The ICI Programming Language

Tim Long

CHAPTER 1 *Introduction* ***9***

CHAPTER 2 *A brief tutorial for C programmers* ***11***

Hello world **11**
Program structure **12**
Variables and arithmetic **12**
Lexicon, syntax and flow control statements **13**
Aggragate data types and the nature of objects **14**
Making and manipulating aggragates ***15***
Literal data items **16**
Other operations and core functions **16**
Regular expressions ***16***

CHAPTER 3 *Some sample programs* ***19***

Ackermann's function **19**
Array access **20**

CHAPTER 4 *ICI Language Reference* ***21***

The lexical analyser ***21***
An introduction to variables, modules and scope ***22***
The parser ***23***
Expressions **25**
Factors ***25***
An introduction to arrays, sets and structs ***26***
Back to expression syntax ***27***
Primary operators ***29***
Terms ***31***
Prefix operators ***31***
Postfix operators ***31***
Binary operators ***32***
Binary operator summary ***34***
Statements **35**
Simple expression statements ***35***
Compound statements ***35***
The if statement ***36***
The while statement ***36***
The do-while statement ***36***
The for statement ***37***
The forall statement ***37***
The switch, case, and default statements ***38***
The break and continue statements ***39***
The return statement ***40***
The try statement ***40***
The critsect statement ***41***
The waitfor statement ***41***

The null statement ***42***
Declaration statements ***43***
Abbreviated function declarations ***44***
Functions ***44***
Objects **47**
Equality ***49***
Structure and set keys ***51***
Structure super types ***52***
An aside on variables and scope ***54***
Base types **55**
array - An ordered sequence of objects ***56***
file - An open file reference ***56***
float - A double precision floating point number ***57***
func - A function ***57***
int - A signed 32 bit integer ***58***
mem - A reference to raw machine memory ***58***
method - A binding of a function and a subject object ***58***
ptr - A reference to a storage location ***59***
regexp - A compiled regular expression ***60***
set - An unordered collection of objects ***60***
string - An ordered sequence of 8 bit characters ***60***
struct - An unordered set of mappings ***61***
Operators **61**
Automatic library loading ***66***

CHAPTER 5 *Object-oriented programming in ICI* ***69***

Sub-classes ***71***
A shorthand for creating instance variables ***74***
Global methods ***74***
Taking advantage of dynamic binding ***75***
Standard global methods ***75***
Multiple inheritance ***76***

CHAPTER 6 *Core language functions* ***77***

Core function summary **77**
Core language functions **79**
`float|int = abs(float|int)` ***79***
`angle = acos(x)` ***80***
`mem = alloc(nwords [, wordz])` ***80***
`array = array(any...)` ***80***
`float = asin(x)` ***80***
`value = assign(struct, key, value)` ***80***
`angle = atan(x)` ***80***
`angle = atan2(y, x)` ***80***
array|struct = build(dims... [, options, content...]) ***80***
`float|struct = calendar(struct|float)` ***81***
`return = call(func [, arg...], args)` ***82***
`float = ceil(x)` ***82***

close(file) ***82***
close(file) ***83***
int = cmp(a, b) ***83***
x = cos(angle) ***83***
float = cputime([float]) ***83***
file = currentfile(["raw"]) ***83***
int = debug([int]) ***84***
del(aggr, key) ***84***
array = dir([path,] [regexp,] [format]) 85
int = eq(obj1, obj2) ***85***
int = eof([file]) ***85***
eventloop() ***85***
exit([string|int|NULL]) ***85***
float = exp(x) ***85***
array = explode(string) ***86***
fail(string) ***86***
value = fetch(struct, key) ***86***
value = float(x) ***86***
float = floor(x) ***86***
flush([file]) ***86***
float = fmod(x, y) ***86***
file = fopen(name [, mode]) ***86***
string = getchar([file]) ***87***
string = getfile([file]) ***87***
string = getline([file]) ***87***
string = gettoken([file [, seps]]) ***87***
array = gettokens([file [, seps [, terms]]]) ***87***
string = gsub(string, string|regexp, string) ***88***
string = implode(array) ***89***
struct = include(string [, scope]) ***89***
value = int(any) ***89***
subpart = interval(str_or_array, start [, length]) ***89***
int = inst|class:isa(any) ***89***
int = isatom(any) ***89***
array = keys(struct) ***90***
any = load(string) ***90***
float = log(x) ***90***
float = log10(x) ***90***
mem = mem(start, nwords [, wordz]) ***90***
file = mopen(mem [, mode]) ***90***
int = nels(any) ***90***
inst = class:new() ***91***
foat = now() ***91***
number = num(x) ***91***
scope = parse(source [, scope]) ***91***
any = pop(array) ***92***
file = popen(string, [flags]) ***92***
float = pow(x, y) ***92***
printf([file,] fmt, args...) ***92***

profile(filename) **93**
any = push(array, any) **94**
put(string [, file]) **94**
int = rand([seed]) **94**
reclaim() **94**
re = regexp(string [, int]) **94**
re = regexpi(string [, int]) **95**
remove(string) **95**
any = rpop(array) **95**
any = rpush(array, any) **95**
current = scope([replacement]) **95**
int = seek(file, int, int) **96**
set = set(any...) **96**
func = signal(string|int [, string|func]) **96**
string = signam(int) **96**
x = sin(angle) **96**
sleep(num) **96**
array = smash(string [, regexp [, replace...] [, include_remainder]) **97**
file = sopen(string [, mode]) **97**
sort(array [, func [, arg]]) **98**
string = sprintf(fmt, args...) **98**
x = sqrt(float) **99**
string = string(any) **99**
struct = struct([super,] key, value...) **99**
string = sub(string, string|regexp, string) **99**
current = super(struct [, replacement]) **100**
int = system(string) **100**
x = tan(angle) **100**
exec = thread(callable, args...) **100**
string = tochar(int) **100**
int = toint(string) **100**
any = top(array [, int]) **100**
int = trace(string) **101**
string = typeof(any) **101**
string = version() **101**
array = vstack() **101**
event = waitfor(event...) **102**
wakeup(any) **102**
Command Line Arguments **102**
argv **102**
argc **102**

CHAPTER 7 *Regular expressions* **103**

Regular Expression Syntax **103**
122

CHAPTER 8 *Interfacing with C and C++* **123**

Universal rules and conventions **123**
Include files and libraries **123**

The nature of ICI objects ***124***
Garbage collection, ici_incref() and ici_decref() ***124***
The error return convention ***125***
ICI's allocation functions ***126***
Common tasks **126**
Writting a simple function that can be called from ICI ***126***
Calling an ICI function from C ***128***
Making new ICI primitive objects ***128***
Writting and compiling a dynamically loading extension module ***128***
Referring to ICI strings from C code ***130***
Accessing ICI array objects from C ***131***
131
Using ICI independently from multiple threads ***131***
Summary of ICI's C API **131**
*void *ici_alloc(size_t z)* ***136***
int ici_argcount(int n) ***136***
*int ici_register_type(type_t *t)* ***137***
137
How it works **137**

CHAPTER 9 *Obsolete features* ***139***

OBSOLETE: Method Calls ### ***139***

CHAPTER 1

Introduction

ICI is a general purpose interpretive programming language that has dynamic typing and flexible data types with the basic syntax, flow control constructs and operators of C. It is designed for use in many environments, including embedded systems, as an adjunct to other programs, as a text-based interface to compiled libraries, and as a cross-platform scripting language with good string-handling capabilities.

The ICI language and source is not copyright in any way.

This document is the basic reference for the core language and functions. There is also an extensive man page that includes details of command line invocation not described here. Additional documentation is provided in ICI source releases. The ICI web site is http://www.zeta.org.au/~atrn/ici/

CHAPTER 2

A brief tutorial for C programmers

This chapter is intended as a quick tutorial introduction to ICI for programmers familiar with C or C++. It does not dwell on formal definitions and exceptions. For precise definitions, see the next chapter: *ICI Language Reference*. Because ICI's syntax and flow control constructs are based on those of C, a C programmer has a particular advantage in learning to use ICI. This tutorial will take advantage of that and move quickly through areas that are unsuprising to C programmers.

This tutorial will also occasionally elude to how things work inside the interpreter as, to a programmer, this can aid comprehension and give an idea of the implications of using certain contructs.

Hello world

The ICI *hello world* program is simply:

```
printf("Hello world.\n");
```

ICI's printf is the same as C's. You can verify your ICI execution environment by placing that single line in a file (often with a *.ici* suffix) and running it with:

```
ici hello.ici
```

And of course on UNIX like systems you can place:

```
#!ici
```

in the first line to allow direct execution of the file.

Program structure

An ICI program file is a sequence of statements. These statements include both executable statements as you would expect to find within C functions, and declaration statements as you would expect to find at the file level of a C program. Thus:

```
printf("Let's define a function.\n");

static
func()
{
    printf("Hello from func.\n");
}

printf("Now let's call it.\n");
func();
```

Will produce:

```
Let's define a function.
Now let's call it.
Hello from func.
```

Variables and arithmetic

Because ICI is a dynamically typed language, the nature of a variable is of course different from those of C. But for typical arithmetic the differences are invisible. All ICI variables *refer* to storage that records both the *type* and *data* of the variable's current value . Thus we can say:

```
x = 1;
```

which makes *x* refer to the integer one. Then

```
x = 1.0;
```

which updates *x* to refer to the float one. Then

```
x = "one";
```

which updates *x* to refer to the string "one".

As a C programmer, you can consider all ICI variables to be typless pointers to objects that record both type and value. But because ICI's built in operators know this is the case, they read and generate the pointed-to values automatically. Thus ordinary arithmetic is unsuprising:

```
fahr = 100.0;
celsius = (5.0 / 9.0) * (fahr - 32.0);
printf("%g deg. F is %g deg. C\n", fahr, celsius);
```

works as one would expect. So most of the time you don't need to consider this at all. All objects in ICI are subject to automatic garbage collection, so no explicit freeing is required.

Because ICI variable are dynamically typed, you don't need to declare them. But ICI supports heirchial modularisation and it is often desireable to declare at what scope a variable lives. Thus we have:

```
extern xxx; /* xxx is visible to other files. */
static sss; /* sss is visible in this file. */
auto aaa;    /* aaa is visible in the local scope. */
```

The word *static* is used in the C sense of the value being persistent. This variable will exist with persistent value as long as functions in this module are still callable. Extern variables are also persistent, they just have more global scope. Consider:

```
static
func(arg)
{
    auto  local;

    local = 10;
    for (i = 0; i < local; ++i)
        printf("%d\n", i * arg);
}
```

This function (which is declared static) has an *auto* variable. Auto variables are, as in C, the variables that spring into existence (on the stack) for the duration of a single execution of a function. The function also uses the variable *i*. If an undeclared variable is assigned to, it is implicitly declared *auto*. That can be dangerous in large programs with many variables of more global scope that may already exist, so as a style rule, implicit autos are normally kept to one or two characters, and more global variables should not be.

Auto variables, and their implicit declaration, also work at the file level. They have a similar (in a sense) semantics. While the file is being parsed, they exsit. But they evaporate afterwards. They are not visible to functions defined within the file. We used implicit auto variables in our farenheit to celsius conversion above.

Lexicon, syntax and flow control statements

ICI's lexicon is (basically) the same as C's. Same tokens, comments (including //) and literal data values. Sorry, no preprocessor.

ICI's syntax is, wherever possible, the same as C's. Naturally differences arise due to the different nature of the environment, as we have seen above.

As we have seen, expressions are as in C. There are of course additional data types, literals, and operators, but these build from the initial C compatible set.

The flow control constructs *if-else*, *while*, *for*, *do-while*, *switch* (including *case* and *default*), *continue*, *break* and *return* all have the same basic syntax and semantics as C. But there is no *goto*.

In addition to these classic C statements forms, ICI has *forall, try-onerror*, *waitfor*, and *critsect*. But before considering these, we will look at aggragate data types and the nature of objects, which is the one aspect a C programmer needs to understand before writting effective ICI code.

Aggragate data types and the nature of objects

ICI supports a number of "aggragate" data types. Principly:

array	Simple ordered collections of values that can be indexed by integers. The first item is at index 0. They can be efficiently pushed and poped at either end.
struct	Mappings from an index (any object) to a value. Also known as associative arrays, dictionaries, maps, hashes, etc in other languages. Adding entries, lookup and deletion are all efficient operations irrespective of the complexity of the objects involved.
set	Simple unordered collections of values.

Each of these hold a collection of references to other objects. There is a significant distinction between these aggrgate types and the simple types such as *int*, *float*, and *string*. These simple types have no *modifiable* internal structure. They are read-only. In fact, when an object of one of these types is required (say as the result of some arithmetic operation) it is looked for in a hash table of all such objects, and the entry found there is used. It is created and added if it does not exist.

Thus we can see that all strings "xyz" in an ICI program are just pointers to the same single object in memory. The same is true for integers (which are 32 bit signed values) and floats (which are double precision values). An object that has been resolved to its single unique (read-only) version is said to be *atomic*.[1]

Aggragates, on the other hand, are internally modifiable in-place.

In ICI, "indexing" an aggragate is the most primitive way of accessing internal elements. But we use the term indexing in a more general sense than simple array indexing. For example, array indexing is unsuprising, so:

```
a[0] = 3;
```

sets the first element of an array *a* to 3. With a sruct *s* we might say:

```
s.value = 3;
```

which sets the *value* fields of the struct to 3. But this is just an "indexing" operation on the struct. In fact it is just a syntactic variation on:

```
s["value"] = 3;
```

1. This type of mechanism is typical for dynamically typed interpretive langauges such as ICI. But it is less common to apply it to numbers.

Arrays, structs and sets are all objects that support indexing to refer to internal values (i.e. object references) for read or write. Each varies only in how they are structured internally, and how they interpret the "key", or index, applied to them.

- Arrays are growable circular buffers of object references that can only be indexed by integers, which are interpreted as an offset from the first element.
- Structs are hash tables that map one object reference to another. The index reference itself is the basis for indexing, not the details of the index object (that is, the indexing operation only looks at the index as a pointer, not at what it points to). But because ints, strings, floats, etc are already resolved to unique pointers based on their values, this behaviour is indistiguashable from full value hashing and comparison for simple (atomic) types.
- Sets are hash tables that merely record the presence or absence of an object in the same manner as structs, but they have no associated value. Although they have an "implicit" value of 1 if the object is in the set.

Arrays, structs and sets all return the special object NULL if the key is not in their current domain.

Making and manipulating aggragates

The simplest ways to make aggragates is the functions *array()*, *set()* and *struct()*. For example:

```
a = array(1, 2.5, "hello");
b = set("bye", 5.5, 9);
c = struct("a", 12, "b", 13);
```

The *struct()* function interprets its arguments pair-wise as key-value pairs. If, after executing the above code, we do:

```
printf("a[2] = %s\n", a[2]);
if (b[9])
    printf("The set b contains 9.\n");
printf("c.a = %d\n", c.a);
```

we will see:

```
a[2] = hello
The set b contains 9.
c.a = 12
```

It is equally common to see these functions used to make empty aggregates that are then added to through further code. For example:

```
things = array();
while ((thing = get_next_thing()) != NULL)
     push(things, thing);
```

Or:

```
node = struct();
node.name = name;
node.left = a;
node.right = b;
```

Literal data items

ICI supports in-line literal aggragates. That is, like an initialisated structure in C, but instead of being tied to a variable declaration, they are self-describing, and can be used anywhere. For example:

```
[array 1, 2, 3]
```

is a term in an expression. Just like a literal string in C:

```
"Hello world.\n"
```

the compiler builds the data structure in memory somewhere and the term evaluates to a reference to it. Examples of *array*, *set*, and *struct* literals are:

```
a = [array 1, 2.5, "hello"];
b = [set "bye", 5.5, 9];
c = [struct a = 12, b = 13];
```

Arrays and *sets* have syntax almost identical to the run-time functions that create the same types. But *structs* have a more convinient syntax for the commonest activity; associating values with named keys.

Be careful not to confuse literals with the run-time functions of the same name. Confusion often arises because at the file level where a statement is parsed, then immediately executed, there isn't much effective difference. But in a loop or function there is a very big difference.

Other operations and core functions

Common to all dynamically typed interpreted languages, execution speed is very different from fully compiled statically typed languages. Achieving useful performance relies heavily on the use of operations and functions that perform the "inner loops" of a program, but are fully compiled and carefully optimised.

ICI is no exception to this principle. So it is wise to be aware of the full repetoir of operations, core functions, and extension modules available. However, in this brief tutorial we won't attempt to enumerate all such features. They are listed in subsequent chapters, and a skim through *Operators* in the *ICI Language Reference* chapter, the *Core language functions*, chapter, and *Some extension modules* is recommended. Having said that, a few of the commonest non-C features and idioms are worth illustrating here.

Regular expressions

Regular expressions are "simple" (atomic) types in ICI, just like ints, floats and strings. A literal a regular expression is delimited by # characters (like a string is delimited by " characters). For example:

```
while ((line = getline()) != NULL)
{
    if (line ~ #^abc#)
        printf("%s\n", line);
}
```

will print all lines starting with *abc*. The ~ operators is *"matches"* and *!~* is *"doesn't match"*. Other operators exist which extract sub-matches. Regular expressions can be very useful for avoiding character-by-character operations on text. They are a very efficient way of matching and breaking up text.

For example, one of my first resorts in dealing with some new regular text file is to load the entire file, use a function called *smash()* to break it up into lexical units based on regular expressions, then rearrange the result into the data I want. Consider doing this to load a "CSV" file (Comma Seperated Fields - each line is comma seperated fields, each field optionally surrounded by double quotes).

```
/*
 * Smash the file into fields and seperators. Each
 * seperator is either a "," or a "\n". Fields are
 * either plain or quoted, but the quotes
 * are removed.
 */
csv = smash
(
    getfile(f),                          /* The file. */
    #(([^,"\n]*)|"([^"\n]*)")([,\n])#, /* Match these. */
    "\\2\\3",                            /* For each.. */
    "\\4"                                /* ..push these*/
);
/*
 * Re-build the linear array into an array of arrays
 * based on the "\n" seperators.
 */
a = array(aa = array());
while (nels(csv) > 0)
{
    push(aa, rpop(csv));
    if (rpop(csv) == "\n")
        push(a, (aa = array()));
}
```

CHAPTER 3

Some sample programs

This chapter contains a small collection of very simple sample programs. These programs are not random. They are based on the a set of simple language benchmark tests used in *The Great Computer Language Shootout* by *Doug Bagley* (http://www.bagley.org/~doug/shootout/) and the *The Great Win32 Computer Language Shootout* by *Aldo Calpini* (http://dada.perl.it/shootout/). These programs have been chosen because at those sites you can view programs written to exactly the same specification in almost any programming language you are likely to know.

The specification of this benchmark suite demmands that some of the programs are implemented the *same way* as they are implemented in the other languages. Others are merely required to do the *same thing*.

Many of the tests take a single optional command-line argument being the integer number of times some loop is to be repeated. This is typically obtained in each program with a line like:

```
n = argv[1] ? int(argv[1]) : 1;
```

Some tests expect input data, which is generally read from standard input. See the sites mentioned above for further details.

No comment will be made on code that should be unsurprising to someone who knows C.

Ackermann's function

This test must be implemented in the same recursive manner in all languages. It is designed to stress recursion by computing Ack(3, N) for various (small) N.

```
static
Ack(M, N)
{
    return M ? (Ack(M - 1, N ? Ack(M, N - 1) : 1)) : N + 1;
}

n := argv[1] ? int(argv[1]) : 1;
```

```
printf("Ack(3,%d): %d\n", n, Ack(3, n));
```

Array access

This test must be implemented in the same way in all languages. It must first build an array full on integers, then repeatedly add them to a second array, with each loop running backwards through the array.

Notice the use of the *build()* function to make the first array. The *"i"* argument to *build()* causes the content to be auto-incrementing integers, the 1 is the start value.

The second call to *build()* makes an array of size *n* with each element initialised to 0. The *"c"* argument means “apply the initialiser(s) cyclicly to leaf elements of the built data”.

```
n = argv[1] ? int(argv[1]) : 1;

x = build(n, "i", 1);
y = build(n, "c", 0);

for (k = 0; k < 1000; ++k)
{
    for (i = n - 1; i >= 0; --i)
        y[i] += x[i];
}

printf("%d %d\n", y[0], y[n - 1]);
```

Count lines/words/characters

This test must count lines, words and characters from standard input and must do the same thing as the versions in other languages. However, it is not allowed to read the whole input at once, but must limit its read to no more than 4K bytes. There is no easy way to do this except by reading lines.

Notice the use of the *smash()* function to get the words of each line. *Smash()* is the most common method of breaking up strings. The \S+ pattern matches one or more non-whitespace characters. If we really wanted the words, the last argument to the *smash()* call would have been *"\\&"* (meaning append the matched portion to the array being built). However, we only want to count the words, so we just push empty strings. This saves the cost of actually extracting and making the string.

```
nl = nw = nc = 0;
while (l = getline())
{
     ++nl;
    nc += nels(l) + 1;
    nw += nels(smash(l, #\S+#, ""));
}
printf("%d %d %d\n", nl, nw, nc);
```

Echo client/server

For this test, each language is required to do the same thing. The specification says it should fork a child process that repeatedly sends a message to the parent (server), which echoes it back to the child (client), which checks it is correct. Because *fork()* is only available in versions of ICI running on UNIX-like systems (in the *sys* module), we actually use a thread here.

This test uses the *net* module, which provides sockets based networking primitives (it is documents seperately).

Notice the use of *waitfor* to wait for the child thread to finish. The *result* field of the child thread will be assigned the return value of the *echo_client* function when it returns. The thread object istself is waited on. A wakeup is done on the thread object on thread termnation.

```
n = argv[1] ? int(argv[1]) : 1;
static data = "Hello there sailor\n";

static
echo_client(n, port)
{
    sock := net.connect(net.socket("tcp/ip"), port);
    for (i := 0; i < n; ++i)
    {
        net.send(sock, data);
        if ((ans := net.recv(sock, nels(data))) != data)
            printf("received \"%s\", expected \"%s\"", ans, data);
    }
    net.close(sock);
    return 1;
}

ssock = net.listen(net.bind(net.socket("tcp/ip"), 0));
client = thread(echo_client, n, net.getportno(ssock));
csock = net.accept(ssock);
t = 0;
while (str = net.recv(csock, nels(data)))
{
    net.send(csock, str);
    t += nels(str);
}
waitfor(client.result; client)
    ;
printf("server processed %d bytes\n", t);
```

Exception mechanisms

For this test, each language is required to implement it the same way. The outer loop calls *hi_function()* which calls *low_function()* which calls *blowup()*. The *blowup()* function throws two types of exceptions, one of which must be caught by *lo_functio()* and the other by *hi_function()*.

