CSCI 3336 Organization of Programming Languages

DATA TYPES

Topics

- Introduction
- Primitive Data Types
- Character String Types
- User-Defined Ordinal Types
- Array Types
- Associative Arrays
- Record Types
- Union Types
- Pointer and Reference Types

Introduction

- A *data type* defines a collection of data objects and a set of predefined operations on those objects

- A *descriptor* is the collection of the attributes of a variable

- An *object* represents an instance of a user–defined (abstract data) type

- One design issue for all data types: What operations are defined and how are they specified?
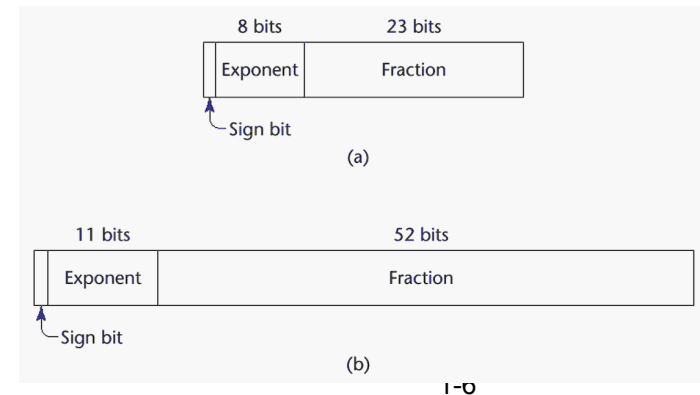
Primitive Data Types

- Almost all programming languages provide a set of *primitive data types*
- Primitive data types: Those not defined in terms of other data types
- Some primitive data types are merely reflections of the hardware
- Others require only a little non-hardware support for their implementation

Primitive Data Types: Integer

- Almost always an exact reflection of the hardware so the mapping is trivial

- There may be as many as eight different integer types in a language

- Java's signed integer sizes: `byte, short, int, long`

Primitive Data Types: Floating Point

- Model real numbers, but only as approximations
- Languages for scientific use support at least two floating-point types (e.g., `float` and `double`; sometimes more
- Usually exactly like the hardware, but not always
- IEEE Floating-Point Standard 754

8 bits   23 bits

Exponent	Fraction

Sign bit

(a)

11 bits   52 bits

Exponent	Fraction

Sign bit

(b)

Primitive Data Types: Complex

- Some languages support a complex type, e.g., C99, Fortran, and Python
- Each value consists of two floats, the real part and the imaginary part
- Literal form (in Python):

    `(7 + 3j)`, where `7` is the real part and `3` is the imaginary part

Primitive Data Types: Decimal

- For business applications (money)
  - Essential to COBOL
  - C# offers a decimal data type
- Store a fixed number of decimal digits, in coded form (BCD)
- *Advantage*: accuracy
- *Disadvantages*: limited range, wastes memory

Primitive Data Types: Boolean

- Simplest of all
- Range of values: two elements, one for "true" and one for "false"
- Could be implemented as bits, but often as bytes
  - Advantage: readability

Primitive Data Types: Character

- Stored as numeric codings
- Most commonly used coding: ASCII
- An alternative, 16-bit coding: Unicode (UCS-2)
  - Includes characters from most natural languages
  - Originally used in Java
  - C# and JavaScript also support Unicode
- 32-bit Unicode (UCS-4)
  - Supported by Fortran, starting with 2003

Character String Types

- Values are sequences of characters
- Design issues:
  - Is it a primitive type or just a special kind of array?
  - Should the length of strings be static or dynamic?

