CS558 Programming Languages

Fall 2023

Lecture 8a

The Type Zoo

```
int x = 17

Z[1023] := 99;

double e = 2.81828

type emp = {name:  string, age:  int}

class Foo extends Bar { ...  }

data Day = Mon | Tue | Wed | Thu | Fri | Sat | Sun

String s = "abc"

type Days = set of Day

fold ::  (a -> b -> b) -> [a] -> b -> b

'a btree = LEAF of 'a | NODE of 'a * 'a btree * 'a btree
```

ORGANIZING THE ZOO

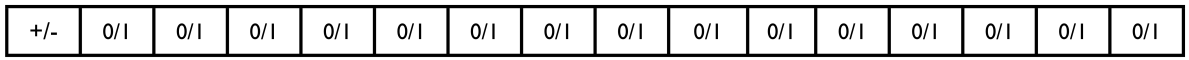
Mathematical View	$\Longleftrightarrow$	Programming Language View	$\Longleftrightarrow$	Machine View

Programming Language view: Types classify values and operations

Mathematical View: Types are sets of values

Machine View: Types describe memory layout of values

EXAMPLE: INTEGERS

$\mathbb{Z} \iff$ `int` $\iff$ 

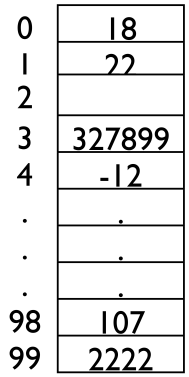
EXAMPLE: RECORDS

$$\mathbb{Z} \times \mathbb{Z} \iff \{\texttt{a:int,b:int}\} \iff$$

a	42
b	999

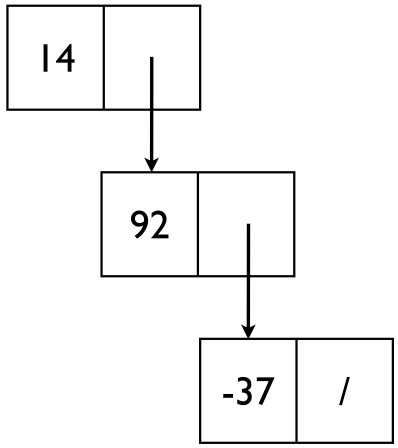
EXAMPLE: ARRAYS

$$\mathbb{N} \to \mathbb{Z} \iff \texttt{int[]} \iff$$



EXAMPLE: SEQUENCES

$\mathbb{Z}^* \iff$ `[Int]` $\iff$



Machine Types

Machine language doesn't distinguish types; all values are just bit patterns until **used**. As such they can be loaded, stored, moved, etc.

But certain type-specific **operations** are supported directly by hardware; the operands are in some sense implicitly typed.

Typical machine-level types:

- **Integers** of various sizes, signedness, etc. with standard arithmetic operations.
- **Bit vectors** of various sizes, with bit-level logic operations (`and`, `or`, etc.)
- **Floating point** numbers of various sizes, with standard arithmetic operations.
- **Pointers** to values stored in memory, with load and store operations.
- **Instructions**, i.e., code, which can be executed.

There is no abstraction at machine level: programs can inspect individual words and bits of any value.

PRIMITIVE ATOMIC TYPES

Most higher-level languages provide some **built-in atomic** types.

- e.g. in Java: `boolean`, `byte`, `short`, `int`, `long`, `char`, `float`, `double`

Built-in type names and operators are keywords or part of initial environment

Language may have special syntax for writing literal values (e.g. numbers, character strings)

Atomic types are usually **abstract**: programs cannot inspect internal details of value representation. e.g.,

- usually cannot extract mantissa or exponent from a floating point number (C/C++ is an exception)

- can't tell what convention is used to encode booleans as numbers (C/C++ is an exception)

Atomic types are often closely based on standard machine-level types, giving them an obvious representation and (usually) efficient implementation for operations

CONSTRUCTED TYPES

Many languages also provide a set of mechanisms for **constructing** new, user-defined types from existing types.

- e.g. Java has array and class definition mechanisms

Each type constructor comes with corresponding mechanisms for:

- constructing values
- inspecting values
- (for mutable types) modifying values

For example, in Java:

- arrays are constructed using `new` and an optional list of initializers; their elements are inspected and modified using subscript notation (e.g. `a[i]`).
- class instances (objects) are constructed using `new` and defined constructor methods; object fields can be accessed using dot notation (e.g. `p.x`).

Abstraction for Constructed Types

Are constructed types **abstract**?