ICI cannot selectively catch exceptions, so in *lo_function()* we must catch and re-throw the exception that is not for us. ICI exceptions are very simple, just being a string.

```
N = argv[1] ? int(argv[1]) : 1;

static HI = 0;
static LO = 0;

static
blowup(n)
{
    fail(n & 1 ? "low" : "hi");
}

static
lo_function (n)
{
    try
        blowup(n);
    onerror
    {
        if (error !~ #low#)
            fail(error);
        ++LO;
    }
}

static
hi_function(n)
{
    try
        lo_function(n);
    onerror
        ++HI;
}

static
some_function(n)
{
    try
        hi_function(n);
    onerror
        fail(error + " -- we shouldn't get here");
}

while (N)
    some_function(N--);

printf("Exceptions: HI=%d / LO=%d\n", HI, LO);
```

Fibonacci numbers

In this test each language is required to compute a fibonacci number by the same recursive method.

```
static
fib(n)
{
    return n < 2 ? 1 : fib(n - 2) + fib(n - 1);
}

printf("%d\n", fib(argv[1] ? int(argv[1]) : 1));
```

Hash (associative array) access

All languages must implement this test the same way. We store the integers from 1..N in an array indexed by the hex string of the integer, then access it with decimal strings. Only some of the decimal strings will strike values we stored under the hex string keys.

```
n = argv[1] ? int(argv[1]) : 1;

x = struct();
for (i = 1; i < n + 1; ++i)
    x[sprintf("%x", i)] = i;

c = 0;
for (i = n; i > 0; --i)
    c += x[string(i)] != NULL;

printf("%d\n", c);
```

Hashes, part II

This is like the above, but isn't swamped by the time to make strings. The strings are made first, then used repeatedly.

```
n = argv[1] ? int(argv[1]) : 1;

h1 = struct();
for (i = 0; i < 10000; ++i)
    h1[sprintf("foo_%d", i)] = i;

h2 = struct();
for (i = 0; i < n; ++i)
{
    forall (v, k in h1)
    {
        if (h2[k] == NULL)
            h2[k] = 0;
        h2[k] += h1[k];
    }
}

printf("%d %d %d %d\n", h1["foo_1"], h1["foo_9999"],
    h2["foo_1"], h2["foo_9999"]);
```

Heapsort

In this test each language is required to implement an in-place heapsort in the same way.

```
static IM = 139968;
static IA = 3877;
static IC = 29573;

static
gen_random(max)
{
    static last = 42;

    return max * (last = (last * IA + IC) % IM) / IM ;
}

static
heapsort(n, ra)
{
    ir = n;
    l = (n >> 1) + 1;
    for (;;)
    {
        if (l > 1)
        {
            rra = ra[--l];
        }
        else
        {
            rra = ra[ir];
            ra[ir] = ra[1];
            if (--ir == 1)
            {
                ra[1] = rra;
                return;
            }
        }
        i = l;
        j = l << 1;
        while (j <= ir)
        {
            if (j < ir && ra[j] < ra[j+1])
                ++j;
            if (rra < ra[j])
            {
                ra[i] = ra[j];
                j += (i = j);
            }
            else
            {
                j = ir + 1;
            }
        }
```

```
            ra[i] = rra;
        }
    }

    N = argv[1] ? int(argv[1]) : 1;
    ary = array();
    for (i = 0; i <= N; ++i)
        ary[i] = gen_random(1.0);
    heapsort(N, ary);
    printf("%.10f\n", ary[N]);
```

Hello world

Couldn't get much simpler than this. We use *put()* which is raw unformatted output, unlike *printf()*.

```
put("hello world\n");
```

CHAPTER 4

ICI Language Reference

The ICI interpreter's *execution engine* calls on the *parser* to read and compile a statement from an input stream. The parser in turns calls on the *lexical analyser* to read tokens. Upon return from the parser the execution engine executes the compiled statement. When the statement has finished execution, the execution engine repeats the sequence.

The lexical analyser

The ICI lexical analyser breaks the input stream into tokens, optionally separated by white-space (which includes comments as described below). The next token is always the longest string of following characters which could possibly be a token. The following are tokens:

```
/      /=     $      @      (      )      {      }
,      ~      ~~     ~~=    ~~~    [      ]      .
*      *=     %      %=     ^      ^=     +      +=
++     -      -=     --     ->     >      >=     >>
>>=    <      <=     <=>    <<     <<=    =      ==
!      !=     !~     &      &&     &=     |      ||
|=     ;      ?      :      :=     :^
```

The following are also tokens:

- The character '#' followed by any sequence of characters except a newline, then another '#'. This token is a *regular-expression*.
- The character ' (single quote) followed by a single character (other than a newline) or a single *backslash character sequence* (described below), followed by another single quote. This token is a *character-code*. A single quote followed by other than the above sequence will result in an error.
- The character " (double quote) followed by any sequence of characters (other than a newline) and *backslash character sequences*, up to another double quote character. This token is a *string*.

A *backslash character sequence* is any of the following:

`\n`	newline (ASCII 0x0A)
`\t`	tab (ASCII 0x09)
`\v`	vertical tab (ASCII 0x0B)
`\b`	back space (ASCII 0x08)
`\r`	carriage return (ASCII 0x0D)
`\f`	form feed (ASCII 0x0C)
`\a`	audible bell (ASCII 0x07)
`\e`	escape (ASCII 0x1B)
`\\`	backslash (ASCII 0x5C)
`\'`	single quote (ASCII 0x27)
`\"`	double quote (ASCII 0x22)
`\?`	question mark (ASCII 0x3F)
`\c`*x*	control-*x*
`\x`*x*`..`	the character with hex code *x*...
`\`*n*	the character with octal code *n*. (1, 2 or 3 octal digits)

Consecutive string-literals, seperated only by white-space, are concatenated to form a single string-literal.

- Any upper or lower case letter, any digit, or '_' (underscore) followed by any number of the same (or other characters which may be involved in a floating point number while that is a valid interpretation). A token of this form may be one of three things:

  If it can be interpreted as an integer, it is an *integer-number.*

  Otherwise, if it can be interpreted as a floating point number, it is a *floating-point-number.*

  Otherwise, it is an *identifier.*

Notice that keywords are not recognised directly by the lexical analyser. Instead, certain identifiers are recognised in context by the parser as described below.

There are two forms of comments (which are white-space). One starts with the characters `/*` and continue until the next `*/`. The other starts with the characters `//` and continues until the next end of line. Also, lines which start with a # character are ignored. (Lines may be terminated with *linefeed*, *carriage return* or *carrige return* plus *linefeed*.)

An introduction to variables, modules and scope

Variables are simple identifiers which have a value associated with them. They are in themselves typeless, depending on the type of the value currently assigned to them.

The term *module* in ICI refers to a collection of functions, declarations and code which share the same variables. Typically each source file is a module, but not necessarily.

In ICI, modules may be nested in a hierarchical fashion. Within a module, variables can be declared as either *static* or *extern*. When a variable is declared as static it is visible to code defined in the module of its definition, and to code defined in sub-modules of that one. This is termed the *scope* of the variable.

When a variable is defined as *extern* it is declared *static* in the parent module. Thus the parent module and all sub-modules of the parent module have that variable in their scope. Variables of this type, whether originally declared extern or static, will be henceforward referred to as static variables.

Static variables are persistent variables. That is they remain in existence even when execution completely leaves their scope, despite not being visible to any executing code. They are visible again when code flow again enters their scope.

The scoping of static variables is strictly governed by the nesting of the modules, not by the flow of execution. For example. Suppose two neighbouring modules (call them module **A** and module **B**) each define a variable called **theVariable**. When some code in module **A** calls a function defined in module **B** and that function refers to **theVariable**; it is referring to the version of **theVariable** defined in module **B**, not the one defined in module **A**.

Variables in sub scopes hide variables of the same name defined in outer scopes.

The second type of variable in ICI is the *automatic*, or *auto*, variable. Automatic variables are not persistent. They last only as long as a module is being parsed or a function is being executed. For instance, each time a function is entered a copy is made of the auto variables which were declared in the function. This group of variables generally only persists during the execution of the function; once the function returns they are discarded.

The parser

The parser uses the lexical analyser to read a source input stream. The parser also has reference to the variable-scope within which this source is being parsed, so that it may define variables.

When encountering a variable definition, the parser will define variables within the current scope. When encountering normal executable code at the outermost level, the parser returns its compiled form to the execution engine for execution.

For some constructs the parser will in turn recursively call upon the execution engine to evaluate a sub-construct within a statement.

The following sections will work through the syntax of ICI with explanations and examples. Occasionally constructs will be used ahead of their full explanation. Their intent should be obvious.

The following notation is used in the syntax in these sections. Note that the syntax given in the text is not always exact, but rather designed to aid comprehension. The exact syntax is given in a later section.

bold	The **bold** text is literal ASCII text.
italic	The *italic* text is a construct further described elsewhere.
[xxx *]*	The xxx is optionally present.
xxx...	The xxx may be present zero or more times.
(xxx | yyy *)*	Either xxx or yyy may be present.

As noted previously there are no reserved words recoginsed by the lexical anaylyser, but certain identifiers will be recognised by the parser in certain syntactic positions (as seen below). While

these identifiers are not otherwise restricted, special action may need to be taken if they are used as simple variable names. They probably should be avoided. The complete list is:

NULL	auto	break	case
continue	default	do	else
extern	for	forall	if
in	onerror	return	static
switch	try	while	

We now turn our attention to the syntax itself.

Firstly consider the basic statement which is the unit of operation of the parser. As stated earlier the execution engine will call on the parser to parse one top-level statement at a time. We split the syntax of a statement into two categories (purely for semantic clarity):

statement *executable-statement*
declaration

That is, a statement is either an *executable-statement* or a *declaration*. We will first consider the *executable-statement*.

These are statements that, at the top-level of parsing, can be translated into code which can be returned to the execution engine. This is by far the largest category of statements:

executable-statement *expression* **;**
compound-statement
if (*expression* **)** *statement*
if (*expression* **)** *statement* **else** *statement*
while (*expression* **)** *statement*
do *statement* **while (** *expression* **) ;**
for (*[expression]* **;** *[expression]* **;** *[expression]* **)** *statement*
forall (*expression [* **,** *expression]* **in** *expression* **)** *statement*
switch (*expression* **)** *compound-statement*
case *parser-evaluated-expression* **:**
default :
break ;
continue ;
return *[expression]* **;**
try *statement* **onerror** *statement*
waitfor (*expression* **;** *expressiion* **)** *statement*
critsect *statement*
;

These are the basic executable statement types. Many of these involve *expression*s, so before examining each statement in turn we will examine the *expression*.

Expressions

We will examine expressions by starting with the most primitive elements of expressions and working back up to the top level.

Factors

The lowest level building block of an expressions is the *factor*:

factor
- *integer-number*
- *character-code*
- *floating-point-number*
- *string*
- *regular-expression*
- *identifier*
- **NULL**
- **(** *expression* **)**
- **[array** *expression-list* **]**
- **[set** *expression-list* **]**
- **[struct** *[(:* | *=) expression* **,** *] assignment-list* **]**
- **[class** *[(:* | *=) expression* **,** *] assignment-list* **]**
- **[func** *function-body* **]**
- **[module** *[(:* | *=) expression* **,** *] statement...* **]**

The constructs *integer-number*, *character-code*, *floating-point-number*, *string*, and *regular-expression* are primitive lexical elements (described above). Each is converted to its internal form and is an object of type *int*, *int*, *float*, *string*, or *regexp* respectively.

A *factor* which is an *identifier* is a variable reference. But its exact meaning depends upon its context within the whole expression. Variables in expressions can either be placed so that their value is being looked up, such as in:

```
a + 1
```

Or they can be placed so that their value is being set, such as in:

```
a = 1
```

Or they can be placed so that their value is being both looked up and set, as in:

```
a += 1
```

Only certain types of expression elements can have their value set. A variable is the simplest example of these. Any expression element which can have its value set is termed an *lvalue* because it can appear on the left hand side of an assignment (which is the simplest expression construct which requires an lvalue). Consider the following two expressions:

```
1 = 2                   /* WRONG */
a = 2                   /* OK */
```

The first is illegal because an integer is not an lvalue, the second is legal because a variable reference is an lvalue. Certain expression elements, such as assignment, require an operand to be an lvalue. The parser checks this.

The next factor in the list above is **NULL**. The keyword NULL stands for the value NULL which is the general undefined value. It has its own type, NULL. Variables which have no explicit initialisation have an initial value of NULL. Its other uses will become obvious later in this document.

Next is the construct **(** *expression* **)**. The brackets serve merely to make the expression within the bracket act as a simple factor and are used for grouping, as in ordinary mathematics.

Finally we have the four constructs surrounded by square brackets. These are textual descriptions of more complex data items; typically known as *literals*. For example the factor:

```
[array 5, 6, 7]
```

is an array of three items, that is, the integers 5, 6 and 7. Each of these square bracketed constructs is a textual description of a data type named by the first identifier after the starting square bracket. A full explanation of these first requires an explanation of the fundamental aggregate types.

An introduction to arrays, sets and structs

There are three fundamental aggregate types in ICI: arrays, sets, and structs. Certain properties are shared by all of these (and other types as will be seen later). The most basic property is that they are each collections of other values. The next is that they may be "indexed" to reference values within them. For example, consider the code fragment:

```
a = [array 5, 6, 7];
i = a[0];
```

The first line assigns the variable a an array of three elements. The second line assigns the variable i the value currently stored at the *first* element of the array. The suffixing of an expression element by an expression in square brackets is the operation of "indexing", or referring to a subelement of an aggregate, and will be explained in more detail below.

Notice that the *first* element of the array has index *zero*. This is a fundamental property of ICI arrays.

The next ICI aggregate we will examine is the set. Sets are unordered collections of values. Elements "in" the set are used as indexes when working with the set, and the values looked up and assigned are interpreted as a booleans. Consider the following code fragment:

```
s = [set 200, 300, "a string"];
if (s[200])
     printf("200 is in the set\n");
if (s[400])
     printf("400 is in the set\n");
if (s["a string"])
     printf("\"a string\" is in the set\n");
s[200] = 0;
if (s[200])
     printf("200 is in the set\n");
```

When run, this will print:

```
200 is in the set
"a string" is in the set
```

Notice that there was no second printing of "200 is in the set" because it was removed from the set on the third last line by assigning zero to it.

Now consider structs. Structs are unordered collections of values indexed by any values. Other properties of structs will be discussed later. The typical indexes of structs are strings. For this reason notational shortcuts exist for indexing structures by simple strings. Also, because each element of a struct is actually an index and value pair, the syntax of a struct literal is slightly different from the arrays and sets seen above. Consider the following code fragment:

```
s = [struct a = 123, b = 456, xxx = "a string"];
printf("s[\"a\"] = %d\n", s["a"]);
printf("s.a = %d\n", s.a);
printf("s.xxx = \"%s\"\n", s.xxx);
```

Will print:

```
s["a"] = 123
s.a = 123
s.xxx = "a string"
```

Notice that on the second line the structure was indexed by the string "a", but that the assignment in the struct literal did not have quotes around the *a*. This is part of the notational shortcut which will be discussed further, below. Also notice the use of *s.a* in place of s["a"]. This is a similar shortcut, also discussed below.

Back to expression syntax

The aggregate literals, which in summary are:

[array *expression-list* **]**
[set *expression-list* **]**
[struct *[(:* | *=) expression* **,** *] assignment-list* **]**
[class *[(:* | *=) expression* **,** *] assignment-list* **]**
[module *[(:* | *=) expression* **,** *] statement...* **]**
[func *function-body* **]**

involve three further constructs, the *expression-list*, which is a comma separated list of expressions; the *assignment-list*, which is a comma separated list of assignments; and the *function-body*, which is the argument list and code body of a function. The syntax of the first of these is:

expression-list
empty
expression [**,** *]*
expression **,** *expression-list*

The *expression-list* is fairly simple. The construct *empty* is used to indicate that the whole list may be absent. Notice the optional comma after the last expression. This is designed to allow a

more consistent formatting when the elements are line based, and simpler output from programmatically produced code. For example:

```
[array
     "This is the first element",
     "This is the second element",
     "This is the third element",
]
```

The assignment list has similar features:

assignment-list	*empty*
	assignment **[** **,** **]**
	assignment **,** *assignment-list*
assignment	*struct-key*
	struct-key **=** *expression*
	struct-key function-body
struct-key	*identifier*
	(*expression* **)**

Each *assignment* is either an assignment to a simple identifier or an assignment to a full expression in brackets. The assignment to an identifier is merely a notational abbreviation for an assignment to a string. The following two struct literals are equivalent:

```
[struct abc = 4]
[struct ("abc") = 4]
```

The syntax of a *function-body* is:

function-body	**(** *identifier-list* **)** *compound-statement*
identifier-list	*empty*
	identifier **[** **,** **]**
	identifier **,** *identifier-list*

That is, an *identifier-list* is an optional comma separated list of *identifiers* with an optional trailing comma. Literal functions are rare in most programs; functions are normally named and defined with a special declaration form which will be seen in more detail below. The following two code fragments are equivalent; the first is the abbreviated notation:

```
static fred(a, b){return a + b;}
```

and:

```
static fred = [func (a, b){return a + b;}];
```

The meaning of functions will discussed in more detail below.

Aggregates in general, and literal aggregates in particular, are fully nestable:

```
[array
```

```
        [struct a = 1, c = 2],
        [set "a", 1.2, 3],
        "a string",
    ]
```

Note that aggregate literals are entirely evaluated by the parser. That is, each expression is evaluated and reduced to a particular value, these values are then used to build an object of the required type. For example:

```
[struct a = sin(0.5), b = cos(0.5)]
```

Causes the functions sin and cos to be called during the parsing process and the result assigned to the keys *a* and *b* in the struct being constructed. It is possible to refer to variables which may be in existence while such a literal is being parsed[1].

This ends our consideration of the lowest level element of an expression, the *factor.*

Primary operators

A simple factor may be adorned with a sequence of *primary-operation*s to form a *primary-expression*. That is:

primary-expression *factor primary-operation...*

primary-operation **[** *expression* **]**
index-operator identifier
index-operator **(** *expression* **)**

index-operator *Any of:*
`. -> : :^`

The first *primary-operation* (above) we have already seen. It is the operation of "indexing" which can be applied to aggregate types. For example, if *xxx* is an array:

```
xxx[10]
```

refers to the element of xxx at index 10. The parser does not impose any type restrictions (because typing is dynamic), although numerous type restrictions apply at execution time (for instance, arrays may only be indexed by integers, and floating point numbers are not able to be indexed at all).

Of the other index operators, **.** *identifier*, is a notational abbreviation of **[** "*identifier*" **]** , as seen previously. The bracketed form is again just a notational variation. Thus the following are all equivalent:

```
xxx["aaa"]
xxx.aaa
```

1. Literal aggregates are analagous to literal strings in K&R C. And likewise they have the property that modifications to the literal are persistent. Returning to the original use of the literal after it has been modified does not magically restore it to its original value.

```
xxx.("aaa")
```

And the following are also equivalent to each other:

```
xxx[1 + 2]
xxx.(1 + 2)
```

Note that factors may be suffixed by any number of *primary-operation*s. The only restriction is that the types must be right during execution. Thus:

```
xxx[123].aaa[10]
```

is legal.

The two constructs

-> identifier
-> **(** *expression* **)**

are again notational variations. In general, constructs of the form:

primary-expression -> identifier
primary-expression -> **(** *expression* **)**

are re-written as:

(***** *primary-expresion* **) .** *identifier*
(***** *primary-expression* **) . (** *expression* **)**

The unary operator ***** used here is the indirection operator, its meaning is discussed later.

The index operators **:** and **:^** index the primary expression to discover a function, the result of the operation is a callable method. These operators and methods are discussed in more detail below.

The last of the *primary-operation*s:

(*expression-list* **)**

is the call operation. Although, as usual, no type checking is performed by the parser; at execution time the thing it is applied to must be callable. For example:

```
my_function(1, 2, "a string")
```

and

```
xxx.array_of_funcs[10]()
```

are both function calls. Function calls will be discussed in more detail below.

This concludes the examination of a *primary-expression*.

Terms

Primary-expressions are combined with prefix and postfix unary operators to make terms:

term	*[prefix-operator...] primary-expression [postfix-operator...]*
prefix-operator	*Any of:* *** & - + ! ~ ++ -- @ $**
postfix-operator	*Any of:* ++ --

That is, a *term* is a *primary-expression* surrounded on both sides by any number of prefix and postfix operators. Postfix operators bind more tightly than prefix operators. Both types bind right-to-left when concatenated together. That is: -!x is the same as -(!x). As in all expression compilation, no type checking is performed by the parser, because types are an execution-time consideration.

Some of these operators touch on subjects not yet explained and so will be dealt with in detail in later sections. But in summary:

Prefix operators

*****	Indirection; applied to a pointer, gives target of the pointer.
&	Address of; applied to any lvalue, gives a pointer to it.
-	Negation; gives negative of any arithmetic value.
+	Positive; no real effect.
!	Logical not; applied to 0 or NULL, gives 1, else gives 0.
~	Bit-wise complement.
++	Pre-increment; increments an lvalue and gives new value.
--	Pre-decrement; decrements an lvalue and gives new value.
@	Atomic form of; gives the (unique) read-only version of any value.
$	Immediate evaluation; see below.

Postfix operators

++	Post-increment; increments an lvalue and gives old value.
--	Post-increment; decrements an lvalue and gives old value.

One of these operators, $, is only a pseudo-operator. It actually has its effect entirely at parse time. The $ operator causes its subject expression to be evaluated immediately by the parser and the result of that evaluation substituted in its place. This is used to speed later execution, to protect against later scope or variable changes, and to construct constant values which are better made with running code than literal constants. For example, an expression involving the square root of two could be written as:

```
x = y + 1.414213562373095;
```

Or it could be written more clearly, and with less chance of error, as:

```
x = y + sqrt(2.0);
```

But this construct will call the square root function each time the expression is evaluated. If the expression is written as:

```
x = y + $sqrt(2.0);
```

The square root function will be called just once, by the parser, and will be equivalent to the first form.

When the parser evaluates the subject of a $ operator it recursively invokes the execution engine to perform the evaluation. As a result there is no restriction on the activity which can be performed by the subject expression. It may reference variables, call functions or even read files. But it is important to remember that it is called at parse time. Any variables referenced will be immediately interrogated for their current value. Automatic variables of any expression which is contained in a function will not be available, because the function itself has not yet been invoked; in fact it is clearly not yet even fully parsed.

The $ operator as used above increased speed and readability. Another common use is to avoid later re-definitions of a variable. For instance:

```
($printf)("Hello world\n");
```

Will use the *printf* function which was defined at the time the statement was parsed, even if it is latter re-defined to be some other function. It is also slightly faster, but the difference is small when only a simple variable look-up is involved. Notice the bracketing which has been used to bind the *$* to the word *printf*. Function calls are primary operations so the *$* would have otherwise referred to the whole function call as it did in the first example.

This concludes our examination of a *term* (remember that the full meaning of other prefix and postfix operators will be discussed in later sections).