Character String Types Operations

- Typical operations:
  - Assignment and copying
  - Comparison (=, >, etc.)
  - Catenation
  - Substring reference
  - Pattern matching

Character String Type in Certain Languages

- C and C++
  - Not primitive
  - Use `char` arrays and a library of functions that provide operations
- SNOBOL4 (a string manipulation language)
  - Primitive
  - Many operations, including elaborate pattern matching
- Fortran and Python
  - Primitive type with assignment and several operations
- Java
  - Primitive via the `String` class
- Perl, JavaScript, Ruby, and PHP
  - Provide built-in pattern matching, using regular expressions

Character String Length Options

- Static: COBOL, Java's `String` class
- *Limited Dynamic Length*: C and C++
  - In these languages, a special character is used to indicate the end of a string's characters, rather than maintaining the length
- *Dynamic* (no maximum): SNOBOL4, Perl, JavaScript
- Ada supports all three string length options

Character String Type Evaluation

- As a primitive type with static length, they are inexpensive to provide--why not have them?

- Dynamic length is nice, but is it worth the expense?

Character String Implementation

- Static length: compile-time descriptor
- Limited dynamic length: may need a run-time descriptor for length (but not in C and C++)

- Dynamic length: need run-time descriptor; allocation/de-allocation is the biggest implementation problem

Compile– and Run–Time Descriptors

Static string
Length
Address

Compile–time descriptor for static strings

Limited dynamic string
Maximum length
Current length
Address

Run–time descriptor for limited dynamic strings

User–Defined Ordinal Types

- An ordinal type is one in which the range of possible values can be easily associated with the set of positive integers
- Examples of primitive ordinal types in Java
  - `integer`
  - `char`
  - `boolean`

Enumeration Types

- All possible values, which are named constants, are provided in the definition
- C# example

```
enum days {mon, tue, wed, thu, fri, sat, sun};
```

- Design issues
  - Is an enumeration constant allowed to appear in more than one type definition, and if so, how is the type of an occurrence of that constant checked?
  - Are enumeration values coerced to integer?
  - Any other type coerced to an enumeration type?

Evaluation of Enumerated Type

- Aid to readability, e.g., no need to code a color as a number

- Aid to reliability, e.g., compiler can check:
  - operations (don't allow colors to be added)
  - No enumeration variable can be assigned a value outside its defined range
  - Ada, C#, and Java 5.0 provide better support for enumeration than C++ because enumeration type variables in these languages are not coerced into integer types

Subrange Types

- An ordered contiguous subsequence of an ordinal type
  - Example: 12..18 is a subrange of integer type
- Ada's design

```
type Days is (mon, tue, wed, thu, fri, sat, sun);
subtype Weekdays is Days range mon..fri;
subtype Index is Integer range 1..100;

Day1: Days;
Day2: Weekday;
Day2 := Day1;
```

Subrange Evaluation

- ## Aid to readability
  - Make it clear to the readers that variables of subrange can store only certain range of values
- ## Reliability
  - Assigning a value to a subrange variable that is outside the specified range is detected as an error

Array Types

- An array is an aggregate of homogeneous data elements in which an individual element is identified by its position in the aggregate, relative to the first element.

Array Design Issues

- What types are legal for subscripts?
- Are subscripting expressions in element references range checked?
- When are subscript ranges bound?
- When does allocation take place?
- What is the maximum number of subscripts?
- Can array objects be initialized?
- Are any kind of slices supported?

Array Indexing

- *Indexing* (or subscripting) is a mapping from indices to elements

  `array_name (index_value_list)` → `an element`

- Index Syntax
  - FORTRAN, PL/I, Ada use parentheses
    - Ada explicitly uses parentheses to show uniformity between array references and function calls because both are *mappings*
  - Most other languages use brackets

Arrays Index (Subscript) Types

- FORTRAN, C: integer only
- Ada: integer or enumeration (includes Boolean and char)
- Java: integer types only
- Index range checking
  - C, C++, Perl, and Fortran do not specify range checking
  - Java, ML, C# specify range checking
  - In Ada, the default is to require range checking, but it can be turned off

Subscript Binding and Array Categories

- *Static*: subscript ranges are statically bound and storage allocation is static (before run-time)
  - Advantage: efficiency (no dynamic allocation)
  - Disadvantage: Storage of array is fixed
  - Arrays that are declared in C and C++ functions that include the **static** modifier.