- User-defined types can't be completely abstract, because we must have access to their components in order to write operations over the type.

- But many languages have mechanisms for limiting component access in order to enforce some level of abstraction, e.g. Java's `private` fields

The line between built-in and user-defined types is not always clear-cut.

- e.g, in Java, the `String` type is really just an ordinary class that happens to be provided in the standard library, except that there is special syntax for string literals and concatenation.

- Built-in constructed type values usually need to be kept abstract in order to guarantee that they behave as specified.

We will examine data abstraction mechanisms in detail later.

MATHEMATICAL VIEW OF TYPE CONSTRUCTORS

It can be enlightening to view type constructors as operators on the underlying **sets** represented by the component types.

Early on in the history of programming languages, it became clear that a small number of type operators suffices to describe most useful data structures:

- Cartesian product ($S_1 \times S_2$)
- Disjoint union ($S_1 + S_2$)
- Mapping (by explicit enumeration or by formula) ($S_1 \rightarrow S_2$)
- Set ($\mathcal{P}^S$)
- Sequence (S^*)
- Recursive definition ($\mu s.T(s)$)

Representation of Constructed Types

Concretely, each language defines the internal **representation** of values of the composite type, based on the type constructor and the types used in the construction.

Historically, most languages have provided just a few constructors, usually with the property that constructed values can be represented and accessed **efficiently** on conventional hardware.

For conventional languages, this is the short list:

- **Records**
- **Unions**
- **Arrays**

Many languages also support manipulation of **pointers** to values of these types, in order to allow operating on data “by reference” and to support recursive structures.

RECORDS = CARTESIAN PRODUCTS

Records, tuples, "structures", etc. Nearly every language has them.

"Take a bunch of existing types and choose one value from each."

Examples (Ada Syntax)

```
type EMP is
  record
    NAME : STRING;
    AGE : INTEGER;
  end record;

E: EMP := (NAME => "ANDREW", AGE => 99);
```

(ML syntax):

```
type emp = string * int  (unlabeled fields)
val e : emp = ("ANDREW",99);

type emp =
   {name: string, age: int}  (labeled fields)
val e : emp = {name="ANDREW",age=99};
```

Mathematically: $emp = string \times \mathbb{Z}$

RECORDS (CONTINUED)

Standard operations: construction, selection, selective update.

Representation: Fields occupy successive memory addresses (perhaps with some padding to maintain hardware-required alignments), so total size is (roughly) sum of field sizes.

Each field lives at a known static offset from the beginning of the record, allowing very fast access using pointer arithmetic.

Because records may be large, they are often **boxed**, i.e., represented by a pointer. The fields within a record may also be represented this way.

DISJOINT UNIONS

Variant records, discriminated records, unions, etc.

“Take a bunch of existing types and choose one value from one type.”

Introduced in Pascal:

```
type RESULT = record
                case found : Boolean of
                  true: (value:integer);
                  false: (error:STRING)
                end;

function search (...) : RESULT;
```

Generally behave like records, with **tag** as an additional field.

A variant value is represented by the tag following by the representation of the particular variant. Its size is thus bounded by the size of the largest possible variant plus the tag size.

VARIANT INSECURITIES

Pascal variant records are **insecure** because it is possible to manipulate the tag independently from the variant contents.

```
tr.value := 101;    { write an integer }
write tr.error;     { but read a string! }

if (tr.found) then begin    { check for integer }
   tr := tr1;       { overwrite with arbitrary RESULT }
   x := tr.value    { might now contain a string }
```

These problems were fixed in Ada by requiring tag and variant contents to be set simultaneously, and inserting a runtime check on the tag before any read of the variant contents.

DISJOINT UNIONS DONE PROPERLY

ML has very clean approach to building and inspecting disjoint unions:

```
datatype result =  FOUND of integer |  NOTFOUND of string

fun search (..) : result  =
   if ... then FOUND 10 else NOTFOUND "problem"

case search(...) of
  FOUND x =>
    print ("Found it : " ^ (Int.toString x))
| NOTFOUND s =>
    print ("Couldn't find it : " ^ s)
```

Here `FOUND` and `NOTFOUND` tags are **not** ordinary fields. The `case` expression **combines** inspection of tag and extraction of values into one operation.

Mathematically: $result = \mathbb{Z} + string$

OBJECTS

Objects in class-based OO languages, e.g. Java, can be viewed as a sort of variant record.

- The object's class identifier is stored at the beginning of the record, and acts like a variant tag, distingushing among different subclasses.

- The remaining record fields correspond to the object's instance variables.