Binary operators

We will now turn to the top level of expressions where *term*s are combined with binary operators:

expression	*term*
	expression infix-operator expression
infix-operator	*Any of:*
	@
	*** / %**
	+ -
	>> <<
	< > <= >=

```
== != ~ !~ ~~ ~~~
&
^
|
&&
||
:
?
= += -= *= /= %= >>= <<= &= ^= |= ~~= <=>
,
```

That is, an *expression* can be a simple *term*, or two *expressions* separated by an *infix-operator.* The ambiguity amongst expressions built from several binary-operator separated expressions is resolved by assigning each operator a precedence and also applying rules for order of binding amongst equal precedence levels[2]. The lines of binary operators in the syntax rules above summarise their precedence. Operators on higher lines have higher precedence than those on lower lines. Thus 1+2*3 is the same as 1+(2*3). Operators which share a line have the same precedence. All operators except those on the second last line group left-to-right. Those on the second last line (the assignment operators) group right-to-left. Thus

```
a * b / c
```

is the same as:

```
(a * b) / c
```

But:

```
a = b += c
```

is the same as:

```
a = (b += c)
```

As with unary operators, the full meaning of each will be discussed in a later section. But in summary:

2. The precedences and rules are identical to those of C.

Binary operator summary

@	Form pointer
*****	Multiplication, Set intersection
/	Division
%	Modulus
+	Addition, Set union
-	Subtraction, Set difference
>>	Right shift (shift to lower significance)
<<	Left shift (shift to higher significance)
<	Logical test for less than, Proper subset
>	Logical test for greater than, Proper superset
<=	Logical test for less than or equal to, Subset
>=	Logical test for greater than or equal to, Superset
==	Logical test for equality
!=	Logical test for inequality
~	Logical test for regular expression match
!~	Logical test for regular expression non-match
~~	Regular expression sub-string extraction
~~~	Regular expression multiple sub-string extraction
&	Bit-wise and
^	Bit-wise exclusive or
|	Bit-wise or
&&	Logical and
||	Logical or
:	Choice separator (must be right hand subject of **?** operator)
?	Choice (right hand expression must use **:** operator)
=	Assignment
+=	Add to
-=	Subtract from
***=**	Multiply by
/=	Divide by
%=	Modulus by
>>=	Right shift by
<<=	Left shift by
&=	And by
^=	Exclusive or by
|=	Or by
~~=	Replace by regular expression extraction
<=>	Swap values
,	Multiple expression separator

This concludes our consideration of *expression*s.

Statements

We will now move on to each of the executable statement types in turn.

Simple expression statements

The simple expression statement:

expression **;**

Is just an expression followed by a semicolon. The parser translates this expression to its executable form. Upon execution the expression is evaluated and the result discarded. Typically the expression will have some side-effect such as assignment, or make a function call which has a side-effect, but there is no explicit requirement that it do so. Typical expression statements are:

```
printf("Hello world.\n");
x = y + z;
++i;
```

Note that an expression statement which could have no side-effects other than producing an error may be completely discarded and have no code generated for it.

Compound statements

The compound statement has the form:

{ *statement...* **}**

That is, a compound statement is a series of any number of statements surrounded by curly braces. Apart from causing all the sub-statements within the compound statement to be treated as a syntactic unit, it has no effect. Thus:

```
printf("Line 1\n");
{
     printf("Line 2\n");
     printf("Line 3\n");
}
printf("Line 4\n");
```

When run, will produce:

```
Line 1
Line 2
Line 3
Line 4
```

Note that the parser will not return control to the execution engine until all of a top-level compound statement has been parsed. This is true in general for all other statement types.

The *if* statement

The *if* statement has two forms:

if (*expression* **)** *statement*
if (*expression* **)** *statement* **else** *statement*

The parser converts both to an internal form. Upon execution, the *expression* is evaluated. If the expression evaluates to anything other than 0 (integer zero) or NULL, the following statement is executed; otherwise it is not. In the first form this is all that happens, in the second form, if the expression evaluated to 0 or NULL the statement following the *else* is executed; otherwise it is not.

The interpretation of both 0 and NULL as false, and anything else as true, is common to all logical operations in ICI. There is no special boolean type.

The ambiguity introduced by multiple if statements with an lesser number of else clauses is resolved by binding else clauses with their closest possible if. Thus:

```
if (a) if (b) dox(); else doy();
```

If equivalent to:

```
if (a)
{
     if (b)
          dox();
     else
          doy();
}
```

The *while* statement

The *while* statement has the form:

while (*expression* **)** *statement*

The parser converts it to an internal form. Upon execution a loop is established. Within the loop the *expression* is evaluated, and if it is false (0 or NULL) the loop is terminated and flow of control continues after the *while* statement. But if the *expression* evaluates to true (not 0 and not NULL) the *statement* is executed and then flow of control moves back to the start of the loop where the test is performed again (although other statements, as seen below, can be used to modify this natural flow of control).

The *do-while* statement

The *do-while* statement has the following form:

do statement **while (** expression **) ;**

The parser converts it to an internal form. Upon execution a loop is established. Within the loop the *statement* is executed. Then the *expression* is evaluated and if it evaluates to true, flow

of control resumes at the start of the loop. Otherwise the loop is terminated and flow of control resumes after the *do-while* statement.

The *for* statement

The *for* statement has the form:

for (*[expression]***;** *[expression]***;** *[expression]* **)** *statement*

The parser converts it to an internal form. Upon execution the first *expression* is evaluated (if present). Then, a loop is established. Within the loop: If the second *expression* is present, it is evaluated and if it is false the loop is terminated. Next the *statement* is executed. Finally, the third *expression* is evaluated (if present) and flow of control resumes at the start of the loop. For example:

```
for (i = 0; i < 4; ++i)
     printf("Line %d\n", i);
```

When run will produce:

```
Line 0
Line 1
Line 2
Line 3
```

The *forall* statement

The *forall* statement has the form:

forall (*expression [* **,***expression]* **in** *expression* **)** *statement*

The parser converts it to an internal form. In doing so the first and second *expression*s are required to be lvalues (that is, capable of being assigned to). Upon execution the first expression is evaluated and that storage location is noted. If the second *expression* is present the same is done for it. The third *expression* is then evaluated and the result noted; it must evaluate to an array, a set, a struct, a string, or NULL; we will call this *the aggregate*. If this is NULL, the *forall* statement is finished and flow of control continues after the statement; otherwise, a loop is established.

Within the loop, an element is selected from the noted aggregate. The value of that element is assigned to the location given by the first expression. If the second expression was present, it is assigned the key used to access that element. Then the statement is executed. Finally, flow of control resumes at the start of the loop.

Each arrival at the start of the loop will select a different element from the aggregate. If no as yet unselected elements are left, the loop terminates. The order of selection is predictable for arrays and strings, namely first to last. But for structs and sets it is unpredictable. Also, while changing the values of the structure members is acceptable, adding or deleting keys, or adding or deleting set elements during the loop will have an unpredictable effect on the progress of the loop.

As an example:

```
forall (colour in [array "red", "green", "blue"])
```

```
    printf("%s\n", colour);
```

when run will produce:

```
red
green
blue
```

And:

```
forall (value, key in [struct a = 1, b = 2, c = 3])
    printf("%s = %d\n", key, value);
```

when run will produce (possibly in some other order):

```
c = 3
a = 1
b = 2
```

Note in particular the interpretation of the value and key for a set. For consistency with the access method and the behavior of structs and arrays, the values are all 1 and the elements are regarded as the keys, thus:

```
forall (value, key in [set "a", "b", "c"])
    printf("%s = %d\n", key, value);
```

when run will produce:

```
c = 1
a = 1
b = 1
```

But as a special case, when the second expression is omitted, the first is set to each "key" in turn, that is, the elements of the set. Thus:

```
forall (element in [set "a", "b", "c"])
    printf("%s\n", element);
```

when run will produce:

```
c
a
b
```

When a forall loop is applied to a string (which is not a true aggregate), the "sub-elements" will be successive one character sub-strings.

Note that although the sequence of choice of elements from a set or struct is at first examination unpredictable, it will be the same in a second forall loop applied without the structure or set being modified in the interim.

The *switch*, *case*, and *default* statements

These statements have the forms:

switch (*expression* **)** *compound-statement*

case *expression* **:**
default **:**

The parser converts the switch statement to an internal form. As it is parsing the compound statement, it notes any *case* and *default* statements it finds at the top level of the compound statement. When a *case* statement is parsed the *expression* is evaluated immediately by the parser. As noted previously for parser evaluated expressions, it may perform arbitrary actions, but it is important to be aware that it is resolved to a particular value just once by the parser. As the *case* and *default* statements are seen their position and the associated expressions are noted in a table.

Upon execution, the *switch* statement's *expression* is evaluated. This value is looked up in the table created by the parser. If a matching *case* statement is found, flow of control immediately moves to immediately after that *case* statement. If there is a *default* statement, flow of control immediately moves to just after that. If there is no matching *case* and no *default* statement, flow of control continues just after the entire *switch* statement.

For example:

```
switch ("a string")
{
case "another string":
     printf("Not this one.\n");
case 2:
     printf("Not this one either.\n");
case "a string":
     printf("This one.\n");
default:
     printf("And this one too.\n");
}
```

When run will produce:

```
This one.
And this one too.
```

Note that the case and default statements, apart from the part they play in the construction of the look-up table, do not influence the executable code of the compound statement. Notice that once flow of control had transferred to the third case statement above, it continued through the default statement as if it had not been present. This behavior can be modified by the *break* statement described below.

It should be noted that the "match" used to look-up the switch expression against the case expressions is the same as that used for structure element look-up. That is, to match, the switch expression must evaluate to the same object as the case expression. The meaning of this will be made clear in a later section.

The *break* and *continue* statements

The *break* and *continue* statements have the form:

break ;
continue ;

The parser converts these to an internal form. Upon execution of a break statement the execution engine will cause the nearest enclosing loop (a while, do, for or forall) or switch statement within the same scope to terminate. Flow of control will resume immediately after the affected statement. Note that a *break* statement without a surrounding loop or *switch* in the same function or module is illegal.

Upon execution of a *continue* statement the execution engine will cause the nearest enclosing loop to move to the next iteration. For *while* and *do* loops this means the test. For *for* loops it means the step, then the test. For *forall* loops it means the next element of the aggregate.

The *return* statement

The *return* statement has the form:

return *[expression]* **;**

The parser converts this to an internal form. Upon execution, the execution engine evaluates the *expression* if it is present. If it is not, the value NULL is substituted. Then the current function terminates with that value as its apparent value in any expression it is embedded in. It is an error for there to be no enclosing function.

The *try* statement

The *try* statement has the form:

try *statement* **onerror** *statement*

The parser converts this to an internal form. Upon execution, the first *statement* is executed. If this statement executes normally flow continues after the *try* statement; the second *statement* is ignored. But if an error occurs during the execution of the first *statement* control is passed immediately to the second *statement*.

Note that "during the execution" applies to any depth of function calls, even to other modules or the parsing of sub-modules. When an error occurs both the parser and execution engine unwind as necessary until an error catcher (that is, a *try* statement) is found.

Errors can occur almost anywhere and for a variety of reasons. They can be explicitly generated with the *fail* function (described below), they can be generated as a side-effect of execution (such as division by zero), and they can be generated by the parser due to syntax or semantic errors in the parsed source. For whatever reason an error is generated, a message (a string) is always associated with it.

When any otherwise uncaught error occurs during the execution of the first *statement*, two things are done:

- Firstly, the string associated with the failure is assigned to the variable *error*. The assignment is made as if by a simple assignment statement within the scope of the *try* statement.
- Secondly, flow of control is passed to the statement following the *onerror* keyword.

Once the second *statement* finishes execution, flow of control continues as if the whole *try* statement had executed normally.

For example:

```
static
div(a, b)
{
     try
          return a / b;
     onerror
          return 0;
}

printf("4 / 2 = %d\n", div(4, 2));
printf("4 / 0 = %d\n", div(4, 0));
```

When run will print:

```
4 / 2 = 2
4 / 0 = 0
```

The handling of errors which are not caught by any *try* statement is implementation dependent. A typical action is to prepend the file and line number on which the error occurred to the error string, print this, and exit.

The *critsect* statement

The *critsect*, or "critical section", statement has the form:

critsect *statement*

The parser converts this to an internal form. Upon execution, the *statement* is executed indivisibly with respect to other threads. Thus:

```
critsect x = x + 1;
```

will increment *x* by 1, even if another thread is doing similar increments. Without the use of the *critsect* statement we could encounter a situation where both threads read the current value of *x* (say 2) at the same time, then both added 1 and stored the result 3, rather than one thread incrementing the value to 3, then the other to 4.

The indivisibility bestowed by a *critsect* statement applies as long as the code it dominates is executing, including all functions that code calls. Even operations that block (such as the *waitfor* statement) will be effected. The indivisibility will be revoked once the *critsect* statement completes. Either through completing normally, or through an error being thrown by the code it is dominating.

The *waitfor* statement

The *waitfor* statement has the form:

waitfor (*expression* **;** *expression* **)** *statement*

The parser converts this to an internal form. Upon execution, a critical section is established that extends for the entire scope of the *waitfor* statement (except for the special exception explained below). Within the scope of this critical section, the *waitfor* statement repeatedly evaluates the first *expression* until it is true (that is, neither 0 nor NULL). Once the first *expression* evaluates to true, control passes to the *statement* (still within the scope of the critical sec-

tion) After executing *statement* the critical section is released and the *waitfor* statement is finished.

However, each time the first *expression* evalutes to a false value, the second *expression* is evaluated and the object that it evaluates to is noted. Then, idivisibly, the current thread sleeps waiting for that object to be signaled (by a call to the *wakeup()* function), and the critical section is supressed (thus allowing other thread to run). The thread will remain asleep until it is woken up by a call to *wakeup()* with the given object as an argument. Each time this occurs, the critical section is again enforced and the process repeats with the evaluation and testing of the first *expression*. While the thread is asleep it consumes no significant CPU time.

The *waitfor* statement is the basic method of inter-thread communication and control in ICI. It is typically used to gate control of some data that is passing from one thread to another. For example, suppose *jobs* is an array that is shared between two processes. In one thread we might write:

```
waitfor (nels(jobs) > 0; jobs)
     job = rpop(jobs);
/*
 * Process job...
 */
```

While in a second thread that is generating jobs we might write:

```
push(jobs, new_job);
wakeup(jobs);
```

In this example, the list object *jobs* is the object we are using to wait on and wakeup, but any object can be used. One technique is to use a commonly agreed string (strings being intrinsically atomic, will naturally be the same object without any explicit commnality between the threads). In some circumstances it may be necessary to apply a *critsect* to the access to the shared data (*jobs* in this example) in the thread doing the waking up.

It if very important to only perform the call to *wakeup()* after the condition that allows release of the wait has been established. To illustrate, suppose we had written:

```
wakeup(jobs);                /* WRONG */
enque(jobs, new_job);        /* WRONG */
```

In this case the waiting thread may have run between the two statements, evaluated the test to false, and gone to sleep again, possible never to wake.

Similarly, the waitfor condition must be a true reflection of a condition that implies a wakeup will occur, at some stage, on the object being waited on. Do not assume that because the thread has woken up, the wakeup has been for the expected reason. For example, it would be wrong to write:

```
wait_once = 0;
waitfor (wait_once++; jobs)      /* WRONG */
     jobs = deque(jobs);
```

The null statement

The null statement has the form:

```
;
```

The parser may convert this to an internal form. Upon execution it will do nothing.

Declaration statements

There are two types of declaration statements:

declaration — *storage-class declaration-list* **;**
storage-class identifier function-body

storage-class — **extern**
static
auto

The first is the general case while the second is an abbreviated form for function definitions. Declaration statements are syntactically equal to any other statement, but their effect is made entirely at parse time. They act as null statements to the execution engine. There are no restriction on where they may occur, but their effect is a by-product of their parsing, not of any execution.

Declaration statements must start with one of the *storage-class* keywords listed above[3]. Considering the general case first, we next have a *declaration-list*.

declaration-list — *identifier [= expression]*
declaration-list **,** *identifier [* **=** *expression]*

That is, a comma separated list of identifiers, each with an optional initialisation, terminated by a semicolon. For example:

```
static a, b = 2, c = [array 1, 2, 3];
```

The storage class keyword establishes which scope the variables in the list are established in, as discussed earlier. Note that declaring the same identifier at different scope levels is permissible and that they are different variables.

A declaration with no initialisation first checks if the variable already exists at the given scope. If it does, it is left unmodified. In particular, any value it currently has is undisturbed. If it does not exist it is established and is given the value NULL.

A declaration with an initialisation establishes the variable in the given scope and gives it the given value even if it already exists and even if it has some other value.

Note that initial values are parser evaluated expressions. That is they are evaluated immediately by the parser, but may take arbitrary actions apart from that. For example:

```
static
fibonacci(n)
{
     if (n <= 1)
          return 1;
     return fibonacci(n - 1) + fibonacci(n - 2);
}
```

3. Note that, unlike C, function definitions must be prefixed by a storage class. As executable code may occur anywhere, this is required to distinguish them from a function call.

```
static fib10 = fibonacci(10);
```

The declaration of *fib10* calls a function. But that function has already been defined so this will work.

Note that the scope of a static variable is (normally) the entire module it is parsed in. For example:

```
static
func()
{
     static aStatic = "The value of a static.";
}

printf("%s\n", aStatic);
```

when run will print:

```
The value of a static.
```

That is, despite being declared within a function, the declaration of *aStatic* has the same effect as if it had been declared outside the function. Also notice that the function has not been called. The act of parsing the function caused the declaration to take effect.

The behavior of extern variables has already been discussed, that is, they are declared as static in the parent module. The behavior of auto variables, and in particular their initialisation, will be discussed in a later section.

Abbreviated function declarations

As seen above there are two forms of declaration. The second:

storage-class identifier function-body

is a shorthand for:

storage-class identifier = **[func** *function-body* **] ;**

and is the normal way to declare simple functions. Examples of this have been seen above.

Functions

As with most ICI constructs there are two parts to understanding functions; how they are parsed and how they execute.

When a function is parsed four things are noted:

- the names and positions of the formal parameters;
- the names and initialisation of auto variables;
- the static scope or class in which the function is declared;
- the code generated by the statements in the function.

The formal parameters (that is, the identifiers in the bracket enclosed list just before the compound statement) are actually implicit auto variable declarations. Each of the identifiers is declared as an auto variable without an initialisation, but in addition, its name and position in the list is noted.

Upon execution (that is, upon a function call), the following takes place:

- The auto variables, as noted by the parser, along with any initialisations, are copied as a group. This copy forms the auto variables of this invocation.
- Any actual parameters (that is, expressions provided by the caller) are matched positionally with the formal parameter names, and the value of those expressions are assigned to the auto variables of those names.
- If there were more actual parameters than formal parameters, and there is an auto variable called *vargs*, the remaining argument values are formed into an array which is assigned to *vargs*.
- If this is a a method call (see below) the auto variable *this* is set to the subject object of the call, and the auto variable *class* is set to the class (if any).
- The variable-scope is set such that the auto variables are the inner-most scope.
- Successive outer scopes are set to the static scope, or, if this is a method call, the class noted when the function was parsed.
- The flow of control is diverted to the code generated by parsing the function.

A *return* statement executed within the function will cause the function to return to the caller and act as though its value were the expression given in the return statement. If no expression was given in the return statement, or if execution fell through the bottom of the function, the apparent return value is NULL. In any event, upon return the scope is restored to that of the caller. All internal references to the group of automatic variables are lost (although as will be seen later explicit program references may cause them to remain active).

Simple functions have been seen in earlier examples. We will now consider further issues.

It is very important to note that the parser generates a prototype set of auto variables which are copied, along with their initial values, when the function is called. The value which an auto variable is initialised with is a parser evaluated expression just like any other initialisation. It is not evaluated on function entry. But on function entry the value the parser determined is used to initialise the variable. For example:

```
static myVar = 100;

static
myFunc()
{
     auto anAuto = myVar;

     printf("%d\n", anAuto);
     anAuto = 500;
}

myFunc();
myVar = 200;
myFunc();
```

When run will print:

```
100
100
```

Notice that the initial value of *anAuto* was computed just once, changing *myVar* before the second call did not affect it. Also note that changing *anAuto* during the function did not affect its subsequent re-initialisation on the next invocation.

As stated above, formal parameters are actually uninitialised auto variables. Because of the behavior of variable declarations it is possible to explicitly declare an auto variable as well as include it in the formal parameter list. In addition, such an explicit declaration may have an initialisation. In this case, the explicit initialisation will be effective when there is no actual parameter to override it. For example:

```
static
print(msg, file)
{
     auto file = stdout; /* Default value. */

     fprintf(file, "%s\n", msg);
}

print("Hello world");
print("Hello world", stderr);
```

In the first call to the function *print* there is no second actual parameter. In this case the explicit initialisation of the auto variable *file* (which is the second formal parameter) will have its effect unmolested. But in the second call to *print* a second argument is given. In this case this value will over-write the explicit initialisation given to the argument and cause the output to go to *stderr*.

As indicated above there is a mechanism to capture additional actual parameters which were not mentioned in the formal parameter list. Consider the following example:

```
static
sum()
{
     auto vargs;
     auto total = 0;
     auto arg;

     forall (arg in vargs)
          total += arg;
     return total;
}

printf("1+2+3 = %d\n", sum(1, 2, 3));
printf("1+2+3+4 = %d\n", sum(1, 2, 3, 4));
```

Which when run will produce:

```
1+2+3 = 6
1+2+3+4 = 10
```

In this example the unmatched actual parameters were formed into an array and assigned to the auto variable *vargs*, a name which is recognised specially by the function call mechanism.

And also consider the following example where a default initialisation to *vargs* is made. In the following example the function *call* is used to invoke a function with an array of actual parameters, the function *array* is used to form an array at run-time, and addition is used to concatenate arrays; all these features will be further explained in later sections:

```
static
debug(fmt)
{
     auto fmt = "Reached here.\n";
     auto vargs = [array];

     call(fprintf, array(stderr, fmt) + vargs);
}

debug();
debug("Done that.\n");
debug("Result = %d, total = %d.\n", 123, 456);
```

When run will print:

```
Reached here.
Done that.
Result = 123, total = 456.
```

In the first call to *debug* no arguments are given and both explicit initialisations take effect. In the second call the first argument is given, but the initialisation of *vargs* still takes effect. But in the third call there are unmatched actual parameters, so these are formed into an array and assigned to *vargs*, overriding its explicit initialisation.

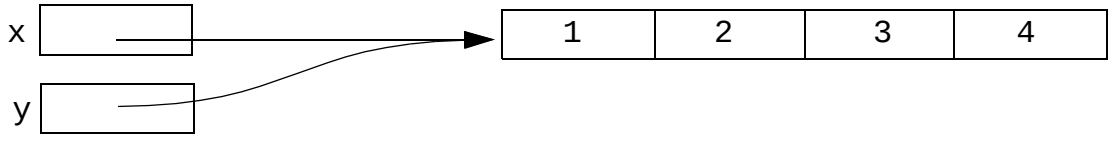
Objects

Up till now few exact statements about the nature of values and data have been made. We will now examine values in more detail. Consider the following code fragment:

```
static x;
static y;

x = [array 1, 2, 3, 4];
y = x;
```

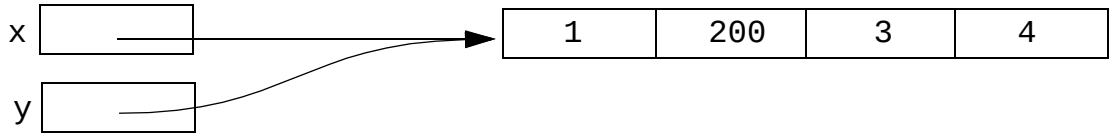
After execution of this code the variable *x* refers to an array. The assignment of *x* to *y* causes *y* to refer to the same array. Diagrammatically:



If the assignment:

```
y[1] = 200;
```

is performed, the result is:

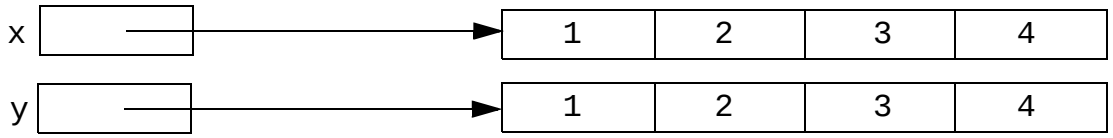


We say that *x* and *y* refer to the same object. Now consider the following code fragment:

```
static x;
static y;

x = [array 1, 2, 3, 4];
y = [array 1, 2, 3, 4];
```

Diagrammatically:

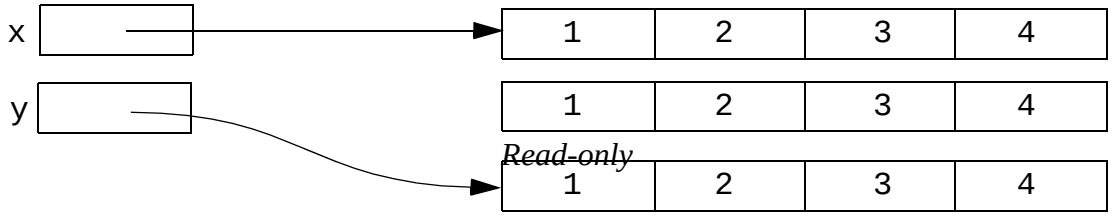


In this case, *x* and *y* refer to different objects, despite that fact they are equal.

Now consider one of the unary operators which was only briefly mentioned in the sections above. The @ operator returns a read-only version of the sub-expression it is applied to. Consider the following statement:

```
y = @y;
```

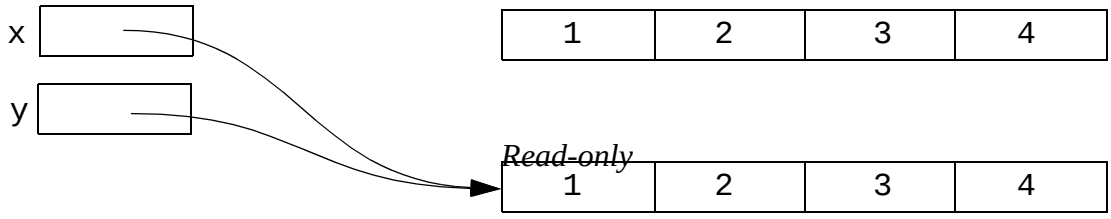
After this has been executed the result could be represented diagrammatically as:



The middle array now has no reference to it and the memory associated with it will be collected by the interpreter's standard garbage collection mechanism. Now consider the following statement:

```
x = @x;
```

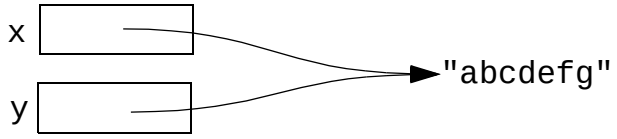
This is similar to the previous statement, except that this time *x* is replaced by a read-only version of its old value. But the result of this operation is:



Notice that *x* now refers to the same read-only array that *y* refers to. This is a fundamental property of the @ operator. It returns *the unique* read-only version of its argument value. Such read-only objects are referred to as *atomic* objects. The array which *x* used to refer to was non-atomic, but the array it refers to now is an atomic array. Aggregate types such as arrays, sets and structs are generally non-atomic, but atomic versions can be obtained (as seen above). But most other types, such as integers floats, strings and functions are intrinsically atomic. That is, no matter how a number, say 10, is generated, it will be the same object as every other number 10 in the interpreter. For-instance, consider the following example:

```
x = "ab" + "cdefg";
y = "abcde" + "fg";
```

After this is executed the situation can be represented diagrammatically as:



It is important to understand when objects are the same object, when they are different and the effects this has.

Equality

We saw above how two apparently identical arrays were each distinct objects. But these two arrays were *equal* in the sense of the equality testing operator ==. If two values are the same object they are said to be *eq*[4], and there is a function of that name to test for this condition. Two objects are *equal* (that is ==) if:

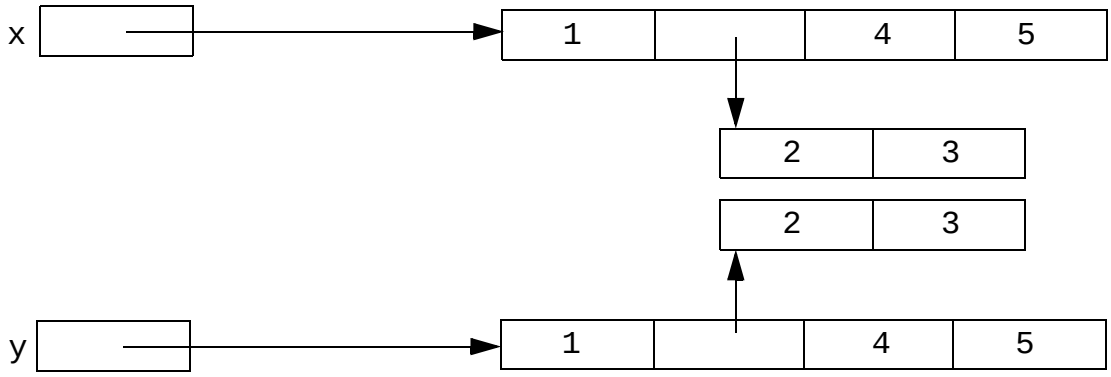
- they are the same object (i.e. *eq*); or
- they are both arithmetic (int and float) and have equivalent numeric values; or
- they are aggregates of the same type and all the sub-elements are the same objects (i.e. *eq)*.

This definition of equality is the basis for resolving the merging of aggregates into unique read-only (atomic) versions. Two aggregates will resolve to the same atomic object if they are *equal*. That is, they must contain exactly the same objects as sub-elements, not just equal objects. For example:

4.As in LISP.

```
static x = [array 1, [array 2, 3], 4, 5];
static y = [array 1, [array 2, 3], 4, 5];
```

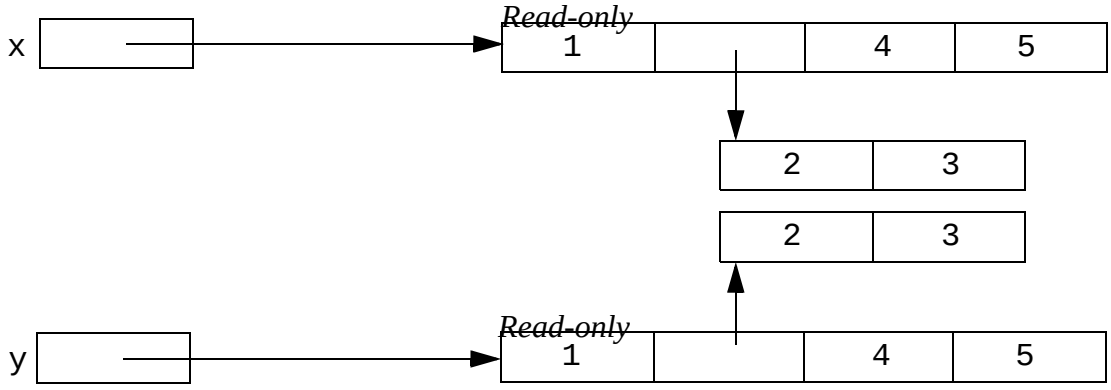
Could be represented diagrammatically as:



Now, if the following statements were executed:

```
x = @x;
y = @y;
```

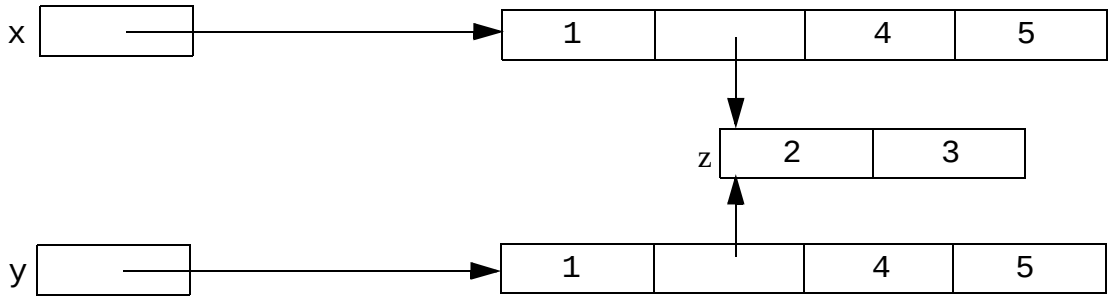
The result could be represented diagrammatically as:



That is, both *x* and *y* refer to new read-only objects, but they refer to different read-only objects because they have an element which is not the same object. The simple integers are the same objects because integers are intrinsically atomic objects. But the two sub-arrays are distinct objects. Being equal was not sufficient. The top-level arrays needed to have exactly the same objects as contents to make *x* and *y* end up referring to the same read-only array. In contrast to this consider the following similar situation:

```
static z = [array 2, 3];
static x = [array 1, z, 4, 5];
static y = [array 1, z, 4, 5];
```

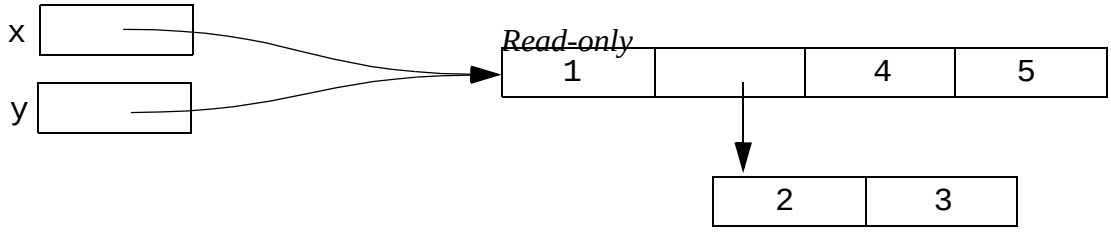
This could be represented diagrammatically as:



Now, if the following statements were executed:

```
x = @x;
y = @y;
```

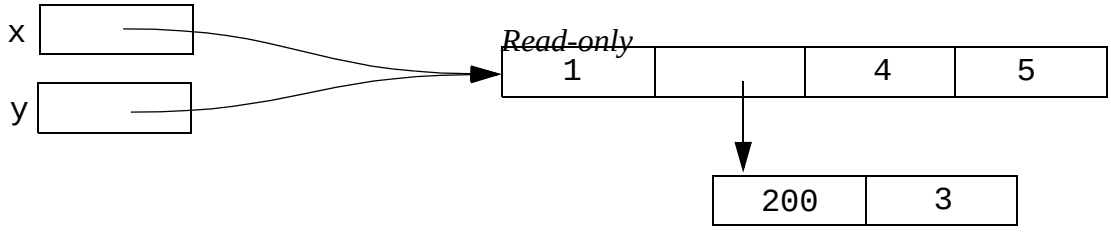
The result could be represented diagrammatically as:



In this case both *x* and *y* refer to the same read-only array because the original arrays where equal, that is, all their elements were the same objects. Notice that one of the elements is still a *writeable* array. The read-only property is only referring to the top level array. The sub-array can be changed, but the reference to it from the top level array can not. Thus:

```
x[1][0] = 200;
```

will result in:



whereas the statement:

```
x[1] = 200;
```

will just result in an error.

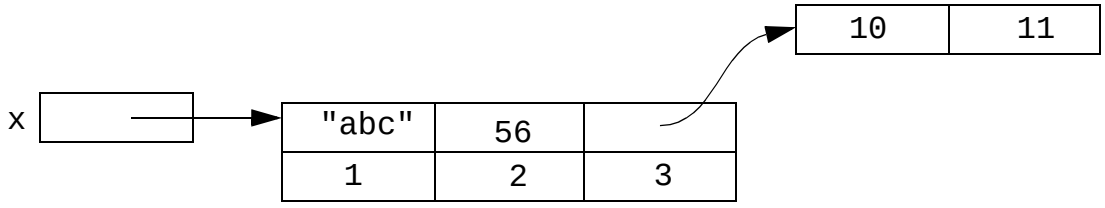
Structure and set keys

Any object, not just a string, can be used as a key in a structure. For instance:

```
static x = [struct];
static z = [array 10, 11];

x["abc"] = 1;
x[56] = 2;
x[z] = 3;
```

Could be represented diagrammatically as:



And the assignment:

```
x[z] = 300;
```

would replace the *3* in the above diagram with *300*. But the assignment:

```
x[[array 10, 11]] = 300;
```

would result in a new element being added to the structure because the array given in the above statement is a different object from the one which *z* refers to.

Similarly, elements of sets may be any objects.

Indexing structures by complex aggregates is as efficient as indexing by intrinsically atomic types such as strings and integers.

Structure super types

Up till now structures have been described as simple lookup tables which map a key, or index, to a value. But a structure may have associated with it a *super structure*.

The function *super* can be used to discover the current super of a struct and to set a new super. With just one argument it returns the current super of that struct, with a second argument it also replaces the super by that value.

When a key is being looked-up in a structure for reading, and it is not found and there is a *super struct*, the key is further looked for in the super struct, if it is found there its value from that struct is returned. If it is not found it will be looked for in the next super struct etc. If no structures in the *super chain* contain the key, the special value NULL is returned.

When a key is being looked up in a structure for writing, it will similarly be searched for in the super chain. If it is found in a writeable structure the value in the structure in which it was found will be set to the new value. If it was never found, it will be added along with the given value to the very first struct, that is, the structure at the base, or root, of the super chain.

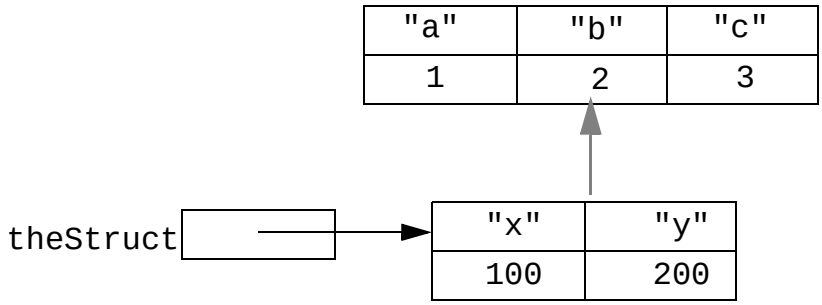
Consider the following example:

```
static theSuper = [struct a = 1, b = 2, c = 3];
```

```
static theStruct = [struct x = 100, y = 200];

super(theStruct, theSuper);
```

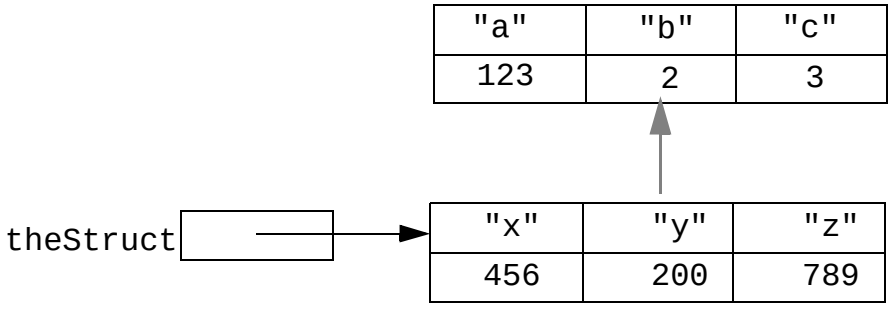
After this statement the situation could be represented diagrammatically as:



then if the following statements were executed:

```
theStruct.a = 123;
theStruct.x = 456;
theStruct.z = 789;
```

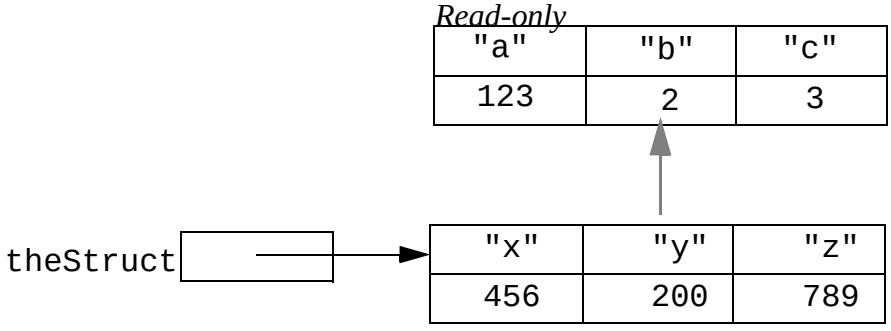
the situation could be diagrammatically represented as:



If a super struct is not writeable (that is, it is atomic) values will not be written in it and will lodge in the base structure instead. Thus consider what happens if we replace the super structure in the previous example by its read-only version:

```
super(theStruct, @theSuper);
```

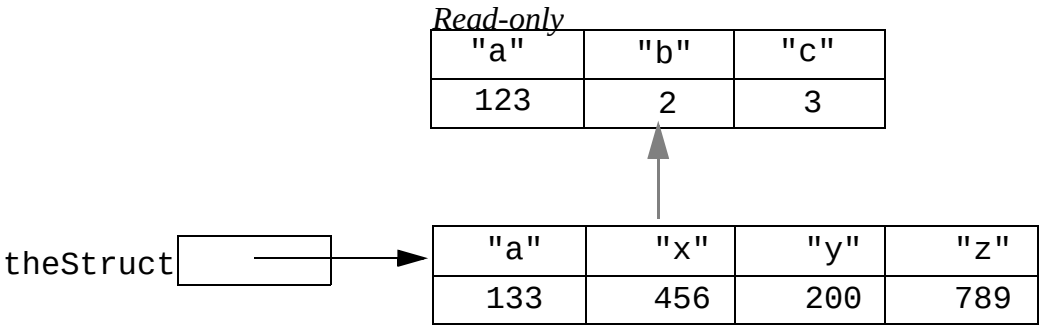
The situation could now be represented diagrammatically as:



If the assignment statement:

```
theStruct.a += 10;
```

were executed, the value of the element *a* will first be *read* from the super structure, this value will then have ten added to it, and the result will be *written* back into the base structure; because the super structure is read-only and cannot be modified. The finally situation can be represented diagrammatically as:



Note that many structs may share the same super struct. Thus a single read-only super struct can be used hold initial values; saving explicit initialisations and storage space.

The function *assign* may be used to set a value in a struct explicitly, without reference to any super structs; and the function *fetch* may be used to read a value from a struct explicitly, without reference to any super structs.

Within a *struct-literal* a colon prefixed expression after the *struct* identifier is used as the super struct. For example, the declarations used in the previous example could be written as:

```
static theSuper = [struct a = 1, b = 2, c = 3];
static theStruct = [struct:theSuper, x = 100, y = 200];
```

An aside on variables and scope

Now that structs and their super have been described a more precise statement about variables and scope can be made.

ICI variables are entries in ordinary structs. At all times, the execution context holds a reference to a struct that is the current scope for the lookup of simple variables. An un-adorned identifier in an expression is just an implicit reference to an element of the current scope structure. The inheritance and name hiding of the variable scope mechanism is a product of the super chain.

During both module parseing and function execution, the auto variables are the entries in the base structure. The super of this is the struct containing the static variables. The next super struct contains the externs, and successive super structs are successive outer scopes.

Auto, static and extern declarations make explicit assignments to the appropriate structure.

The function *scope* can be used to obtain the current scope structure; and to set it (use with care).

###But there is a difference in the handling of undefined entries. Whereas normal lookup of undefined entries in a structure produces a default value of NULL or implicit creation, the implicit lookup of undefined variables triggers an attempt to dynamically load a library to define the variable (see *Undefined variables and dynamic loading* below), and failing that, produce an error ("%s undefined").

Base types

ICI supports a base set of standard data types. Each is identified by a simple name. In summary these are:

array	An ordered sequence of objects.
file	An open file reference.
float	A double precision floating point number.
func	A function.
int	A signed 32 bit integer.
list	An ordered set of objects.
mem	References to raw machine memory.
method	A binding of a function and a subject object.
ptr	A reference to a storage location.
regexp	A compiled regular expression.
set	An unordered collection of objects.
string	An ordered sequence of 8 bit characters.
struct	An unordered set of mappings from one object to another.

Many of these base types have been alluded to in previous sections. The following sections describe each type in more detail.

It should be noted that indexing and calling are the only operations that are an intrinsic property of each base type. Other behaviours of base types are a product of operators and functions that perform their various functions when supplied with operands of particular types. For this reason the following descriptions typically describe what data an instance of each base type holds, what happens when it is indexed or called, and may briefly mention the functions and operators that are highly relevant to the type. See following sections on operators and core functions for a complete picture.

In the following text, the word "efficient" typically means in constant time or memory, although occasional internal housekeeping may occur.

array - An ordered sequence of objects

An array is a contigouse (in memory) block of object references. The first object is refered to with index zero, subsequent elements of the block are referenced by successive integers. The index must always be an integer, else the indexing operation will fail. Reading at indicies not in the block results in a NULL value. Writting at negative indicies fails, while writting at indicies beyond the current end of the block silently extends the block, and NULL fills the span between the old end and the newly written element. The function *nels()* can be used to reveal the number of elements currently in the array (which is also the index of the first element beyond the current length of the array).

Arrays offer the most memory efficient method of storing collections of objects.[5] The functions *push(), pop(), rpush(), rpop()* and *top()* are of note. They allow arrays to be used as efficient stacks and queues. They are all of constant order time[6]. Most other functions and operations on arrays are *O(n)*. For example, array addition is *O(n + m)* where *n* and *m* are the lengths of the two arrays.

The *rpush()* and *rpop()* functions push and pop items from the front of the array (that is, near index zero). But the first item is always considered to be at index zero. Pushing and poping items on the front of an array effectively changes the index of all the items in the array.

See also the functions *array()* to create an array at run-time, and the parse-time in-line literal form of arrays *[array ...]*. The function *sort()* can be used to sort the elements of an array.

Arrays form the fundamental basis for operand, execution and scope stacks in the ICI internal execution engine, as well as the storage of compiled code. Although this is not visible to the ICI programmer.

file - An open file reference

A file object is a reference and interface to some lower level file-like object. Most commonly a real file supported by the operating system, but not necessarilly so. The actual file object holds a reference to the basic file object, and references to its primitive access methods and operations. Those primitive methods are directly represented by the intrinsic ICI functions: *close(), eof(), flush(), getchar(), put()*, and *seek()*. In addition, the functions *getline(), getfile(), gettoken()*, and *gettokens()* efficiently build on these to read higher level constructs than simple bytes from a file. These functions can greatly increse the efficieny of file parsing over explicit per-character operations. The function *printf()* provides efficient formated output to files.

File objects are generally created by “open” functions; such as the archetypal *fopen()* function that opens or creates a host operating system file. Also note the *sopen()* function that allows an ICI string object to be opened as a file. The variables *stdin*, *stdout*, and *stderr* are generally created in the outer-most scope at interpreter startup and refer to the associated files of the current process. Also, various functions, such as *printf()*, will, if no explicit file argument is supplied, use the current value of the appropriate variable. These functions do this by looking up the name in the current scope, so it is possible to locally overide their default file usage.

5. On 32 bit machines, the raw per-element overhead is typically 4 bytes; although there is often slop at the end of the block to allow efficient growth.
6. Arrays are internally implemented as gowable circular buffers.

File objects are intrinsicly atomic based on identity of the lower level file structure and access methods. But in normal situations two seperate *fopen()* operations on the same operating system file result in different lower level structures and thus distinct file objects.

Files can not be indexed or called.

Note that an unreferenced file object will, eventually, be collected by the ICI interpreter's garbage collector, at which point it will be closed (if it is not already closed). But the indeterminate timming of garbage collections makes it inadvisable to rely on this mechanism to close files. In general, files should be explicitly closed to release lower level resources at a deterministic time. A file object is still a valid object after it has been closed, except no I/O operations will work on it any more.

(There is currently a pausity of functions to support reading and writting binary files. This will be corrected in future revisions. TML)

float - A double precision floating point number

A float hold a double precision floating point number (in the local machine's native format). Floats are intrinsically atomic based on their value (that is, all floats with a particular value are references to the same memory location). Floats can not be indexed or called and their utility is entirely based on the operators and functions that accept and return them.

func - A function

A function holds a reference to executable code, and a name suitable for diagnostics. In reality there are two types of function objects: Functions that reference native machine code, and functions the reference interpreted ICI code. But they are both called "func".

Function objects that link to interpreted ICI code also hold the names of the formal parameters and a prototype of the local scope structure that will be copied and used each time the function is invoked.

Function objects are intrinsically atomic based on the identity of all their components. The ICI parser also makes code atomic so in theory equal functions will be identical objects, but in practice such items as source file line information embedded in executable code frustrate this.

Function objects can, of course, be called. The semantics of this operation has been described above. Function objects are also the basis of methods in classes, the difference merely existing in their preparation by the parser, and the semantics of calling through a method.

Function objects can be indexed by some specific names to discover some of the internal elements. Specifically:

name	Returns a name that has been assigned to the function. In the case of the "abbreviated function declaraction" described above, it will be the identifier associated with the function. In the case of an in-line function literal, it will be the name *_funcname_*. In the case of a function implemented in native machine code, it will be an author assigned name.

autos	Returns the (atomic) prototype auto scope struct of this function, or NULL for functions implemented in native machine code. The super of this struct reveals the static scope of this function (the parsed class for a method).
args	Returns the (atomic) array of formal parameter names, or NULL for functions implemented in native machine code.

int - A signed 32 bit integer

An int hold a 32 bit signd integer (in the local machine's native format). Ints are intrinsically atomic based on their value (that is, all ints with a particular value are references to the same memory location). Ints can not be indexed or called and their utility is entirely based on the operators and functions that accept and return them.

mem - A reference to raw machine memory

A mem object references byte or word structured native machine memory. The mem object holds the base address of a region of raw machine memory, the wor-size it is to be accessed with (1, 2 or 4 bytes per word), and number of words that can be accessed.

The base address identifies the first word, and this can be accessed (as an integer) at index zero. Successive integers indexes reference successive words, up to the limit. Reading outside the bounds returns NULL, writting outside the bounds causes an error. Words are read and written in the native machine format (endieness in particular). One and two byte words are read as unsigned quantities. When writting, non-zero bits above the word size are simply discarded.

Mem objects can be used for simple, but dense, unstructured data storage. But they are most commonly used in interfaces to native machine code or hardware. The functions *mem()* and *alloc()* can be used to create mem objects. Although *mem()*, which allows access to arbitrary native machine addresses, may be disallowed in some systems. The function *alloc()* allocates memory for the mem object to refer to which is freed when the mem object is garbage collected.

Mem objects are intrinsically atomic based on the address, word size and number of elements.

method - A binding of a function and a subject object

A method holds a reference to a callable object, and a subject object. The subject object is typically a struct of some class (that is, its super is the class). The callable object is typically a function of the class or one of its super classes.

Method objects can be called, the semantics of which are described above.

Method objects can be indexed to discover their internal elements. Specifically:

subject	Returns the sibject object of the method. This is typically a struct.
callable	Returns the callable object of the method. This is typicaly a function.

Method objects are created by the ***:*** operator, typically preparatory to the invocation of a class function. But in most situations the parser will generate a special type of shortcut function invo-

cation to avoid the run-time creation of an ephemeral method object. So in practice method objects are quiet rare.

ptr - A reference to a storage location

Pointers are references to storage locations. Storage locations are the elements of anything which can be indexed. That is, array elements, set elements, struct elements and others. Variables (which are just struct elements) can be pointed to.

Pointers hold two objects, one is the object pointed into, the other is the key used to access the location in question.

The & operator is used to obtain a pointer to a location. Thus if the following were executed:

```
static x;
static y = [array 1, 2, 3];
static p1 = &x;
static p2 = &y[1];
```

The variable *p1* would be a pointer to *x* and the variables *p2* would be a pointer to the second element of *y*. Reference to the object a pointer points to can be obtained with the * operator. Thus if the following were executed:

```
*p1 = 123;
*p2 = 456;
printf("x = %d, y[1] = %d\n", x, y[1]);
```

the output would be:

```
x = 123, y[1] = 456
```

The generation of a pointer does not affect the location being pointed to. In fact the location may not even exist yet. When a pointer is referenced the same operation takes place as if the location was referenced explicitly. Thus a search down the super chain of a struct may occur, or an array may be extended to include the index being written to, etc.

In addition to simple indirection (that is the * operator), pointers may be indexed. But the index values must be an integer, and the key stored as part of the pointer must also be an integer. When a pointer is indexed, the index is added to the key which is stored as part of the pointer, the sum forms the actual index to use to when referencing the aggregate recorded by the pointer. For instance, continuing the example above:

```
p2[1] = 789;
```

would set the last element of the array to 789, because the pointer currently references element 1, and the given index is 1, and 1 + 1 is 2 which is the last element. The index arithmetic provided by pointers will work with any types, as long as the indexes are integers, thus:

```
static s = [struct (20) = 1, (30) = 2, (40) = 3];
static p = &s[30];

p[-10] = -1;
p[0] = -2;
p[10] = -3;
```

Would replace each of the elements in the struct *s* by their negative value.

Pointers can be called, but this is an obsolete facility and may be removed in future versions.

regexp - A compiled regular expression

A regexp object holds a regular expression and its compiled form. Regular expressions describe text patterns against which actual text can be matched to discover if the actual text matches the pattern. They can also be used to extract sub-strings of the actual text based on the pattern matching. For more details on the syntax and semantics of regular expressions, see the chapter on the subject below.

Regular expressions are created by the *regexp()* and *regexpi()* functions, and by the parser from regular expression literals (that is, *#...#*). Text can be matched against regular expressions by the operators *~*, *!~*, *~~*, and *~~~*, and by the functions *sub()*, *gsub()* and *smash()*.

Regular expressions can be indexed by two specific names:

pattern	Returns the original pattern as a string.
options	Returns an integer bit mask of the options applied in making the regular expression.

Regular expressions are intrinsically atomic based on the identity of the original pattern.

set - An unordered collection of objects

A set is an unordered collection of object references. Any single object can either be in a given set, or not in the set. It can not be in the set multiple times. Adding and removing objects from sets is an efficient constant time operation, and each distinct object in the set imposes a small fixed memory cost (both access speed and memory cost is slightly higher than the per element cost of an array). The type and complexity of an object being added or removed from a set has no effect on the efficiency of the operation.[7]

Sets can be used in different ways. In some circumstances they are used simply as unordered aggragates of other objects. In other circumstances they are used more as algebraic sets to record which objects have a certain property. In this regard they can be particularly useful because objects can be noted as having a particular property without modifying the internals of the object at all.

string - An ordered sequence of 8 bit characters

A string holds an ordered sequence of 8 bit characters. Strings are intrinsically atomic based on their value (that is, all strings with a particular value are references to the same memory location). Strings can be indexed by an *int* (read only) to reveal a one-character sub-string, or an empty string if negative or beyond the end of the string. Most of the utility of strings derive from the functions and operators than can be applied to them.

7. Sets are implemented as hash tables of object references; object references are native machine pointers. Actual memory requirements is typically 4 bytes per entry, plus an additional overall overhead of from 50% to 25%.

Strings are one of the commonest structure keys. Variables are identified by strings (there is no seperate “name” or “variable” type in ICI).

struct - An unordered set of mappings

A struct is an unordered set of mappings. That is, a struct records object references that are regarded as *keys* and for each such key, a corresponding *value*, which is also an object reference.[8] A struct also records a *super* struct, which is a reference to a subsequent struct. The details of structure indexing are described above. See “Structure and set keys” on page 51.

Structures form the fundamental basis for variables and scoping in ICI.

Adding, removing and looking up obejcts in a struct is an efficient constant time operation (although is O(n) with respect to searches up the super chain). The type and complexity of an object being added or removed from a set has no effect on the efficiency of the operation.

Operators

The following table details each of the unary and binary operators with all of the types they may be applied to. Within the first column the standard type names are used to stand for operands of that type, along with *any* to mean any type and *num* to mean an *int* or a *float*. In general, where an *int* and a *float* are combined in an arithmetic operation, the *int* is first converted to a *float* and then the operation is performed.

The following table is in precedence order.

***ptr**	Indirection: The result references the thing the pointer points to. The result is an lvalue.
&*any*	Address of: The result is a pointer to *any*. If *any* is an lvalue the pointer references that storage location. If *any* is not an lvalue but is a *term* other than a bracketed non-*term*, as described in the syntax above, a one element array containing *any* will be fabricated and a pointer to that storage location returned. For example: `p = &1;` sets p to be a pointer to the first element of an un-named array, which currently contains the number 1.
-*num*	Negation: Returns the negation of *num*. The result is the same type as the argument. The result is not an lvalue.
+*any*	Has no effect except the result is not an lvalue.

8. Structs are implemented as hash tables of object references, with each entry recording a value associated with the key. Actual memory requirements is typically 8 bytes per entry, plus an additional overall overhead of from 50% to 25%.

! *any*	Logical negation: If *any* is 0 (integer) or NULL, 1 is returned, else 0 is returned.
~*int*	Bit-wise complement: The bit-wise complement of *int* is returned.
++*any*	Pre-increment: Equivalent to **(***any* **+= 1)**. *any* must be an lvalue and obey the restrictions of the binary **+** operator. See **+** below.
--*any*	Pre-decrement: Equivalent to **(***any* **-= 1)**. *any* must be an lvalue and obey the restrictions of the binary **-** operator. See **-** below.
@*any*	Atomic form of: Returns the unique, read-only form of *any*. If *any* is already atomic, it is returned immediately. Otherwise an atomic form of *any* is found or generated and returned; this is of execution time order equal to the number of elements in *any*. See the section on objects above for more explanation.
$*any*	Immediate evaluation: Recognised by the parser. The sub-expression *any* is immediately evaluated by invocation of the execution engine. The result of the evaluation is substituted directly for this expression term by the parser.
any++	Post-increment: Notes the value of any, then performs the equivalent of **(***any* **+= 1)**, except any is only evaluated once, and finally returns the original noted value. any must be an lvalue and obey the restrictions of the binary + operator. See + below.
any--	Post-increment: Notes the value of any, then performs the equivalent of **(***any* **-= 1)**, except any is only evaluated once, and finally returns the original noted value. any must be an lvalue and obey the restrictions of the binary - operator. See - below.
any1 **@** *any2*	*Form pointer: Returns a pointer object formed from its operands with the pointer's aggregate being set from any1 and the pointer's key from any2.*
num1 ***** *num2*	Multiplication: Returns the product of the two numbers, if both nums are ints, the result is int, else the result is float.
set1 ***** *set2*	Set intersection: Returns a set that contains all elements that appear in both *set1* and *set2*.
num1 / *num2*	Division: Returns the result of dividing *num1* by *num2*. If both numbers are ints the result is int, else the result is float. If *num2* is zero the error *division by 0* is generated, or *division by 0.0* if the result would have been a float.
int1 **%** *int2*	Modulus: Returns the remainder of dividing *int1* by *int2*. If *int2* is zero the error *modulus by 0* is generated.
num1 **+** *num2*	Addition: Returns the sum of *num1* and *num2*. If both numbers are *ints* the result is *int*, else the result is *float*.

ptr + *int*

Pointer addition: *ptr* must point to an element of an indexable object whose index is an *int*. Returns a new pointer which points to an element of the same aggregate which has the index which is the sum of *ptr*'s index and *int*. The arguments may be in any order.

string1 + *string2*

String concatenation: Returns the string which is the concatenation of the characters of *string1* then *string2*. The execution time order is proportional to the total length of the result.

array1 + *array2*

Array concatenation: Returns a new array which is the concatenation of the elements from *array1* then *array2*. The execution time order is proportional to the total length of the result. Note the difference between the following:

```
a += [array 1];
push(a, 1);
```

In the first case a is replaced by a newly formed array which is the original array with one element added. But in the second case the *push* function (see below) appends an element to the array *a* refers to, without making a new array. The second case is much faster, but modifies an existing array.

struct1 + *struct2*

Structure concatenation: Returns a new struct which is a copy of *struct1*, with all the elements of *struct2* assigned into it. Obeys the semantics of copying and assignment discussed in other sections with regard to super structs.. The execution time order is proportional to the sum of the lengths of the two arguments.

set1 + *set2*

Set union: Returns a new set which contains all the elements from both sets. The execution time order is proportional to the sum of the lengths of the two arguments.

num1 - *num2*

Subtraction: Returns the result of subtracting *num2* from *num1*. If both numbers are ints the result is *int*, else the result is *float*.

set1 - *set2*

Set subtraction: Returns a new set which contains all the elements of *set1*, less the elements of *set2*. The execution time order is proportional to the sum of the lengths of the two arguments.

ptr1 - *ptr2*

Pointer subtraction: *ptr1* and *ptr2* must point to elements of indexable objects whose indexs are *ints*. Returns an *int* which is the the index of *ptr1* less the index of *ptr2*.

int1 >> *int2*

Right shift: Returns the result of right shifting *int1* by *int2*. Equivalent to division by 2*******int2*. *int1* is interpreted as a signed quantity.

int1 << *int2*

Left shift: Returns the result of left shifting *int1* by *int2*. Equivalent to multiplication by 2*******int2*.

num1 < *num2*

Numeric test for less than: Returns 1 if *num1* is less than *num2*, else 0.

set1 < *set2* — Test for subset: Returns 1 if *set1* contains only elements that are in *set2*, else 0.

string1 < *string2* — Lexical test for less than: Returns 1 if *string1* is lexically less than *string2*, else 0.

ptr1 < *ptr2* — Pointer test for less than: *ptr1* and *ptr2* must point to elements of indexable objects whose indexes are *ints*. Returns 1 if *ptr1* points to an element with a lesser index than *ptr2*, else 0.

The >, <= and >= operators work in the same fashion as <, above. For sets > tests for one set being a superset of the other. The <= and >= operators test for proper sub- or super-sets. That is one set can contain only those elements contained in the other set but cannot be equal to the other set.

any1 == *any2* — Equality test: Returns 1 if *any1* is equal to *any2*, else 0. Two objects are equal when: they are the same object; or they are both arithmetic (*int* and *float*) and have equivalent numeric values; or they are aggregates of the same type and all the sub-elements are the same objects.

any1 **!=** *any2* — Inequality test: Returns 1 if *any1* is not equal to *any2*, else 0. See above.

string ~ *regexp* — Logical test for regular expression match: Returns 1 if *string* can be matched by *regexp*, else 0. The arguments may be in any order.

string **!~** *regexp* — Logical test for regular expression non-match: Returns 1 if *string* can not be matched by *regexp*, else 0. The arguments may be in any order.

string ~~ *regexp* — Regular expression sub-string extraction: Returns the sub-string of *string* which is matched by the first bracket enclosed portion of *regexp*, or NULL if there is no match or *regexp* does not contain a (...) portion. The arguments may be in any order. For example, a "basename" operation can be performed with:

```
argv[0] ~~= #([^/]*)$#;
```

string ~~~ *regexp* — Regular expression multiple sub-string extraction: Returns an array of the the sub-strings of *string* which are matched by the (...) enclosed portions of *regexp*, or NULL if there is no match. The arguments may be in any order.

int1 **&** *int2* — Bit-wise and: Returns the bit-wise and of *int1* and *int2*.

int1 ^ *int2* — Bit-exclusive or: Returns the bit-wise exclusive or of *int1* and *int2*.

int1 | *int2* — Bit-wise or: Returns the bit-wise or of *int1* and *int2*.

any1 **&&** *any2*

Logical and: Evaluates the expression which determines *any1*, if this evaluates to 0 or NULL (i.e. *false*), 0 is returned, else *any2* is evaluated and returned[9]. Note that if *any1* does not evaluate to a *true* value, the expression which determines *any2* is never evaluated.

any1 || *any2*

Logical or: Evaluates the expression which determines *any1*, if this evaluates to other than 0 or NULL (i.e. *true*), 1 is returned, else *any2* is evaluated and returned. Note that if *any1* does not evaluate to a *false* value, the expression which determines *any2* is never evaluated.

any1 **?** *any2* **:** *any3*

Choice: If *any1* is neither 0 or NULL (i.e. *true*), the expression which determines *any2* is evaluated and returned, else the expression which determines *any3* is evaluated and returned. Only one of *any2* and *any3* are evaluated. The result may be an lvalue if the returned expression is. Thus:

```
flag ? a : b = value
```

is a legal expression and will assign *value* to either *a* or *b* depending on the state of *flag*.

any1 = *any2*

Assignment: Assigns *any2* to *any1*. *any1* must be an lvalue. The behavior of assignment is a consequence of aggregate access as discussed in earlier sections. In short, an lvalue (in this case *any1*) can always be resolved into an aggregate and an index into the aggregate. Assignment sets the element of the aggregate identified by the index to *any2*. The returned result of the whole assignment is *any1*, after the assignment has been performed.

The result is an lvalue, thus:

```
++(a = b)
```

will assign *b* to *a* and then increment *a* by 1.

Note that assignment operators (this and following ones) associate right to left, unlike all other binary operators, thus:

```
a = b += c -= d
```

Will subtract *d* from *c*, then add the result to *b*, then assign the final value to *a*.

+= -= *= /= %= >>= <<= &= ^= |= ~~=

Compound assignments: All these operators are defined by the rewriting rule:

9. Note that this is different from C where the result is always completely resolved to a 0 or 1. Use !! to force a 0/1 value from a generic true/false.

any1 op= any2

is equivalent to:

any1 = any1 op any2

except that *any1* is not evaluated twice. Type restrictions and the behavior or *op* will follow the rules given with that binary operator above. The result will be an lvalue (as a consequence of = above). There are no further restrictions. Thus:

```
a = "Hello";
a += " world.\n";
```

will result in the variable *a* referring to the string:

```
"Hello world.\n".
```

any1 <=> *any2*

Swap: Swaps the current values of *any1* and *any2*. Both operands must be lvalues. The result is *any1* after the swap and is an lvalue, as in other assignment operators. Also like other assignment operators, associativity is right to left, thus:

```
a <=> b <=> c <=> d
```

rotates the values of *a*, *b* and *c* towards *d* and brings *d*'s original value back to *a*.

any1 **,** *any2*

Sequential evaluation: Evaluates *any1*, then *any2*. The result is *any2* and is an lvalue if *any2* is. Note that in situations where comma has meaning at the top level of parsing an expression (such as in function call arguments), expression parsing precedence starts at one level below the comma, and a comma will not be recognised as an operator. Surround the expression with brackets to avoid this if necessary.

Automatic library loading

During execution, should the ICI execution engine fail to find a variable it is attempting to read within the current scope, it will attempt to load a library based on the name of that variable. Such a library may be a host specific dynamically loaded native machine code library, an ICI module, or both.

In attempting to load an ICI module, a file name of the form:

ici*var*`.ici`

is considered, where *var* is the as yet undefined variable name. This file is searched for on the current host specific search path. If found, a new extern, static and auto scope is established and the new extern scope struct is assigned to *var* in the outermost writable scope available. That outermost writable scope also forms the super of the new extern scope. The module is then parsed with the given scope, after which the variable lookup is repeated. In normal practice this

will mean that the loaded module has an outer scope holding all the normal ICI primitives and a new empty extern scope. The intent of this mechanism is that the loaded module should define all its published functions in its extern scope. References by an invoking program to functions and other objects of the loaded module would always be made explicitly through the *var* which references the module. For example, a program might contain the fragment:

```
query = cgi.decode_query();
cgi.start_page("Query results");
```

where "cgi" is undefined, but the file *ici4cgi.ici* exists on the search path and includes function definitions such as:

```
extern
decode_query()
{
    ...
}

extern
start_page(title)
{
    ...
}
```

Upon first encountering the variable *cgi* in the code fragment the module *ici4cgi.ici* will be parsed and its extern scope assigned to the new variable *cgi* in the outermost scope of the program (that is, the most global scope). The lookup of the variable *cgi* is then repeated, this time finding the structure which contains the function *decode_query*. The second, and all subsequent, use of the variable *cgi* will be satisfied immediately from the already loaded module.

In attempting to load a host specific dynamically loaded native machine code library, a file name of the form:

ici4*var.ext*

is considered, where *var* is the as yet undefined variable name and *ext* is the normal host extension for such libraries (typically *.dll* for Windows and *.so* for UNIX like systems). The *4* is the major ICI version number. This file is searched for on the current host specific search path. If found the file is loaded into the ICI interpreter's address space using the local host's dynamic library loading mechanism. An initialisation function in the loaded library may return an ICI object (see below). Should an object be returned, it is assigned to *var* in the outermost writable scope available. Further, should the returned variable be a structure, additional loading of an ICI module of the same name is allowed (as described above) and the returned struct forms the structure for externs in that load.

CHAPTER 5

Object-oriented programming in ICI

In object-oriented ICI programs, “objects” are structs that have specific properties. This is a bit confusing because I have been using the term “object” to refer to any ICI primitive type. This is historical. To avoid further confusion I will use “class” and “instance” explicitly instead of “object” when talking about object-oriented techniques.

ICI supports object-oriented programming by building on the properties of structs to implement scoping in the same way that vanilla function calls do. The principal feature that supports object-oriented programming in ICI is calls to *methods* as opposed to calls to *functions*. Contrasting the two:

- a call to a function causes an implicit switch to the scope of the function for the duration of the call, whereas
- a call to a method causes an implicit switch to the scope of the instance and its class for the duration of the call.

A method is a primitive ICI object that is a pairing of a subject object (the instance), and a function.

Consider the following simple fragment which creates a class:

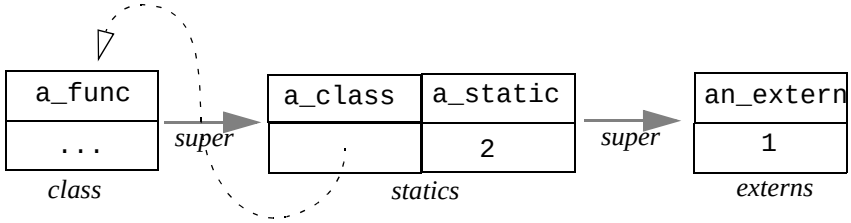
```
extern an_extern = 1;

static a_static = 2;

static a_class = [class

    a_func(arg)
    {
        this.value := arg + 1;
        return value + 2;
    }
];
```

After executing this code, *a_class* will refer to a new struct which is unremarkable except that it's super has been automatically set to be the static scope. Diagramatically:

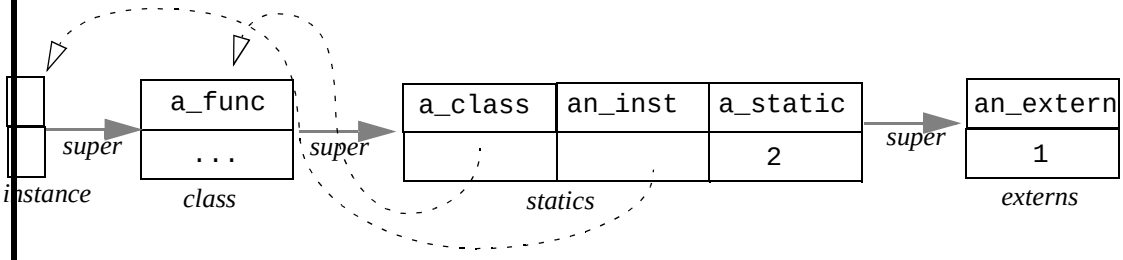


We can create an instance of the class by invoking the *new* method on the class. For example:

```
an_inst = a_class:new();
```

The *new* method is a class method that exists in the global scope, so all classes effectively inherit it from there.

The new instance is, again, a struct that it unremarkable except that it's super has been set to the class. In this simple example there are, as yet, no instance variables. So the instance is an empty struct. Diagramatically:

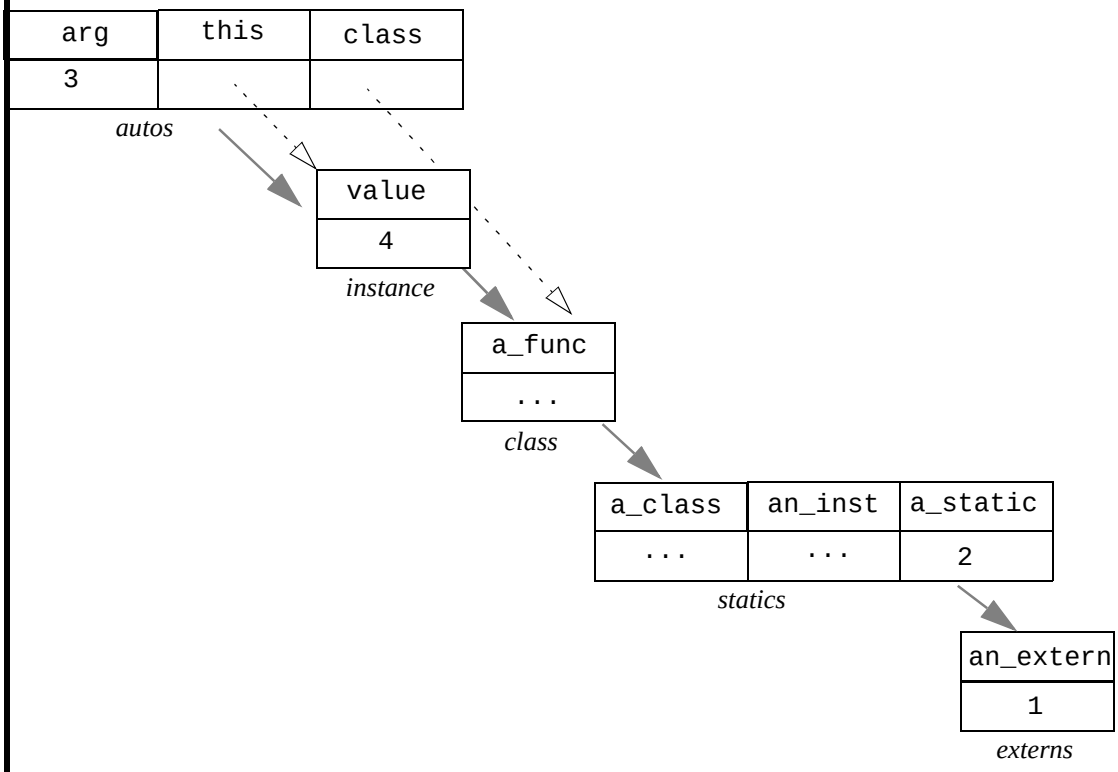


We are now in a position to invoke the *a_func* method on our new instance with, say:

```
x = an_inst:a_func(3);
```

The transfer of control into the function creates a struct for auto variables as usual, but rather than making the super of this struct the static scope the function was defined in, it is set to the

instance that is the subject of this method call. Also, the local variables *this* and *class* are set automatically. Diagramatically, just after the first line of code in the function is executed:



After execution, *x* will be 6. Notice the use of the `:=` operator and the explicit use of *this* to force the creation of *value* in the instance. Otherwise it would have implicitly appeard as a local variable. This is, of course, only required when the instance variable doesn't already exist.

The instance is a normal struct. Thus we can reference the *value* instance variable with:

```
an_inst.value
```

Note that the instance has the class and outer scopes in its super chain. Thus we can also refer to:

```
an_inst.a_func
an_inst.a_static
an_inst.an_extern
```

Sub-classes

Sub-classes are class structs that have another class as their super. The following example illustrates a number of aspects of sub-classing:

```
static sub_class = [class:a_class

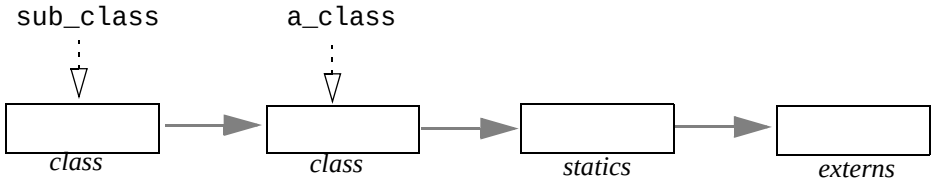
    a_class_variable = 0,

    new(name)
    {
        o = this:^new();
```

```
            o.name := name;
            o.a_count := 0;
            return o;
        }

        a_func()
        {
            this:^a_func();
            ++a_count;
        }
    ];
```

After parsing we have a variable *sub_class* whoes super is *a_class*. Diagramatically:



To make a new instance of the sub-class we would execute:

```
subclass_inst = sub_class:new("a name");
```

The *new* function was defined in the sub-class, overidding the global *new* function. In this case *new* is class function that expects to be called on the class itself, not an instance of the class. There is nothing that distinguishes class functtions from ones that operates on an instance except their operation and documentation.

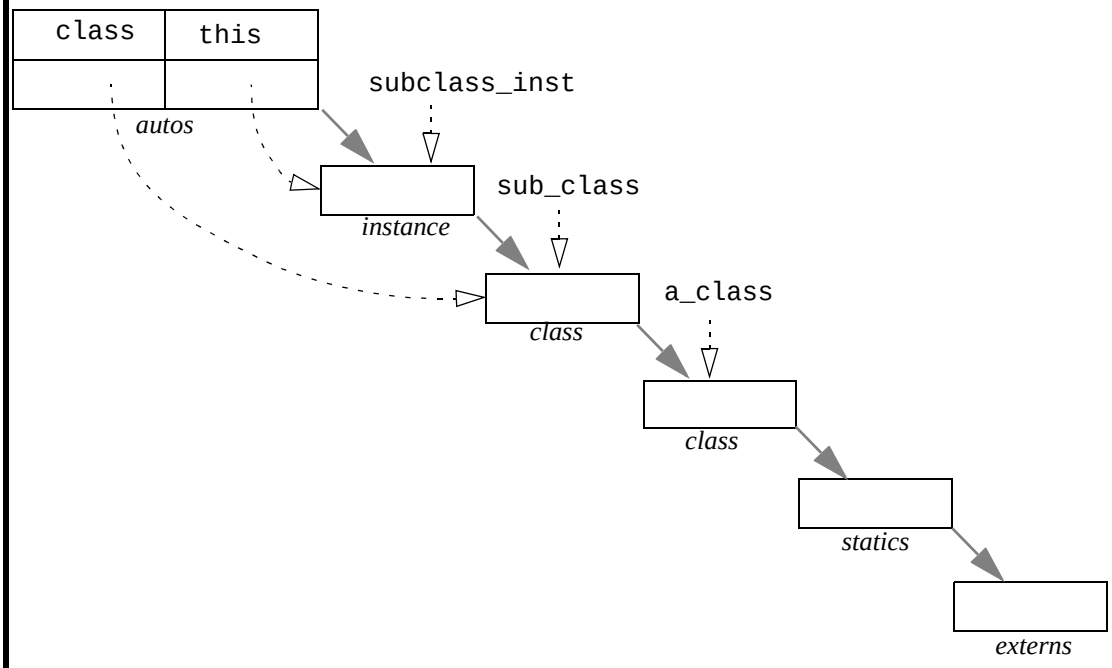
To complete its operation, the *new* function coded here needs to call the *new* of the super-class. To do this it uses the `:^` operator which forms a method, but using the super of the current value of the *class* variable. There isn't actually a *new* coded in the super-class, but it will find the global *new.*

To work with sub-classes and overidden function it is important to understand how the *this* and *class* variables are set in method calls.

Consider the call:

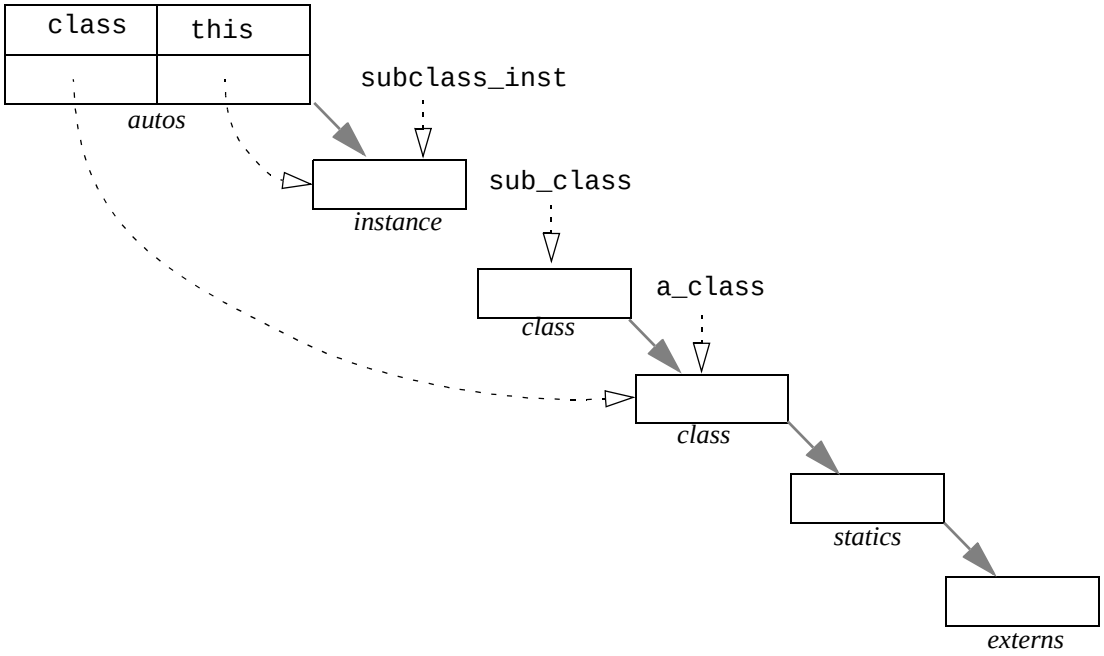
```
subclass_inst:a_func();
```

Before the first line of code is executed, the scope will look like this:



The *class* variable has been set by the method call mechanism to the class of the function being called. Functions being parsed within the scope of a class definition record their class, so it was not the super of the instance that set the class variable, but the class recorded by the function.

The first thing at the sub-class *a_func* function does is call the same function in its super-class. Upon arrival in that function, the scope will look like:



In short, the *class* variable is always the class of the function, irrespective of any sub-classing the instance may be derived from (or any funny business done by changing the super of the instance).

Finally, note that class variables can simply be included in the class definition (as shown by *a_class_variable* in the example). They exist in the class and have no effect on any instance.

A shorthand for creating instance variables

The global *new* function supports a feature that allows creation and simple initialisation of instance variables without explicitly coding a *new* function. The global *new* function scans all of the structures in the the scope chain of the instance it is creating in order from outer-most to the instance's super. For each variable called *proto* it finds whoes values is a struct, it uses entries in the struct to create initial entries in the new instance structure. For example:

```
node = [class

    proto = [struct parent, nchild = 0],
    ...
];

binary_node = [class:node

    proto = [struct left, right, nchild = 2],
    ...
];

n = binary_node:new();
```

After execution, *n* will refer to a struct with *parent*, *left*, and *right* set to NULL, and *nchild* set to 2. Notice that the value for *nchild* in the sub-class overrode the value given in the parent class.

Global methods

As has been seen, the static scope present when a class is defined forms the super for the class. In effect, the outer scopes can be considered outer classes. Functions defined in those scopes may, if appropriately coded, be class functions for these hypothetical top-level classes. For example, we could define a default debug method that we expect some classes to override:

```
extern
dump()
{
    forall (k, v in this)
        printf("%s=%s, ", string(k), string(v));
    printf("\n");
}
```

This function would be available to all instances of all classes. The class of such a function is the scope it was defined in.

Taking advantage of dynamic binding

All name binding is dynamic in ICI. This leads to a number of common constructs that are worthy of highlighting, because they are not seen in statically bound languages such as C++.

The commonest of these is polymorphic functions that work equally well with any object instance that falls within the scope of their definition, irrespective of its class. We saw a simple example of this above with the *dump* function. That function had no prerequisites on the object it was applied to. But in real applications it is more common to define functions that state they will do *blah*, providing the instance they are applied to has fields called *whatever*, that can be interpreted in *such-and-such* a way. For example:

```
/*
 * Return the distance across the diagonal of the
 * boundig box for any object that support a bounding
 * box recorded as xmin, xmax, ymin, ymax.
 */
extern
bbox_diagonal()
{
    dx = xmax - xmin;
    dy = ymax - ymin;
    return sqrt(dx * dx + dy * dy);
}

/*
 * Grow the bounding box of the object to ensure it
 * will account for a r radian rotation of any object
 * contained within the original bounding box. The
 * bounding box is assumed to be recorded in xmin,
 * xmax, ymin, ymax.
 */
extern
bbox_grow_for_rotation(r)
{
    ...
}
```

ICI does not support multiple inheritance as such. But it is common and useful to use composite classes and/or global methods that provide the same effect.

There are various was of arranging this, but all are based on the principle that ### to be continued.

Standard global methods

The standard global methods available to all ICI instances or classes are summarised below. See the chapter on core language functions for detailed descriptions of each:

inst`:isa(`*class*`)`	Returns 1 if *inst* is or is derived from *class*, else 0. May be applied to an instance or a class.
class`:new()`	Returns a new instance of *class*.

inst`:respondsto(`*name*`)`

Returns 1 if *inst* supports a function called *name*, else 0. May be applied to an instance or a class.

Multiple inheritance

ICI's object-oriented features do not provide multiple inheritance as such. But it is somtimes useful to build a class from more than one building block class. This can be achieved fairly simply, providing it is not done blindly. In particular, you must be aware of potential name clashes and of the structure of each class. By way of example, consider a program that has drawable objects like splines, text and images. Each of these is a sub-class of a general drawable class. In this same program we have a matrix class that is used in various ways in the program. Now, we wish the image sub-class (but not the others) to also behave as a matrix. A simple (but not pure) way of achieving this is simply to add all the functions in the matrix class to the image class. When matrix methods are applied to images, the class the functions see will be their original class, but the instance and it's class heirachy are what further operations are applied to. ### to be continued.

CHAPTER 6

Core language functions

Core function summary

The following list summarises the standard functions. Following this is a detailed descriptions of each of them.

```
   float|int = abs(float|int)
       float = acos(number)
         mem = alloc(int [, int])
       array = array(any...)
       float = asin(number)
         any = assign(struct, any, any)
       float = atan(number)
       float = atan2(number, number)
array|struct = build(dims... [, options, content...])
float|struct = calendar(struct|float)
         any = call(func [, arg...], args)
       float = ceil(number)
               chdir(string)
               close(file)
         int = cmp(a, b)
         any = copy(any)
       float = cos(number)
       float = cputime([foat])
        file = currentfile([string])
         int = debug([int])
             = del(struct, any)
       array = dir([path], [, regexp] [, format])
         int = eof(file)
         int = eq(any, any)
```

```
                    = eventloop()
                    = exit([int|string|NULL])
              float = exp(number)
              array = explode(string)
                      fail(string)
                any = fetch(struct, any)
              float = float(any)
              float = floor(number)
                int = flush([file])
              float = fmod(number, number)
               file = fopen(string [, string])
             string = getchar([file])
             string = getcwd()
             string = getenv(string)
             string = getfile([file])
             string = getline([file])
             string = gettoken([file|string [,string]])
              array = gettokens([file|string [,string [,string]]])
             string = gsub(string, regexp, string)
             string = implode(array)
             struct = include(string [, struct])
                int = int(any)
      string|array = interval(string|array, int [, int])
                int = inst|class:isa()
                int = isatom(any)
              array = keys(struct)
                any = load(string)
              float = log(number)
              float = log10(number)
                mem = mem(int, int [,int])
               file = mopen(string [, string])
                int = nels(any)
               inst = class:new(...)
              float = now()
         int|float = num(string|int|float)
             struct = parse(file|string [, struct])
                any = pop(array)
               file = popen(string [, string])
              float = pow(number, number)
                      printf([file,] string [, any...])
                      profile(filename)
                any = push(array, any)
                      put(string [, file])
                      putenv(string [, string])
                int = rand([int])
```

```
                  reclaim()
          regexp = regexp(string)
          regexp = regexpi(string)
                  remove(string)
                  rename(string, string)
             int = inst|class:respondsto(string)
             any = rpop(array)
                  rpush(array, any)
          struct = scope([struct])
             int = seek(file, int, int)
             set = set(any...)
   string|func = signal(int|string [, func|string])
          string = signam(int)
           float = sin(number)
                  sleep(number)
           array = smash(string [, regexp [, string...] [, int]]);
            file = sopen(string [, string])
                  sort(array, func [, arg])
          string = sprintf(string [, any...])
           float = sqrt(number)
          string = string(any)
          struct = struct(any, any...)
          string = sub(string, regexp, string)
          struct = super(struct [, struct])
             int = system(string)
           float = tan(number)
            exec = thread(callable [, args...])
          string = tochar(int)
             int = toint(string)
             any = top(array [, int])
             int = trace(string)
          string = typeof(any)
          string = version()
           array = vstack()
file|int|float = waitfor(file|int|float...)
                  wakeup(any)
```

Core language functions

`float|int = abs(float|int)`

Returns the absolute value of its argument. The result is an int if the argument is an int, a float if it is a float.

angle = acos(x)

Returns the arc cosine of *x* in the range 0 to pi.

mem = alloc(nwords [, wordz])

Returns a new *mem* object referring to *nwords* (an int) of newly allocated and cleared memory. Each word is either 1, 2, or 4 bytes as specified by *wordz* (an int, default 1). Indexing of *mem* objects performs the obvious operations, and thus pointers work too.

array = array(any...)

Returns an array formed from all the arguments. For example:

```
array()
```

will return a new empty array; and

```
array(1, 2, "a string")
```

will return a new array with three elements, *1*, *2*, and *"the string"*.

This is the run-time equivalent of the array literal. Thus the following two expressions are equivalent:

```
$array(1, 2, "a string")
```

```
[array 1, 2, "a string"]
```

float = asin(x)

Returns the arc sine of *x* in the range -pi/2 to pi/2.

value = assign(struct, key, value)

Sets the element of *struct* identified by *key* to *value*, ignoring any super struct. Returns *value*.

angle = atan(x)

Returns the arc tangent of *x* in the range -pi/2 to pi/2.

angle = atan2(y, x)

Returns the angle from the origin to the rectangular coordinates *x*, *y* (floats) in the range -pi to pi.

array|struct = build(dims... [, options, content...])

Build allows construction of a regular data structure such as a multi-dimensional array or an array of structures. *dims...* is a sequence of dimension specifications. For example:

```
build(20, 10);
```

returns a 20×10 array of NULLs (that is, an array of 20 arrays, each of size 10).

Each dimension specification is eiher:

an int	causing an *array* of that many elements to be made and have every element set through recursive application on subsequent dimensions, or
an array	causing a *struct* with the elements of the array as keys to be made and each value set through recursive application on subsequent dimensions.

So, for example:

```
build(10, [array "x", "y"], 2)
```

Returns an array of ten structures, each with fields *x* and *y*. Each field is set to an array of length 2.

If *options* and *content...* are supplied, they may be used to supply initialising data to the leaf fields of the data structure rather than the default NULL. *Options* is a string, which may be:

"c"	Cyclical. The content is used and assigned cyclicly to leaf items.
"r"	Restart. The content is used and assigned cyclicly, but the content list is also restarted from the first item on the commencement of each bottom level aggragate.
"l"	Last repeates. The content is used and assigned in sequence to leaf items, but once it is exhausted, the last content item is used repeatedly for subsequent leaf items.
"a"	Arrays. Each of the content items must be an array. Content is taken firstly from the first element of each array in turn, then from the second element of each in turn etc. If any array is too short, NULL is used as the value.
"i"	Integer increment. The content is incrementing integer values. The first content value, is given is the start value, default 0. The second content value, if given, is the step, default 1.

So, for example, supposing *names* is an array of names of some sort:

```
build(names, [array "count", "sum"], "c", 0, 0.0)
```

will return a struct which, when indexed by a name reveals a struct with fields *count* and *sum* initialised to 0 and 0.0 respectively.

Also:

```
build(100, "i")
```

will return an array filled with the integers from 0 to 99.

Finally, if *names* is an array of names of some sort and *values* is a corresponding array of values:

```
build(nels(names), [array "name", "value"], "a", names, values)
```

will transponse them into an array of structs, each with a *name* and *value* field.

float|struct = calendar(struct|float)

Converts between calendar time and arithmetic time. An arithmetic time is expressed as a signed float time in seconds since 0:00, 1st Jan 2000 UTC. The calendar time is expressed as a structure with fields revealing the local (including current daylight saving adjustment) calendar date and time. Fields in the calendar structure are:

second	The float number of seconds after the minute.
minute	The int number of minutes after the hour.
hour	The int number of hours since midnight.
day	The day of the month (1..31).
month	The int month number, Jan is 0.
year	The int year.
wday	The day since Sunday (0..6)
yday	Days since 1st Jan.

When converting from a local calendar time to an arithmetic time, the fields *sec*, *min*, *hour*, *mday*, *mon*, *year* are used. They need not be restricted to their nomal ranges.

return = call(func [, arg...], args)

Calls the function *func* with the given arguments (*arg...)* and arguments taken from the array *args*. If *args* is NULL it is ignored. Else it must be an array. Returns the return value of the function.

This is often used to pass on an unknown argument list. For example:

```
static
db()
{
     auto vargs;

     if (debug)
          return call(printf, stderr, vargs);
}
```

chdir(path)

Change the current working directory to the specified path.

`float = ceil(x)`

Returns $\lceil x \rceil$ (the smallest integral value greater than or equal to *x*) as a float, where *x* is a number (int or float).

`close(file)`

Close the given file, releasing low level system resources. After this operation the file object is still a valid object, but I/O operations on it will fail. Files object that are lost and collected by the garbage collector will be closed. But due to the indeterminate timming of this, it is preferable to close them explicitly.

On some files and systems this may block, but will allow thread switching while blocked.

`close(file)`

Close the given file, releasing low level system resources. After this operation the file object is still a valid object, but I/O operations on it will fail. Files object that are lost and collected by the garbage collector will be closed. But due to the indeterminate timming of this, it is preferable to close them explicitly.

On some files and systems this may block, but will allow thread switching while blocked.

`int = cmp(a, b)`

Returns -1, 0 or 1 depending if $a < b$, $a == b$, or $a > b$. The operands may be any type for which the < and > operators are defined. This is the default comparison function for *sort()*.

`x = cos(angle)`

Returns the cosine of *angle* (a float interpreted in radians).

`float = cputime([float])`

Returns the accumulated CPU time of the current process in seconds. The precision and accuracy is system dependent.

If *float* is supplied it specifies a new origin, relative to the value being returned, from which subsequent calls are measured. Mostly commnly the value 0.0 is used here.

`file = currentfile(["raw"])`

Returns a file associated with the innermost parsing context, or NULL if there is no module being parsed. By default *currentfile()* returns a new file object that gives “cooked” access that layers on top of the parser’s access to the file. This maintains line number tracking and normalises differing newline conventions to single newline characters even for binary files. If the string `"raw"` is given as an argument, the underlying file that is being parsed is returned directly, bypassing such operations.

This function can be used to include data in a program source file which is out-of-band with respect to the normal parse stream. But to do this it is necessary to know up to what character in the file in question the parser has consumed.

In general: after having parsed any simple statement the parser will have consumed up to and including the terminating semicolon, and no more. Also, after having parsed a compound statement the parser will have consumed up to and including the terminating close brace and no more. For example:

```
static help = gettokens(currentfile(), "", "!")[0]

;This is the text of the help message.
It follows exactly after the ; because
that is exactly up to where the parser
will have consumed. We are using the
gettokens() function (as described below)
to read the text.
!

static otherVariable = "etc...";
```

In the examples shown above, the default cooked mode is used so that line numbers are tracked and stay in sync for subsequence diagnostics. If the raw mode was used the parser would never see the data read out-of-band and would not realise how many lines have been skipped, thus giving inaccurate reports of line numbers on errors later in the file.

This function can also be used to parse the rest of a file within an error catcher. For example:

```
try
     parse(currentfile("raw"), scope())
onerror
     printf("That didn't work, but never mind.\n");

static this = that;
etc();
```

The functions *parse* and *scope* are described below.

int = debug([int])

Returns the current debug status, and, if an int is supplied as an argument, set it to that value.

###The debug mechanism requires more documentation.

del(aggr, key)

Deletes an element of *aggr*, which must be a struct, a set or an array, as identified by *key.* Any super structs are ignored. For structs and sets this is an efficient operation. For arrays it is O(n) where *n* is the length of the array. Returns NULL.

For example:

```
static s = [struct a = 1, b = 2, c = 3];
static v, k;
```

```
forall (v, k in s)
    printf("%s=%d\n", k, v);
del(s, "b");
printf("\n");
forall (v, k in s)
    printf("%s=%d\n", k, v);
```

When run would produce (possibly in some other order):

```
a=1
c=3
b=2

a=1
c=3
```

`array = dir([path,] [regexp,] [format])`

Read directory named in path (a string, defaulting to ".", the current working directory) and return the entries that match the pattern as an array of strings (or all names if no pattern passed). The format string identifies what sort of entries should be returned. If the format string is passed then a path MUST be passed (to avoid any ambiguity) but path may be NULL meaning the current working directory (same as "."). The format string uses the following characters,

f	Return file names.
d	Return directory names.
a	Return all names (which includes things other than files and directories, e.g., hidden or special files).

The default format specifier is "f".

Note that when using *dir()* to traverse directory hierarchies that the "." and ".." names are returned when listing the names of sub-directories, these will need to be avoided when traversing.

`int = eq(obj1, obj2)`

Returns 1 (one) if *obj1* and *obj2* are the same object, else 0 (zero).

`int = eof([file])`

Returns non-zero if end of file has been read on *file*. If *file* is not given the current value of *stdin* in the current scope is used.

`eventloop()`

Enters an internal event loop and never returns. The exact nature of the event loop is system specific. Some dynamically loaded modules require an event loop for their operation. Allows thread switching while blocked.

exit([string|int|NULL])

Causes the interpreter to finish execution and exit. If no parameter, the empty string or NULL is passed the exit status is zero. If an integer is passed that is the exit status. If a non-empty string is passed then that string is printed to the interpreter's standard error output and an exit status of one used. This is implementation dependent and may be replaced by a more general exception mechanism. Avoid.

float = exp(x)

Returns the exponential function of *x*.

array = explode(string)

Returns an array containing each of the integer character codes of the characters in *string*.

fail(string)

Causes an error to be raised with the message *string* associated with it. See the section on error handling in the *try* statement above. For example:

```
if (qf > 255)
     fail(sprintf("Q factor %d is too large", qf));
```

value = fetch(struct, key)

Returns the *value* from *struct* associated with *key*, ignoring any super structs. Returns NULL is *key* is not an element of *struct*.

value = float(x)

Returns a floating point interpretation of *x*, or 0.0 if no reasonable interpretation exists. *x* should be an int, a float, or a string, else 0.0 will be returned.

float = floor(x)

Returns $\lfloor x \rfloor$ (the largest integral value less than or equal to *x*) as a float, where *x* is a number (int or float).

flush([file])

Flush causes data that has been written to the file (or stdout if absent), but not yet delivered to the low level host environment, to be deliverd immediately.

On some files and systems this may block, but will allow thread switching while blocked.

`float = fmod(x, y)`

Returns the float remainder of x/y where *x* and *y* are numbers (int or float). That is, $x - i \times y$ for some integer *i* such that the result has the same sign as *x* and magnitude less than *y*.

`file = fopen(name [, mode])`

Opens the named file for reading or writing according to *mode* and returns a file object that may be used to perform I/O on the file. *Mode* is the same as in C and is passed directly to the C library **`fopen`** function. If mode is not specified **`"r"`** is assumed.

On some files and systems this may block, but will allow thread switching while blocked.

`string = getchar([file])`

Reads a single character from *file* and returns it as a string. Returns NULL upon end of file. If *file* is not given, the current value of *stdin* in the current scope is used.

On some files and systems this may block, but will allow thread switching while blocked.

`string = getcwd()`

Returns the name of the current working directory.

`string = getfile([file])`

Reads all remaining data from *file* and returns it as a string. Returns an empty string upon end of file. If *file* is not given, the current value of *stdin* in the current scope is used.

On some files and systems this may block, but will allow thread switching while blocked.

`string = getline([file])`

Reads a line of text from *file* and returns it as a string. Any end-of-line marker is removed. Returns *NULL* upon end of file. If *file* is not given, the current value of *stdin* in the current scope is used.

On some files and systems this may block, but will allow thread switching while blocked.

`string = gettoken([file [, seps]])`

Read a token (that is, a string) from *file*.

Seps must be a string. It is interpreted as a set of characters which do not from part of the token. Any leading sequence of these characters is first skipped. Then a sequence of characters not in seps is gathered until end of file or a character from seps is found. This terminating character is not consumed. The gathered string is returned, or NULL if end of file was encountered before any token was gathered.

If *file* is not given the current value of *stdin* in the current scope is used.

If *seps* is not given the string " \t\n" is assumed.

Currently, even if blocked while reading a file *gettoken* is indivisible with repect to other threads. This may be corrected in future versions.

`array = gettokens([file [, seps [, terms]]])`

Read tokens (that is, strings) from *file*. The tokens are character sequences separated by *seps* and terminated by *terms*. Returns an array of strings, NULL on end of file.

If *seps* is a string, it is interpreted as a set of characters, any sequence of which will separate one token from the next. In this case leading and trailing separators in the input stream are discarded.

If *seps* is an integer it is interpreted as a character code. Tokens are taken to be sequences of characters separated by exactly one of that character.

Terms must be a string. It is interpreted as a set of characters, any one of which will terminate the gathering of tokens. The character which terminated the gathering will be consumed.

If *file* is not given the current value of *stdin* in the current scope will be used.

If *seps* is not given the string " \t" is assumed.

If *terms* is not given the string "\n" is assumed.

For example:

```
forall (token in gettokens(currentfile()))
    printf("<%s>", token)
;   This    is my line    of data.
printf("\n");
```

when run will print:

```
<This><is><my><line><of><data.>
```

Whereas:

```
forall (token in gettokens(currentfile(), ':', "*"))
    printf("<%s>", token)
;:abc::def:ghi:*
printf("\n");
```

when run will print:

```
<><abc><><def><ghi><>
```

Currently, even if blocked while reading a file *gettokens* is indivisible with repect to other threads. This may be corrected in future versions.

`string = gsub(string, string|regexp, string)`

gsub performs text substitution using regular expressions. It takes the first parameter, matches it against the second parameter and then replaces the matched portion of the string with the third parameter. If the second parameter is a string it is converted to a regular expression as if the regexp function had been called. Gsub does the replacement multiple times to replace all occurrances of the pattern. It returns the new string formed by the replacement. If there is no match this is original string. The replacement string may contain the special sequence "\&" which is replaced by the string that matched the regular expression. Parenthesized portions of the regular expression may be matched by using *n* where *n* is a decimal digit.

`string = implode(array)`

Returns a *string* formed from the concatenation of elements of *array.* Integers in the *array* will be interpreted as character codes; strings in the array will be included in the concatenation directly. Other types are ignored.

`struct = include(string [, scope])`

Parses the code contained in the file named by the string into the scope. If scope is not passed the current scope is used. Include always returns the scope into which the code was parsed. The file is opened by calling the current definition of the ICI fopen() function so path searching can be implemented by overriding that function.

`value = int(any)`

Returns an integer interpretation of *x*, or 0 if no reasonable interpretation exists. *x* should be an int, a float, or a string, else 0 will be returned.

`subpart = interval(str_or_array, start [, length])`

Returns a sub-interval of *str_or_array*, which may be either a string or an array.

If *start* (an integer) is positive the sub-interval starts at that offset (offset 0 is the first element). If *start* is negative the sub-interval starts that many elements from the end of the string (offset -1 is the last element, -2 the second last etc).

If *length* is absent, all the elements from the *start* are included in the interval. Otherwise that many elements are included (or till the end, whichever is smaller).

For example, the last character in a string can be accessed with:

```
last = interval(str, -1);
```

And the first three elements of an array with:

```
first3 = interval(ary, 0, 3);
```

int = inst|class:isa(any)

Returns 1 if *inst* or *class* or any of their super classes is equal to *any*, else 0. That is, if *inst* or *class* is a, or is a sub-class of, *any*.

int = isatom(any)

Return 1 (one) if *any* is an atomic (read-only) object, else 0 (zero). Note that integers, floats and strings are always atomic.

array = keys(struct)

Returns an array of all the keys from *struct*. The order is not predictable, but is repeatable if no elements are added or deleted from the struct between calls and is the same order as taken by a *forall* loop.

any = load(string)

Attempt to load a library named by *string*. This is the explicit form of the automatic library loading described in “Automatic library loading” on page 66. The library is loaded in the same way and the resulting object returned. (Actually, this is the real core mechanism. The automatic mechanis calls the function *load()* in the current scope to load the module. Thus overidding *load()* allows control to be gained over the automatic mechnism.)

float = log(x)

Returns the natural logarithm of *x* (a float).

float = log10(x)

Returns the log base 10 of *x* (a float).

mem = mem(start, nwords [, wordz])

Returns a memory object which refers to a particular area of memory in the ICI interpreter's address space. Note that this is a highly dangerous operation. Many implementations will not include this function or restrict its use. It is designed for diagnostics, embedded systems and controllers. See the *alloc* function above.

file = mopen(mem [, mode])

Returns a *file*, which when read will fetch successive bytes from the given *memory object*. The memory object must have an access size of one (see *alloc* and *mem* above). The file is read-only and the *mode*, if passed, must be one of **"r"** or **"rb"**.

`int = nels(any)`

Returns the number of elements in *any.* The exact meaning depends on the type of *any.* If *any* is an:

array	the length of the array is returned; if it is a
struct	the number of key/value pairs is returned; if it is a
set	the number of elements is returned; if it is a
string	the number of characters is returned; and if it is a
mem	the number of words (either 1, 2 or 4 byte quantities) is returned;

and if it is anything else, one is returned.

`inst = class:new()`

Creates a new instance of the given class. In practice *new* is often also defined in sub-classes. This is the global new. *New* supports a mechanism for defining instance variables that alleviates simple classes from defining a *new* method. See “A shorthand for creating instance variables” on page 74.

The new *inst* will be a new struct with *class* as its super.

`float = now()`

Returns the current time expressed as a signed float time in seconds since 0:00, 1st Jan 2000 UTC.

`number = num(x)`

If *x* is an int or float, it is returned directly. If *x* is a string it will be converted to an int or float depending on its appearance; applying octal and hex interpretations according to the normal ICI source parsing conventions. (That is, if it starts with a 0x it will be interpreted as a hex number, else if it starts with a 0 it will be interpreted as an octal number, else it will be interpreted as a decimal number.)

If *x* can not be interpreted as a number the error *%s is not a number* is generated.

`scope = parse(source [, scope])`

Parses *source* in a new variable scope, or, if *scope* (a struct) is supplied, in that scope. *Source* may either be a file or a string, and in either case it is the source of text for the parse. If the parse is successful, the variables scope structure of the sub-module is returned. If an explicit scope was supplied this will be that structure.

If *scope* is not supplied a new struct is created for the auto variables. This structure in turn is given a new structure as its super struct for the static variables. Finally, this structure's super is

set to the current static variables. Thus the static variables of the current module form the externs of the sub-module.

If *scope* is supplied it is used directly as the scope for the sub-module. Thus the base structure will be the struct for autos, its super will be the struct for statics etc.

For example:

```
static x = 123;
parse("static x = 456;", scope());
printf("x = %d\n", x);
```

When run will print:

```
x = 456
```

Whereas:

```
static x = 123;
parse("static x = 456;");
printf("x = %d\n", x);
```

When run will print:

```
x = 123
```

Note that while the following will work:

```
parse(fopen("my-module.ici"));
```

It is preferable in a large program to use:

```
parse(file = fopen("my-module.ici"));
close(file);
```

In the first case the file will eventually be closed by garbage collection, but exactly when this will happen is unpredictable. The underlying system may only allow a limited number of simultaneous open files. Thus if the program continues to open files in this fashion a system limit may be reached before the unused files are garbage collected.

`any = pop(array)`

Returns the last element of *array* and reduces the length of *array* by one. If the array was empty to start with, NULL is returned.

`file = popen(string, [flags])`

Executes a new process, specified as a shell command line as for the *system* function, and returns a file that either reads or writes to the standard input or output of the process according to *mode*. If mode is `"r"`, reading from the file reads from the standard output of the process. If mode is `"w"` writing to the file writes to the standard input of the process. If mode is not speicified it defaults to `"r"`.

On some commands and systems this may block, but will allow thread switching while blocked.

`float = pow(x, y)`

Returns *x^y* where both *x* and *y* are floats.

`printf([file,] fmt, args...)`

Formats a string based on *fmt* and *args* as per *sprintf* (below) and outputs the result to the *file* or to the current value of the *stdout* variable in the current scope if the first parameter is not a file. The current stdout must be a file. See *sprintf*.

On some files and systems this may block, but will allow thread switching while blocked.

`profile(filename)`

Enables profiling within the scope of the current function (must be called within a function). This profiler measures actual ellapsed time so it's only very useful for quite coarse profiling tasks. The *filename* specifies a file to write the profiling records to once it is complete. The profiling completes when the function *profile()* was called from returns. The file contains a re-parsable ICI data structue of the form:

```
auto profile = [struct
    total = <time in ms for this call>,
    call_count = <number of call to this func>,
    calls = [struct <nested profile structs...>],
];
```

For example, the following program:

```
static
count10000()
{
    j = 0;
    for (i = 0; i < 10000; ++i)
        j += i;
}

static
count20000()
{
    count10000();
    count10000();
}

static
prof()
{
    profile("prof.txt");
    count10000();
    count20000();
}
```

```
prof();
```

Would produce a file “prof.txt” file looking something like:

```
auto profile = [struct
 total = 153,
 call_count = 0,
 calls = [struct
  ("count20000()") = [struct
   total = 96,
   call_count = 1,
   calls = [struct
    ("count10000()") = [struct
     total = 96,
     call_count = 2,
     calls = [struct
     ],
    ],
   ],
  ],
  ("count10000()") = [struct
   total = 57,
   call_count = 1,
   calls = [struct
   ],
  ],
 ],
];
```

any = push(array, any)

Appends *any* to *array*, increasing its length in the process. Returns *any*.

put(string [, file])

Outputs string to *file*. If *file* is not passed the current value of *stdout* in the current scope is used.

int = rand([seed])

Returns a pseudo random integer in the range 0..0x7FFF. If *seed* (an int) is supplied the random number generator is first seeded with that number. The sequence is predictable based on a given seed.

reclaim()

Force a garbage collection to occur.

`re = regexp(string [, int])`

Returns a compiled regular expression derived from *string* This is the method of generating regular expressions at run-time, as opposed to the direct lexical form. For example, the following three expressions are similar:

```
str ~ #*\.c#
str ~ regexp("*\\.c");
str ~ $regexp("*\\.c");
```

except that the middle form computes the regular expression each time it is executed. Note that when a regular expression includes a # character the *regexp* function must be used, as the direct lexical form has no method of escaping a #.

The optional second parameter is a bit-set that controls various aspects of the compiled regular expression's behaviour. This value is passed directly to the PCRE package's regular expression compilation function. Presently no symbolic names are defined for the possible values and interested parties are directed to the PCRE documention included with the ICI source code.

Note that regular expressions are intrinsically atomic. Also note that non-equal strings may sometimes compile to the same regular expression.

`re = regexpi(string [, int])`

Returns a compiled regular expression derived from *string* that is case-insensitive. I.e., the regexp will match a string regardless of the case of alphabetic characters. Literal regular expressions to perform case-insensitive matching may be constructed using the special PCRE notation for such purposes, see the capter on regualr expressions for detais.

`int = inst|class:respondsto(name)`

Returns 1 if *inst* or *class* supports a function called *name*, else 0.

`remove(string)`

Deletes the file whose name is given in *string*.

`rename(oldname, newname)`

Change the name of a file. The first parameter is the name of an existing file and the second is the new name that it is to be given.

`any = rpop(array)`

Returns the first element of *array* and removes that element from *array*, thus shortening it by one. If the array was empty to start with, NULL is returned. After this the item that was at index 1 will be at index 0. This is an efficient constant time operation (that is, no actual data copying is done).

any = rpush(array, any)

Inserts *any* as the first element of the array, increasing the length of *array* in the process. After this the item that was at index 0 will be at index 1. The passed *any* is returned unchanged. This is an efficient constant time operation (that is, no actual data copying is done).

current = scope([replacement])

Returns the current scope structure. This is a struct whose base element holds the auto variables, the super of that hold the statics, the super of that holds the externs etc. Note that this is a real reference to the current scope structure. Changing, adding and deleting elements of these structures will affect the values and presence of variables in the current scope.

If a *replacement* is given, that struct replaces the current scope structure, with the obvious implications. This should clearly be used with caution. Replacing the current scope with a structure which has no reference to the standard functions also has the obvious effect.

int = seek(file, int, int)

Set the input/output position for a file and returns the new I/O position or -1 if an error ocurred. The arguments are the same as for the C library's **fseek** function. If the file object does not support setting the I/O position or the seek operation fails an error is raised.

set = set(any...)

Returns a set formed from all the arguments. For example:

```
set()
```

will return a new empty set; and

```
set(1, 2, "a string")
```

will return a new set with three elements, *1*, *2*, and *"the string"*.

This is the run-time equivalent of the set literal. Thus the following two expressions are equivalent:

```
$set(1, 2, "a string")
```

```
[set 1, 2, "a string"]
```

func = signal(string|int [, string|func])

Allows control of signal handling to the process running the ICI interpreter. The first argument is the name or number of a signal. Signal numbers are defined by the system whilst the function signam() may be used to obtain signal names. If no second argument is given the function returns the current handler for the signal. Handlers are either functions or one of the strings "default" or "ignore". If a second argument is given the signal handler's state is set accordingly, either being reset to its default state, ignored or calling the given function when the signal occurs. The previous signal handler is returned in this case.

`string = signam(int)`

Returns the name of a signal given its number. If the signal number is not valid an error is raised.

`x = sin(angle)`

Returns the sine of *angle* (a float interpreted in radians).

`sleep(num)`

Suspends execution of the current thread for *num* seconds. The resolution of *num* is system dependent.

`array = smash(string [, regexp [, replace...] [, include_remainder])`

Returns an array containing expanded replacement strings that are the result of repeatedly applying the regular expression *regexp* to successive portions of *string*. This process stops as soon as the regular expression fails to match or the string is exhausted.

Each time the regular expression is matched against the string, expanded copies of all the *replace* strings are pushed onto the newly created arrary. The expansion is done by performing the following substitutions:

\0	Is substituted with any leading unmatched portion between the end of the last match (or the start of the string if this is the first match) and the first character that was matched by this match.
\&	Is substituted with the portion of the string that was matched by this application of the regular expression.
\1 \2 \3 ...	Is substituted with the portions of the string that were matched by the succsesive bracketed sub-portions of the regular expression.
\\	Is substituted with a single \ character.

If the final argument, *include_remainder*, is supplied and is a non-zero integer, any remaining unmatched portion of the string is also added as a final element of the array. Else any unmatched remainder is discarded.

If *regexp* is not supplied, the regular expression `#\n#` is used. If no *replace* arguments are supplied, the single string `"\0"` is used. Thus by default smash will break the given string into its newline delimited portions (although it will discard any final undelimited line unless *include_remainder* is specified).

For example:

```
lines = smash(getfile(f), 1);
```

will result in an array of all the lines of the file, with newlines characters discarded. While:

```
smash("ab cd ef", #(.) #, "x\\0", 1);
```

will result in an array of the form:

```
[array "xa", "xc", "ef"]
```

Notice that it is generally necessary to use two backslash characters in literal strings to obtain the single backslash required here.

file = sopen(string [, mode])

Returns a *file*, which when read will fetch successive characters from the given *string*. The file is read-only and the *mode*, if passed, must be one of **"r"** or **"rb"**.

Files are, in general, system dependent. This is the only standard routine which opens a file. But on systems that support byte stream files, the function *fopen* will be set to the most appropriate method of opening a file for general use. The interpretation of *mode* is largely system dependent, but the strings *"r"*, *"w"*, and *"rw"* should be used for read, write, and read-write file access respectively.

sort(array [, func [, arg]])

Sort the content of the array using the heap sort algorithm with func as the comparison function. The comparison function is called with two elements of the array as parameters, *a* and *b*, and the optional *arg*. If *a* is equal to *b* the function should return zero. If *a* is less than *b*, -1, and if *a* is greater than *b*, 1.

For example,

```
static compare(a, b, arg)
{
    return a < b ? -1 : a > b;
}

static a = array(1, 3, 2);

sort(a, compare);
```

If *arg* is not provided, NULL is passed. If *func* is not provided, the current value of *cmp* in the current scope is used. See *cmp()*.

string = sprintf(fmt, args...)

Return a formatted string based on *fmt* (a string) and *args*. Most of the usual % format escapes of ANSI C printf are supported. In particular; the integer format letters *diouxXc* are supported, but if a float is provided it will be converted to an int. The floating point format letters *feEgG* are supported, but if the argument is an int it will be converted to a float. The string format letter, *s* is supported and requires a string. Finally the *%* format to get a single *%* works.

The flags, precision, and field width options are supported. The indirect field width and precision options with * also work and the corresponding argument must be an int.

For example:

```
sprintf("%08X <%4s> <%-4s>", 123, "ab", "cd")
```

will produce the string:

```
0000007B <  ab> <cd  >
```

and

```
sprintf("%0*X", 4, 123)
```

will produce the string:

```
007B
```

x = sqrt(float)

Returns the square root of *float*.

string = string(any)

Returns a short textual representation of *any*. If *any* is an int or float it is converted as if by a *%d* or *%g* format. If it is a string it is returned directly. Any other type will returns its type name surrounded by angle brackets, as in <*struct*>.

struct = struct([super,] key, value...)

Returns a new structure. This is the run-time equivalent of the struct literal. If there are an odd number of arguments the first is used as the super of the new struct; it must be a struct. The remaining pairs of arguments are treated as key and value pairs to initialise the structure with; they may be of any type. For example:

```
struct()
```

returns a new empty struct;

```
struct(anotherStruct)
```

returns a new empty struct which has *anotherStruct* as its super;

```
struct("a", 1, "b", 2)
```

returns a new struct which has two entries *a* and *b* with the values *1* and *2*; and

```
struct(anotherStruct, "a", 1, "b", 2)
```

returns a new struct which has two entries *a* and *b* with the values *1* and *2* and a super of *anotherStruct*.

Note that the super of the new struct is set *after* the assignments of the new elements have been made. Thus the initial elements given as arguments will not affect values in any super struct.

The following two expressions are equivalent:

```
$struct(anotherStruct, "a", 1, "b", 2)
```

```
[struct:anotherStruct, a = 1, b = 2]
```

string = sub(string, string|regexp, string)

Sub performs text substitution using regular expressions. It takes the first parameter, matches it against the second parameter and then replaces the matched portion of the string with the third parameter. If the second parameter is a string it is converted to a regular expression as if the regexp function had been called. Sub does the replacement once (unlike gsub). It returns the new string formed by the replacement. If there is no match this is the original string. The replacement string may contain the special sequence "\&" which is replaced by the string that matched the regular expression. Parenthesized portions of the regular expression may be matched by using *n* where *n* is a decimal digit.

current = super(struct [, replacement])

Returns the current super struct of *struct*, and, if *replacement* is supplied, sets it to a new value. If *replacement* is NULL any current super struct reference is cleared (that is, after this *struct* will have no super).

int = system(string)

Executes a new process, specified as a shell command line using the local system's command interpreter, and returns an integer result code once the process completes (usually zero indicates normal successful completion).

This will block while the process runs, but will allow thread switching while blocked.

x = tan(angle)

Returns the tangent of *angle* (a float interpreted in radians).

exec = thread(callable, args...)

Creates a new ICI thread and calls *callable* (typically a function or method) with *args* in the new ICI execution context in that thead. Returns an execution context object ("exec"). When the thread terminates (by returning from the called function) this object is woken up with *wakeup()*.

string = tochar(int)

Returns a one character string made from the character code specified by *int*.

int = toint(string)

Returns the character code of the first character of *string*.

any = top(array [, int])

Returns the last element of *array* (that is, the top of stack). Or, if *int* is supplied, objects from deep in the stack found by adding *int* to the index of the last element. Thus:

```
top(a, 0)
```

and

```
top(a)
```

are equvalent, while

```
top(a, -1)
```

returns the second last element of the array. Returns NULL if the access is beyond the limits of the array.

int = trace(string)

Enables diagnostic tracing of internal interpreter activity or program flow. The *string* consists of space seperated option words. There is a global enable/disable flag for tracing, and if is enables, a number of sub-flags indicating what should be traced. Trace output is printed to the interpreters standard error output. The options are interpreted as follows:

lexer	Flags tracing of every character read by the lexical analyser.
expr	### To be completed (and checked in source).
calls	
funcs	
all	
mem	
src	
gc	
none	
off	
on	

string = typeof(any)

Returns the type name (a string) of *any.* See the section on types above for the possible type names.

`string = version()`

Returns a version string of the form.