Subscript Binding and Array Categories

- *Fixed stack-dynamic*: subscript ranges are statically bound, but the allocation is done at declaration time during execution.
  - Advantage: space efficiency
  - Disadvantages: Require allocation and deallocation time.
  - Arrays that are declared in C and C++ without the static modifier.

Subscript Binding and Array Categories (continued)

- *Stack-dynamic*: subscript ranges and the storage allocation is dynamic (done at run-time)
  - Advantage: flexibility (the size of an array need not be known until the array is to be used)
  - Disadvantage: Once the subscript ranges are bound and storage is allocated they remain fixed
  - ADA arrays can be stack-dynamic
    ```
    Get (List_Len)
    declare
            List : array (1 .. List_Len) of Integer;
            begin
            …
            end;
    ```

Subscript Binding and Array Categories (continued)

- *Fixed heap-dynamic*: similar to fixed stack-dynamic: storage binding is dynamic but fixed after allocation (i.e., binding is done when requested and storage is allocated from heap, not stack)
  - Advantage: Flexibility – array's size always fits the problem
  - Disadvantage: Allocation time from the heap is longer than from stack.
  - C and C++ provide fixed heap-dynamic arrays – malloc and free in C and new and delete in C++
  - In Java all non-generic arrays
  - C# also provides the same kind of arrays

Subscript Binding and Array Categories (continued)

- Heap-dynamic: binding of subscript ranges and storage allocation is dynamic and can change any number of times
  - Advantage: flexibility (arrays can grow or shrink during program execution)
  - Disadvantages: Allocation and deallocation takes longer – and it may happen many times during the execution of the program.
  - ArrayList in Java
  - Perl, Python, Ruby, and Javascript

Array Initialization

- Some language allow initialization at the time of storage allocation
  - C, C++, Java, C# example

  ```
  int list [] = {4, 5, 7, 83}
  ```

  - Character strings in C and C++

  ```
  char name [] = "freddie";
  ```

  - Arrays of strings in C and C++

  ```
  char *names [] = {"Bob", "Jake", "Joe"];
  ```

  - Java initialization of String objects

  ```
  String[] names = {"Bob", "Jake", "Joe"};
  ```

Heterogeneous Arrays

- A *heterogeneous array* is one in which the elements need not be of the same type
- Supported by Perl, Python, JavaScript, and Ruby

Array Initialization

- ## C–based languages
  - `int list [] = {1, 3, 5, 7}`
  - `char *names [] = {"Mike", "Fred","Mary Lou"};`

- ## Ada
  - `List : array (1..5) of Integer :=`
    `    (1 => 17, 3 => 34, others => 0);`

- ## Python
  - `List comprehensions`
    `list = [x ** 2 for x in range(12) if x % 3 == 0]`
    `    puts [0, 9, 36, 81] in list`

Arrays Operations

- APL provides the most powerful array processing operations for vectors and matrixes as well as unary operators (for example, to reverse column elements)
- Ada allows array assignment but also catenation
- Python's array assignments, but they are only reference changes. Python also supports array catenation and element membership operations
- Ruby also provides array catenation
- Fortran provides *elemental* operations because they are between pairs of array elements
  - For example, + operator between two arrays results in an array of the sums of the element pairs of the two arrays

Rectangular and Jagged Arrays

- A rectangular array is a multi-dimensioned array in which all of the rows have the same number of elements and all columns have the same number of elements

- A jagged matrix has rows with varying number of elements
  - Possible when multi-dimensioned arrays actually appear as arrays of arrays

- C, C++, and Java support jagged arrays

- Fortran, Ada, and C# support rectangular arrays (C# also supports jagged arrays)

Slices

- A slice is some substructure of an array; nothing more than a referencing mechanism