The class tag is used to control dynamic dispatch to the class's methods, and (depending on the language) might be accessible more directly (e.g. Java's `instanceof`).

- The class tag cannot be altered after the object is created, so there is no danger of insecurity.

ARRAYS

The oldest type constructor, found in Fortran (the first real high-level language); crucial for numerical computations.

Basic representation idea: a **table** laid out in adjacent memory locations permitting **indexed access** to any element in constant time, using the hardware's ability to compute memory addresses.

Mathematically: A finite **mapping** from an **index set** to **element set**.

Index set is nearly always a set of integers $0..n-1$, or some other discrete set isomorphic to such a set.

Multidimensional arrays (**matrices**) can be built in several ways:

- using an index set of tuples of integers (e.g. in Fortran)
- using an element set of arrays (e.g. in C/C++/Java)

Array Size and Bounds Checking

Is the size of an array part of its type? Some older languages (e.g. Fortran) took this attitude, but most modern languages are more flexible, and allow the size to be set independently for each array value when the array is first created:

- as a local variable, e.g., in Ada:

```
function fred(size:integer);
  var bill: array(0..size) of real;
```

- or on the heap, e.g., in Java:

```
int[] bill = new int[size];
```

Arrays are often large, and hence better to treat as boxed.

The major security issue for arrays is **bounds checking** of index values. In general, it's not possible to check all bounds at compile time (though often possible in particular cases). Runtime checks are always possible; this may be costly, but it is usually worth it!

FUNCTIONS AND MAPPINGS

Mathematical mappings can also be represented by an algorithmic **formula**.

A **function** gives a “recipe” for computing a **result** value from an **argument** value.

A program function can describe an infinite mapping.

But differs from mathematical function in that:

- it must be specified by an explicit algorithm
- executing the function may have side-effects on variables.

As we have seen, it can be very handy to manipulate functions as first-class values.

SEQUENCES

What about data structures of essentially **unbounded** size, such as **sequences** (or **lists**)?

"Take an arbitrary number of values of some type."

Such data structures require special treatment: they are typically represented by small segments of data linked by pointers, and dynamic storage allocation (and deallocation) is required.

The basic operations on a sequence include

- **concatenation** (especially concatenating a single element onto the head or tail of an existing sequence); and
- **extraction** of elements (especially the head).

An important example is the (unbounded) **string**, a sequence of chars.

Best representation depends heavily on what nature and frequency of various operations. Hard to give single, uniformly efficient implementation.

DEFINING SEQUENCES

Unless the programming language supports sequences directly, the programmer must define them using a **recursive** definition.

For example, a list of integers is either

- **empty**, or
- has a **head** which is an integer and **tail** which is itself a list of integers.

In mathematical notation, we can write

$$list_{\mathbb{Z}} = empty + (\mathbb{Z} \times list_{\mathbb{Z}})$$

or, avoiding the need to name the type,

$$\mu t.(empty + (\mathbb{Z} \times t))$$

where the **fixpoint** operator $\mu s.T(s)$ defines the smallest type s such that $s = T(s)$.

ML lets us write down this type almost directly:

```
datatype intlist =  EMPTY | CELL of int * intlist
```

Recursive Types

Recursive definitions can be used to define and operate on more complex types, in which the type being defined appears more than once in the definition.

For example, binary trees with integers at internal nodes and leaves could be defined mathematically as

$$bintree = \mu t.(\mathbb{Z} + (t \times \mathbb{Z} \times t))$$

As you know from the very first lab, these trees can be defined in Scala using two case classes to represent the disjoint union:

```
abstract case class Tree
case class Leaf(v: Int) extends Tree
case class Node(l: Tree, v: Int, r: Tree) extends Tree
```

In ML we can simply write:

```
datatype tree =
  Node of tree * int * tree
| Leaf of int
```

EFFICIENCY VS. RICH PRIMITIVE TYPES

Must language designers be slaves to hardware?

Historically, most mainstream, general-purpose languages have only provided built-in types that can be given simple hardware implementations with **efficient** and **predictable** performance.

But more modern languages are including more complicated primitive types, because they are so useful. Examples:

- "Bignum" representations for arbitrary-precision numbers (Scheme, Python, Haskell, etc.)
- Strings (Java, Python, Perl, etc.)
- "Associative arrays" in which index set can be an arbitrary type rather than just integers (Awk, Perl, JavaScript, Python, etc.)
- Lists (LISP, Scheme, Haskell, Python, etc.)
- Sets (Pascal, SETL, Python, etc.)