static storage** is space created for the variable prior to the start of execution. In C/C++ and Java, the keyword `static` is used to modify a declaration to indicate this type of storage. Static variables are around for the entire execution life of a program. For class member data, static means there is only one instance of the data, regardless of how many instances of the class object exist (even zero). In a sense, static class data is global to the class objects. The way to think of static variables is that the compiler arranges for the execution image to have some space for these variables prior to beginning execution.
**Stack storage**, or **automatic storage** is used for local variables. C/C++ actually have a keyword `auto` to indicate this storage, but since it is the default, you almost never see it used. Storage for such variables is automatically allocated as part of the stack. It is usually space in an activation frame that is placed on the stack for a block (procedure or just control block) and this is placed on the stack at execution time at the point in time where the block or procedure is entered (becomes active). At exit from the block or procedure, the activation frame is taken off the stack. So the lifetime of automatic variables is similar to their scope, except that there may be holes in the scope where the variable is hidden by a nested scope, but the variable’s storage is still around.

**Heap storage** or **dynamic storage** is storage explicitly requested during the execution of a program according to instructions coded by the programmer. The storage is in an area of memory used by the heap manager, which allocates, frees, and coalesces chunks of memory. This storage is obtained in C code by a call to `malloc`, which is a standard C library routine. In C++ code, the C++ operator `new` is used to allocate space for objects of a given type (and automatically calls the object’s constructor, too). In Java, the operator `new` is used to create an object as well. Objects on the heap have a lifetime that is automatically determined in Java to be as long as they are needed, i.e., as long as they are being referenced. When an object can no longer be referenced, the Java VM will make it available for garbage collection, which is periodically run to recover space that is no longer needed. In C/C++, objects on heap must be explicitly freed. C++ makes this easier with destructors that get called when an object goes out of scope, but it has no garbage collection for everything. Rather, the paradigm in C++ is that you can implement your own garbage collection (as well as heap allocation) that is tailored to the needs of the class. This allows for customization in the event that the one size fits all allocation/de-allocation strategy of Java is not suitable for every application’s needs.