- Slices are only useful in languages that have array operations

Slice Examples

- Fortran 95

  ```
  Integer, Dimension (10) :: Vector
  Integer, Dimension (3, 3) :: Mat
  Integer, Dimension (3, 3) :: Cube
  ```
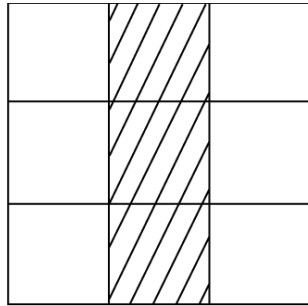
  `Vector (3:6)` is a four element array

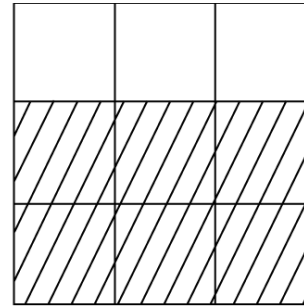- Ruby supports slices with the `slice` method

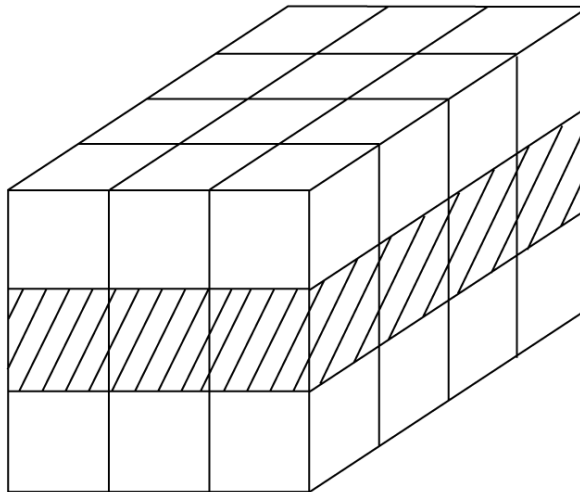 `list.slice(2, 2)` returns the third and fourth elements of `list`

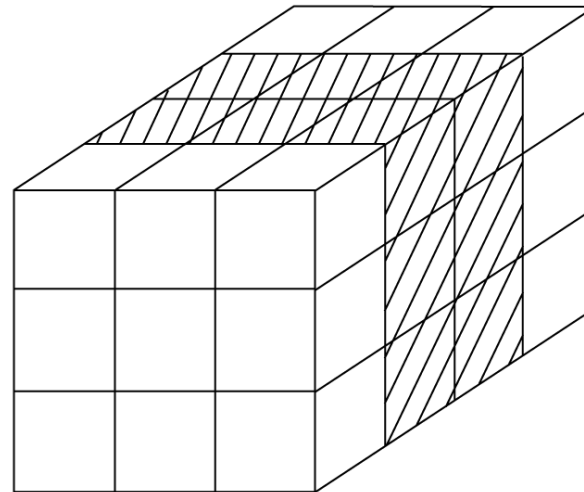# Slices Examples in Fortran 95

MAT (1:3, 2)

MAT (2:3, 1:3)

CUBE (2, 1:3, 1:4)

CUBE (1:3, 1:3, 2:3)

Implementation of Arrays

- Access function maps subscript expressions to an address in the array
- Access function for single-dimensioned arrays:

address(list[k]) = address (list[lower_bound])

+ ((k-lower_bound) * element_size)

Accessing Multi–dimensioned Arrays

- ## Two common ways:
    - Row major order (by rows) – used in most languages
    - column major order (by columns) – used in Fortran

Locating an Element in a Multi-dimensioned Array

- General format
  Location (a[i,j]) = address of a [row_lb,col_lb] +
  (((i – row_lb) * n) + (j – col_lb)) * element_size

Compile–Time Descriptors

Array
Element type
Index type
Index lower bound
Index upper bound
Address

Multidimensioned array
Element type
Index type
Number of dimensions
Index range 1
⋮
Index range n
Address

Single-dimensioned array   Multi-dimensional array

Associative Arrays

- An *associative array* is an unordered collection of data elements that are indexed by an equal number of values called *keys*

  - User-defined keys must be stored

- Design issues:

  - What is the form of references to elements?
  - Is the size static or dynamic?

- Built-in type in Perl, Python, Ruby, and Lua

  - In Lua, they are supported by tables

Associative Arrays in Perl

- **Names begin with `%`; `literals` are delimited by parentheses**

  ```
  %hi_temps = ("Mon" => 77, "Tue" => 79,
     "Wed" => 65, …);
  ```

- **Subscripting is done using braces and keys**

  ```
  $hi_temps{"Wed"} = 83;
  ```

  - **Elements can be removed with `delete`**

    ```
    delete $hi_temps{"Tue"};
    ```

Record Types

- A *record* is a possibly heterogeneous aggregate of data elements in which the individual elements are identified by names

- Design issues:

  - What is the syntactic form of references to the field?

  - Are elliptical references allowed

Definition of Records in COBOL

- COBOL uses level numbers to show nested records; others use recursive definition

```
01 EMP-REC.
    02 EMP-NAME.
        05 FIRST PIC X(20).
        05 MID   PIC X(10).
        05 LAST  PIC X(20).
    02 HOURLY-RATE PIC 99V99.
```

Definition of Records in Ada

- **Record**

```
type Emp_Rec_Type is record
    First: String (1..20);
    Mid: String (1..10);
    Last: String (1..20);
    Hourly_Rate: Float;
end record;
Emp_Rec: Emp_Rec_Type;
```

References to Records

- Record field references
  1. COBOL
  field_name OF record_name_1 OF ... OF record_name_n
  2. Others (dot notation)
  record_name_1.record_name_2. ... record_name_n.field_name

- Fully qualified references must include all record names

- Elliptical references allow leaving out record names as long as the reference is unambiguous, for example in COBOL

  FIRST, FIRST OF EMP-NAME, and FIRST OF EMP-REC are elliptical references to the employee's first name

Operations on Records

- Assignment is very common if the types are identical

- Ada allows record comparison

- Ada records can be initialized with aggregate literals

- COBOL provides `MOVE CORRESPONDING`
  - Copies a field of the source record to the corresponding field in the target record

Evaluation and Comparison to Arrays

- Records are used when collection of data values is heterogeneous
- Access to array elements is much slower than access to record fields, because subscripts are dynamic (field names are static)
- Dynamic subscripts could be used with record field access, but it would disallow type checking and it would be much slower

Implementation of Record Type

Offset address relative to the beginning of the records is associated with each field
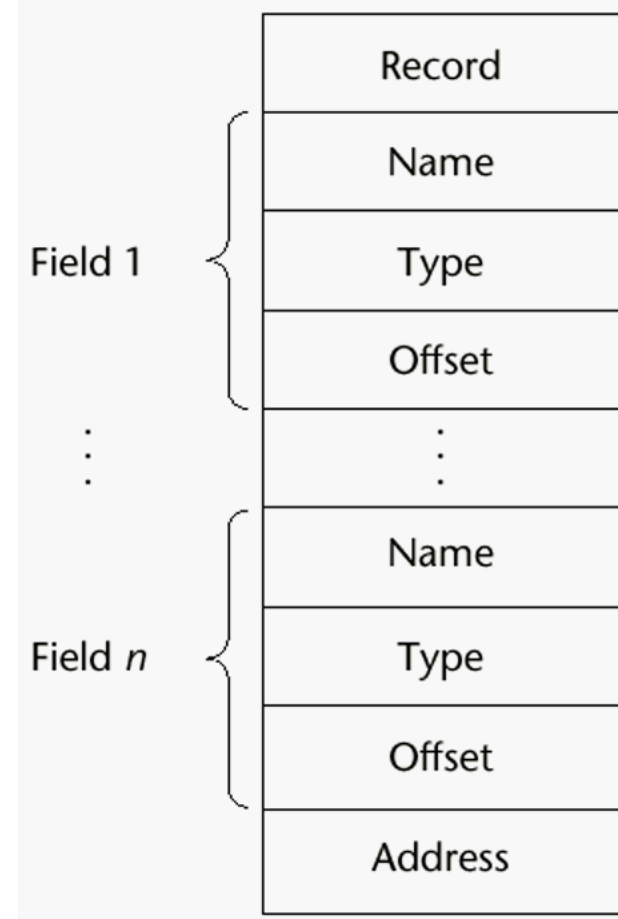
Pointer and Reference Types

- A *pointer* type variable has a range of values that consists of memory addresses and a special value, *nil*
- Provide the power of indirect addressing
- Provide a way to manage dynamic memory
- A pointer can be used to access a location in the area where storage is dynamically created (usually called a *heap*)

Design Issues of Pointers

- What are the scope of and lifetime of a pointer variable?
- What is the lifetime of a heap-dynamic variable?
- Are pointers restricted as to the type of value to which they can point?
- Are pointers used for dynamic storage management, indirect addressing, or both?
- Should the language support pointer types, reference types, or both?

Pointer Operations

- Two fundamental operations: assignment and dereferencing
- Assignment is used to set a pointer variable's value to some useful address
- Dereferencing yields the value stored at the location represented by the pointer's value
  - Dereferencing can be explicit or implicit
  - C++ uses an explicit operation via *

    ```
    j = *ptr
    ```
    sets j to the value located at `ptr`

Pointer Assignment Illustrated
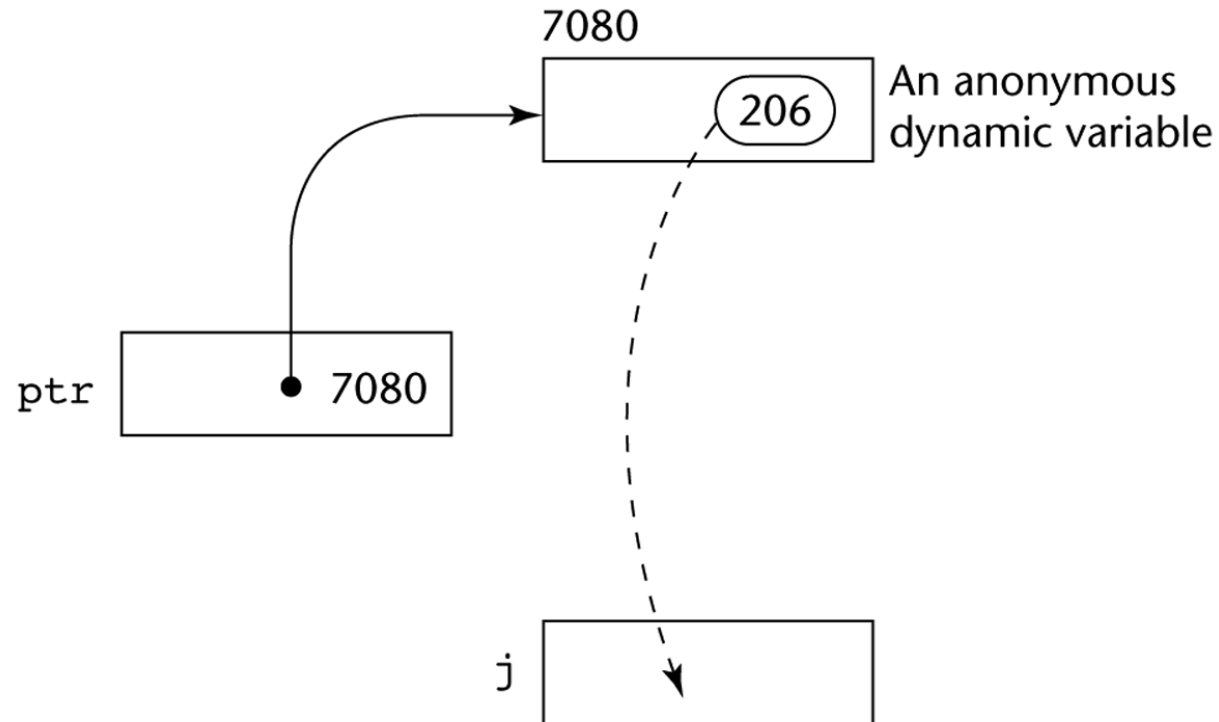
The assignment operation j = *ptr

Problems with Pointers

- **Dangling pointers (dangerous)**
  - A pointer points to a heap-dynamic variable that has been deallocated

- **Lost heap-dynamic variable**
  - An allocated heap-dynamic variable that is no longer accessible to the user program (often called *garbage*)
    - Pointer `p1` is set to point to a newly created heap-dynamic variable
    - Pointer `p1` is later set to point to another newly created heap-dynamic variable
    - The process of losing heap-dynamic variables is called *memory leakage*

Pointers in C and C++

- Extremely flexible but must be used with care
- Pointers can point at any variable regardless of when or where it was allocated
- Used for dynamic storage management and addressing
- Pointer arithmetic is possible
- Domain type need not be fixed (`void *`)

  `void *` can point to any type and can be type checked (cannot be de-referenced)

Pointer Arithmetic in C and C++

```
float stuff[100];
float *p;
p = stuff;
```

`*(p+5)` **is equivalent to** `stuff[5]` **and** `p[5]`
`*(p+i)` **is equivalent to** `stuff[i]` **and** `p[i]`

Reference Types

- C++ includes a special kind of pointer type called a *reference type* that is used primarily for formal parameters
  - Advantages of both pass-by-reference and pass-by-value
- Java extends C++'s reference variables and allows them to replace pointers entirely
  - References are references to objects, rather than being addresses
- C# includes both the references of Java and the pointers of C++

Evaluation of Pointers

- Dangling pointers and dangling objects are problems as is heap management

- Pointers are like `goto`'s--they widen the range of cells that can be accessed by a variable

- Pointers or references are necessary for dynamic data structures--so we can't design a language without them

Type Checking

- Generalize the concept of operands and operators to include subprograms and assignments

- *Type checking* is the activity of ensuring that the operands of an operator are of compatible types

- A *compatible type* is one that is either legal for the operator, or is allowed under language rules to be implicitly converted, by compiler– generated code, to a legal type
  - This automatic conversion is called a *coercion*.

- A *type error* is the application of an operator to an operand of an inappropriate type

Type Checking (continued)

- If all type bindings are static, nearly all type checking can be static

- If type bindings are dynamic, type checking must be dynamic

- A programming language is *strongly typed* if type errors are always detected

- Advantage of strong typing: allows the detection of the misuses of variables that result in type errors

Strong Typing

Language examples:

- FORTRAN 95 is not: parameters, `EQUIVALENCE`
- C and C++ are not: parameter type checking can be avoided; unions are not type checked
- Ada is, almost (`UNCHECKED CONVERSION` is loophole)

  (Java and C# are similar to Ada)

Name Type Equivalence

- *Name type equivalence* means the two variables have equivalent types if they are in either the same declaration or in declarations that use the same type name

- Easy to implement but highly restrictive:
  - Subranges of integer types are not equivalent with integer types
  - Formal parameters must be the same type as their corresponding actual parameters

Structure Type Equivalence

- *Structure type equivalence* means that two variables have equivalent types if their types have identical structures
- More flexible, but harder to implement

Type Equivalence (continued)

- Consider the problem of two structured types:
  - Are two record types equivalent if they are structurally the same but use different field names?
  - Are two array types equivalent if they are the same except that the subscripts are different? (e.g. [1..10] and [0..9])
  - Are two enumeration types equivalent if their components are spelled differently?
  - With structural type equivalence, you cannot differentiate between types of the same structure   (e.g. different units of speed, both float)

Summary

- The data types of a language are a large part of what determines that language's style and usefulness
- The primitive data types of most imperative languages include numeric, character, and Boolean types
- The user-defined enumeration and subrange types are convenient and add to the readability and reliability of programs
- Arrays and records are included in most languages
- Pointers are used for addressing flexibility and to control dynamic storage management