```
@(#)ICI 4.0.0 config-file build-date config-str (opts...)
```

For example:

```
@(#)ICI 4.0.0, conf-w32.h, Feb 22 2002, Microsoft Win32
platforms (math trace system pipes sockets dir dload
startupfile debugging )
```

`array = vstack()`

Returns a representation of the call stack of the current program at the time of the call. It can be used to perform stack tracebacks and related debugging operations. The result is an array of structures, each of which is a variable scope (see *scope*) structure of succesively deeper nestings of the current function nesting.

`event = waitfor(event...)`

###waitfor has the same name as the new waitfor statement. But I doubt anybody is using this function. Can we retire it? TML

Blocks (waits) until an *event* indicated by any of its arguments occurs, then returns that argument. The interpretation of an event depends on the nature of each argument. A file argument is triggered when input is available on the file. A float argument waits for that many seconds to expire, an int for that many millisecond (they then return 0, not the argument given). Other interpretations are implementation dependent. Where several events occur simultaneously, the first as listed in the arguments will be returned.

Note that in some implementations some file types may always appear ready for input, despite the fact that they are not.

`wakeup(any)`

Wakes up all ICI threads that are waiting for *any* (and thus allow them re-evaluate their wait expression).

Command Line Arguments

Versions of ICI on systems that support passing parameters from the command line provide two predefined variables, argv and argc, for accessing these arguments.

On Win32 platforms ICI performs wildcard expansion in the traditional MS-DOS fashion. Arguments containing wildcard meta-characters, ‘?’ and ‘*’, may be protected by enclosing them in single or double quotes.

argv

An array of strings containing the command line arguments. The first element is the name of the ICI program and subsequent elements are the arguments passed to that program.

argc

The count of the number of elements in argv. Initially equal to nels(argv).

CHAPTER 7

Regular expressions

Regular Expression Syntax

ICI uses Philip Hazel's PCRE (Perl-compatible regular expressions) package. The following is extracted from the file `pcre.3.txt` included with the PCRE distribution. This document is intended to be used with the PCRE C functions and makes reference to a number of constants that may be used as option specifiers to the C functions (all such constants are prefixed with the string "`PCRE_`"). These constants are not available in the ICI interface at time of writing although the `regexp()` function does allow a numeric option specific to be passed.

The syntax and semantics of the regular expressions supported by PCRE are described below. Regular expressions are also described in the Perl documentation and in a number of other books, some of which have copious examples. Jeffrey Friedl's "Mastering Regular Expressions", published by O'Reilly (ISBN 1-56592-257-3), covers them in great detail. The description here is intended as reference documentation.

A regular expression is a pattern that is matched against a subject string from left to right. Most characters stand for themselves in a pattern, and match the corresponding characters in the subject. As a trivial example, the pattern

The quick brown fox

matches a portion of a subject string that is identical to itself. The power of regular expressions comes from the ability to include alternatives and repetitions in the pattern. These are encoded in the pattern by the use of meta-characters, which do not stand for themselves but instead are interpreted in some special way.

There are two different sets of meta-characters: those that are recognized anywhere in the pattern except within square brackets, and those that are recognized in square brackets. Outside square brackets, the meta-characters are as follows:

\ general escape character with several uses

^ assert start of subject (or line, in multiline mode)

$ assert end of subject (or line, in multiline mode)

. match any character except newline (by default)

[start character class definition

| start of alternative branch

(start subpattern

) end subpattern

? extends the meaning of (

also 0 or 1 quantifier

also quantifier minimizer

* 0 or more quantifier

+ 1 or more quantifier

{ start min/max quantifier

Part of a pattern that is in square brackets is called a “character class”. In a character class the only meta-characters are:

\ general escape character

^ negate the class, but only if the first character

- indicates character range

] terminates the character class

The following sections describe the use of each of the meta-characters.

BACKSLASH

The backslash character has several uses. Firstly, if it is followed by a non-alphameric character, it takes away any special meaning that character may have. This use of backslash as an escape character applies both inside and outside character classes.

For example, if you want to match a “*” character, you write “*” in the pattern. This applies whether or not the following character would otherwise be interpreted as a meta-character, so it is always safe to precede a non-alphameric with “\”

to specify that it stands for itself. In particular, if you want to match a backslash, you write "\\".

If a pattern is compiled with the PCRE_EXTENDED option, whitespace in the pattern (other than in a character class) and characters between a "#" outside a character class and the next newline character are ignored. An escaping backslash can be used to include a whitespace or "#" character as part of the pattern.

A second use of backslash provides a way of encoding non-printing characters in patterns in a visible manner. There is no restriction on the appearance of non-printing characters, apart from the binary zero that terminates a pattern, but when a pattern is being prepared by text editing, it is usually easier to use one of the following escape sequences than the binary character it represents:

\a alarm, that is, the BEL character (hex 07)

\cx "control-x", where x is any character

\e escape (hex 1B)

\f formfeed (hex 0C)

\n newline (hex 0A)

\r carriage return (hex 0D)

\t tab (hex 09)

\xhh character with hex code hh

\ddd character with octal code ddd, or backreference

The precise effect of "\cx" is as follows: if "x" is a lower case letter, it is converted to upper case. Then bit 6 of the character (hex 40) is inverted. Thus "\cz" becomes hex 1A, but "\c{" becomes hex 3B, while "\c;" becomes hex 7B.

After "\x", up to two hexadecimal digits are read (letters can be in upper or lower case).

After "\0" up to two further octal digits are read. In both cases, if there are fewer than two digits, just those that are present are used. Thus the sequence "\0\x\07" specifies two binary zeros followed by a BEL character. Make sure you supply two digits after the initial zero if the character that follows is itself an octal digit.

The handling of a backslash followed by a digit other than 0 is complicated. Outside a character class, PCRE reads it and any following digits as a decimal number. If the number is less than 10, or if there have been at least that many previous capturing left parentheses in the expression, the entire sequence is taken as a back reference. A description of how this works is given later, following

the discussion of parenthesized subpatterns.

Inside a character class, or if the decimal number is greater than 9 and there have not been that many capturing subpatterns, PCRE re-reads up to three octal digits following the backslash, and generates a single byte from the least significant 8 bits of the value. Any subsequent digits stand for themselves. For example:

\040 is another way of writing a space

\40 is the same, provided there are fewer than 40 previous capturing subpatterns

\7 is always a back reference

\11 might be a back reference, or another way of writing a tab

\011 is always a tab

\0113 is a tab followed by the character "3"

\113 is the character with octal code 113 (since there can be no more than 99 back references)

\377 is a byte consisting entirely of 1 bits

\81 is either a back reference, or a binary zero followed by the two characters "8" and "1"

Note that octal values of 100 or greater must not be introduced by a leading zero, because no more than three octal digits are ever read.

All the sequences that define a single byte value can be used both inside and outside character classes. In addition, inside a character class, the sequence "\b" is interpreted as the backspace character (hex 08). Outside a character class it has a different meaning (see below).

The third use of backslash is for specifying generic character types:

\d any decimal digit

\D any character that is not a decimal digit any whitespace character

\S any character that is not a whitespace character

\w any "word" character

\W any "non-word" character

Each pair of escape sequences partitions the complete set of characters into two

disjoint sets. Any given character matches one, and only one, of each pair.

A "word" character is any letter or digit or the underscore character, that is, any character which can be part of a Perl "word". The definition of letters and digits is controlled by PCRE's character tables, and may vary if locale-specific matching is taking place (see "Locale support" above). For example, in the "fr" (French) locale, some character codes greater than 128 are used for accented letters, and these are matched by \w.

These character type sequences can appear both inside and outside character classes. They each match one character of the appropriate type. If the current matching point is at the end of the subject string, all of them fail, since there is no character to match.

The fourth use of backslash is for certain simple assertions. An assertion specifies a condition that has to be met at a particular point in a match, without consuming any characters from the subject string. The use of subpatterns for more complicated assertions is described below. The backslashed assertions are

\b word boundary

\B not a word boundary

\A start of subject (independent of multiline mode)

\Z end of subject or newline at end (independent of multiline mode)

\z end of subject (independent of multiline mode)

These assertions may not appear in character classes (but note that "\b" has a different meaning, namely the backspace character, inside a character class).

A word boundary is a position in the subject string where the current character and the previous character do not both match \w or \W (i.e. one matches \w and the other matches \W), or the start or end of the string if the first or last character matches \w, respectively.

The \A, \Z, and \z assertions differ from the traditional circumflex and dollar (described below) in that they only ever match at the very start and end of the subject string, whatever options are set. They are not affected by the PCRE_NOTBOL or PCRE_NOTEOL options. If the startoffset argument of pcre_exec() is non-zero, \A can never match. The difference between \Z and \z is that \Z matches before a newline that is the last character of the string as well as at the end of the string, whereas \z matches only at the end.

CIRCUMFLEX AND DOLLAR

Outside a character class, in the default matching mode, the circumflex character is an assertion which is true only if the current matching point is at the start of the subject string. If the startoffset argument of pcre_exec() is non-zero, circumflex can never match. Inside a character class, circumflex has an entirely different meaning (see below).

Circumflex need not be the first character of the pattern if a number of alternatives are involved, but it should be the first thing in each alternative in which it appears if the pattern is ever to match that branch. If all possible alternatives start with a circumflex, that is, if the pattern is constrained to match only at the start of the subject, it is said to be an "anchored" pattern. (There are also other constructs that can cause a pattern to be anchored.)

A dollar character is an assertion which is true only if the current matching point is at the end of the subject string, or immediately before a newline character that is the last character in the string (by default). Dollar need not be the last character of the pattern if a number of alternatives are involved, but it should be the last item in any branch in which it appears. Dollar has no special meaning in a character class.

The meaning of dollar can be changed so that it matches only at the very end of the string, by setting the PCRE_DOLLAR_ENDONLY option at compile or matching time. This does not affect the \Z assertion.

The meanings of the circumflex and dollar characters are changed if the PCRE_MULTILINE option is set. When this is the case, they match immediately after and immediately before an internal "\n" character, respectively, in addition to matching at the start and end of the subject string. For example, the pattern /^abc$/ matches the subject string "def\nabc" in multiline mode, but not otherwise. Consequently, patterns that are anchored in single line mode because all branches start with "^" are not anchored in multiline mode, and a match for circumflex is possible when the startoffset argument of pcre_exec() is non-zero. The PCRE_DOLLAR_ENDONLY option is ignored if PCRE_MULTILINE is set.

Note that the sequences \A, \Z, and \z can be used to match the start and end of the subject in both modes, and if all branches of a pattern start with \A is it always anchored, whether PCRE_MULTILINE is set or not.

FULL STOP (PERIOD, DOT)

Outside a character class, a dot in the pattern matches any one character in the subject, including a non-printing character, but not (by default) newline. If the PCRE_DOTALL option is set, then dots match newlines as well. The handling of dot is entirely independent of the handling of circumflex and dollar, the only relationship being that they both involve newline characters. Dot has no special meaning in a character class.

SQUARE BRACKETS

An opening square bracket introduces a character class, terminated by a closing square bracket. A closing square bracket on its own is not special. If a closing square bracket is required as a member of the class, it should be the first data character in the class (after an initial circumflex, if present) or escaped with a backslash.

A character class matches a single character in the subject; the character must be in the set of characters defined by the class, unless the first character in the class is a circumflex, in which case the subject character must not be in the set defined by the class. If a circumflex is actually required as a member of the class, ensure it is not the first character, or escape it with a backslash.

For example, the character class [aeiou] matches any lower case vowel, while [^aeiou] matches any character that is not a lower case vowel. Note that a circumflex is just a convenient notation for specifying the characters which are in the class by enumerating those that are not. It is not an assertion: it still consumes a character from the subject string, and fails if the current pointer is at the end of the string.

When caseless matching is set, any letters in a class represent both their upper case and lower case versions, so for example, a caseless [aeiou] matches "A" as well as "a", and a caseless [^aeiou] does not match "A", whereas a caseful version would.

The newline character is never treated in any special way in character classes, whatever the setting of the PCRE_DOTALL or PCRE_MULTILINE options is. A class such as [^a] will always match a newline.

The minus (hyphen) character can be used to specify a range of characters in a character class. For example, [d-m] matches any letter between d and m, inclusive. If a minus character is required in a class, it must be escaped with a backslash or appear in a position where it cannot be interpreted as indicating a range, typically as the first or last character in the class.

It is not possible to have the literal character "]" as the end character of a range. A pattern such as [W-]46] is interpreted as a class of two characters ("W" and "-") followed by a literal string "46]", so it would match "W46]" or "-46]". However, if the "]" is escaped with a backslash it is interpreted as the end of range, so [W-\]46] is interpreted as a single class containing a range followed by two separate characters. The octal or hexadecimal representation of "]" can also be used to end a range.

Ranges operate in ASCII collating sequence. They can also be used for characters specified numerically, for example [\000-\037]. If a range that includes letters is used when caseless matching is set, it matches the letters in either case.

For example, [W-c] is equivalent to [][\^_`wxyzabc], matched caselessly, and if character tables for the "fr" locale are in use, [\xc8-\xcb] matches accented E characters in both cases.

The character types \d, \D, \s, \S, \w, and \W may also appear in a character class, and add the characters that they match to the class. For example, [\dABCDEF] matches any hexadecimal digit. A circumflex can conveniently be used with the upper case character types to specify a more restricted set of characters than the matching lower case type. For example, the class [^\W_] matches any letter or digit, but not underscore.

All non-alphameric characters other than \, -, ^ (at the start) and the terminating] are non-special in character classes, but it does no harm if they are escaped.

VERTICAL BAR

Vertical bar characters are used to separate alternative patterns. For example, the pattern

```
gilbert|sullivan
```

matches either "gilbert" or "sullivan". Any number of alternatives may appear, and an empty alternative is permitted (matching the empty string). The matching process tries each alternative in turn, from left to right, and the first one that succeeds is used. If the alternatives are within a subpattern (defined below), "succeeds" means matching the rest of the main pattern as well as the alternative in the subpattern.

INTERNAL OPTION SETTING

The settings of PCRE_CASELESS, PCRE_MULTILINE, PCRE_DOTALL, and PCRE_EXTENDED can be changed from within the pattern by a sequence of Perl option letters enclosed between "(?" and ")". The option letters are

i for PCRE_CASELESS

m for PCRE_MULTILINE

s for PCRE_DOTALL

x for PCRE_EXTENDED

For example, (?im) sets caseless, multiline matching. It is also possible to unset these options by preceding the letter with a hyphen, and a combined setting and unsetting such as (?im-sx), which sets PCRE_CASELESS and

PCRE_MULTILINE while unsetting PCRE_DOTALL and PCRE_EXTENDED, is also permitted. If a letter appears both before and after the hyphen, the option is unset.

The scope of these option changes depends on where in the pattern the setting occurs. For settings that are outside any subpattern (defined below), the effect is the same as if the options were set or unset at the start of matching. The following patterns all behave in exactly the same way:

(?i)abc a(?i)bc ab(?i)c abc(?i)

which in turn is the same as compiling the pattern abc with PCRE_CASELESS set. In other words, such "top level" settings apply to the whole pattern (unless there are other changes inside subpatterns). If there is more than one setting of the same option at top level, the rightmost setting is used.

If an option change occurs inside a subpattern, the effect is different. This is a change of behaviour in Perl 5.005. An option change inside a subpattern affects only that part of the subpattern that follows it, so

(a(?i)b)c

matches abc and aBc and no other strings (assuming PCRE_CASELESS is not used). By this means, options can be made to have different settings in different parts of the pattern. Any changes made in one alternative do carry on into subsequent branches within the same subpattern. For example,

(a(?i)b|c)

matches "ab", "aB", "c", and "C", even though when matching "C" the first branch is abandoned before the option setting. This is because the effects of option settings happen at compile time. There would be some very weird behaviour otherwise.

The PCRE-specific options PCRE_UNGREEDY and PCRE_EXTRA can be changed in the same way as the Perl-compatible options by using the characters U and X respectively. The (?X) flag setting is special in that it must always occur earlier in the pattern than any of the additional features it turns on, even when it is at top level. It is best put at the start.

SUBPATTERNS

Subpatterns are delimited by parentheses (round brackets), which can be nested. Marking part of a pattern as a subpattern does two things:

1. It localizes a set of alternatives. For example, the pattern

```
cat(aract|erpillar|)
```

matches one of the words “cat”, “cataract”, or “caterpillar”. Without the parentheses, it would match “cataract”, “erpillar” or the empty string.

2. It sets up the subpattern as a capturing subpattern (as defined above). When the whole pattern matches, that portion of the subject string that matched the subpattern is passed back to the caller via the ovector argument of pcre_exec(). Opening parentheses are counted from left to right (starting from 1) to obtain the numbers of the capturing subpatterns.

For example, if the string “the red king” is matched against the pattern

```
the ((red|white) (king|queen))
```

the captured substrings are “red king”, “red”, and “king”, and are numbered 1, 2, and 3.

The fact that plain parentheses fulfil two functions is not always helpful. There are often times when a grouping subpattern is required without a capturing requirement. If an opening parenthesis is followed by “?:”, the subpattern does not do any capturing, and is not counted when computing the number of any subsequent capturing subpatterns. For example, if the string “the white queen” is matched against the pattern

```
the ((?:red|white) (king|queen))
```

the captured substrings are “white queen” and “queen”, and are numbered 1 and 2. The maximum number of captured substrings is 99, and the maximum number of all subpatterns, both capturing and non-capturing, is 200.

As a convenient shorthand, if any option settings are required at the start of a non-capturing subpattern, the option letters may appear between the “?” and the “:”. Thus the two patterns

```
(?i:saturday|sunday)
```

```
(?:(?i)saturday|sunday)
```

match exactly the same set of strings. Because alternative branches are tried from left to right, and options are not reset until the end of the subpattern is reached, an option setting in one branch does affect subsequent branches, so the above patterns match “SUNDAY” as well as “Saturday”.

REPETITION

Repetition is specified by quantifiers, which can follow any of the following

items:

a single character, possibly escaped

the . metacharacter

a character class

a back reference (see next section)

a parenthesized subpattern (unless it is an assertion - see below)

The general repetition quantifier specifies a minimum and maximum number of permitted matches, by giving the two numbers in curly brackets (braces), separated by a comma. The numbers must be less than 65536, and the first must be less than or equal to the second. For example:

```
z{2,4}
```

matches “zz”, “zzz”, or “zzzz”. A closing brace on its own is not a special character. If the second number is omitted, but the comma is present, there is no upper limit; if the second number and the comma are both omitted, the quantifier specifies an exact number of required matches. Thus

```
[aeiou]{3,}
```

matches at least 3 successive vowels, but may match many more, while

```
\d{8}
```

matches exactly 8 digits. An opening curly bracket that appears in a position where a quantifier is not allowed, or one that does not match the syntax of a quantifier, is taken as a literal character. For example, {,6} is not a quantifier, but a literal string of four characters.

The quantifier {0} is permitted, causing the expression to behave as if the previous item and the quantifier were not present.

For convenience (and historical compatibility) the three most common quantifiers have single-character abbreviations:

* is equivalent to {0,}

\+ is equivalent to {1,}

? is equivalent to {0,1}

It is possible to construct infinite loops by following a subpattern that can match no characters with a quantifier that has no upper limit, for example:

```
(a?)*
```

Earlier versions of Perl and PCRE used to give an error at compile time for such patterns. However, because there are cases where this can be useful, such patterns are now accepted, but if any repetition of the subpattern does in fact match no characters, the loop is forcibly broken.

By default, the quantifiers are "greedy", that is, they match as much as possible (up to the maximum number of permitted times), without causing the rest of the pattern to fail. The classic example of where this gives problems is in trying to match comments in C programs. These appear between the sequences /* and */ and within the sequence, individual * and / characters may appear. An attempt to match C comments by applying the pattern

```
/\*.*\*/
```

to the string

```
/* first command */  not comment  /* second comment */
```

fails, because it matches the entire string due to the greediness of the .* item.

However, if a quantifier is followed by a question mark, then it ceases to be greedy, and instead matches the minimum number of times possible, so the pattern

```
/\*.*?\*/
```

does the right thing with the C comments. The meaning of the various quantifiers is not otherwise changed, just the preferred number of matches. Do not confuse this use of question mark with its use as a quantifier in its own right. Because it has two uses, it can sometimes appear doubled, as in

```
\d??\d
```

which matches one digit by preference, but can match two if that is the only way the rest of the pattern matches.

If the PCRE_UNGREEDY option is set (an option which is not available in Perl) then the quantifiers are not greedy by default, but individual ones can be made greedy by following them with a question mark. In other words, it inverts the default behaviour.

When a parenthesized subpattern is quantified with a minimum repeat count that is greater than 1 or with a limited maximum, more store is required for the compiled pattern, in proportion to the size of the minimum or maximum.

If a pattern starts with .* or .{0,} and the PCRE_DOTALL option (equivalent to Perl's /s) is set, thus allowing the . to match newlines, then the pattern is implic-

itly anchored, because whatever follows will be tried against every character position in the subject string, so there is no point in retrying the overall match at any position after the first. PCRE treats such a pattern as though it were preceded by \A. In cases where it is known that the subject string contains no newlines, it is worth setting PCRE_DOTALL when the pattern begins with .* in order to obtain this optimization, or alternatively using ^ to indicate anchoring explicitly.

When a capturing subpattern is repeated, the value captured is the substring that matched the final iteration. For example, after

```
(tweedle[dume]{3}\s*)+
```

has matched "tweedledum tweedledee" the value of the captured substring is "tweedledee". However, if there are nested capturing subpatterns, the corresponding captured values may have been set in previous iterations. For example, after

```
/(a|(b))+/
```

matches "aba" the value of the second captured substring is "b".

BACK REFERENCES

Outside a character class, a backslash followed by a digit greater than 0 (and possibly further digits) is a back reference to a capturing subpattern earlier (i.e. to its left) in the pattern, provided there have been that many previous capturing left parentheses.

However, if the decimal number following the backslash is less than 10, it is always taken as a back reference, and causes an error only if there are not that many capturing left parentheses in the entire pattern. In other words, the parentheses that are referenced need not be to the left of the reference for numbers less than 10. See the section entitled "Backslash" above for further details of the handling of digits following a backslash.

A back reference matches whatever actually matched the capturing subpattern in the current subject string, rather than anything matching the subpattern itself. So the pattern

```
(sens|respons)e and \1ibility
```

matches "sense and sensibility" and "response and responsibility", but not "sense and responsibility". If caseful matching is in force at the time of the back reference, then the case of letters is relevant. For example,

```
((?i)rah)\s+\1
```

matches “rah rah” and “RAH RAH”, but not “RAH rah”, even though the original capturing subpattern is matched caselessly.

There may be more than one back reference to the same subpattern. If a subpattern has not actually been used in a particular match, then any back references to it always fail. For example, the pattern

```
(a|(bc))\2
```

always fails if it starts to match “a” rather than “bc”. Because there may be up to 99 back references, all digits following the backslash are taken as part of a potential back reference number. If the pattern continues with a digit character, then some delimiter must be used to terminate the back reference. If the PCRE_EXTENDED option is set, this can be whitespace. Otherwise an empty comment can be used.

A back reference that occurs inside the parentheses to which it refers fails when the subpattern is first used, so, for example, (a\1) never matches. However, such references can be useful inside repeated subpatterns. For example, the pattern

```
(a|b\1)+
```

matches any number of “a”s and also “aba”, “ababaa” etc. At each iteration of the subpattern, the back reference matches the character string corresponding to the previous iteration. In order for this to work, the pattern must be such that the first iteration does not need to match the back reference. This can be done using alternation, as in the example above, or by a quantifier with a minimum of zero.

ASSERTIONS

An assertion is a test on the characters following or preceding the current matching point that does not actually consume any characters. The simple assertions coded as \b, \B, \A, \Z, \z, ^ and $ are described above. More complicated assertions are coded as subpatterns. There are two kinds: those that look ahead of the current position in the subject string, and those that look behind it.

An assertion subpattern is matched in the normal way, except that it does not cause the current matching position to be changed. Lookahead assertions start with (?= for positive assertions and (?! for negative assertions. For example,

```
\w+(?=;)
```

matches a word followed by a semicolon, but does not include the semicolon in the match, and

```
foo(?!bar)
```

matches any occurrence of “foo” that is not followed by “bar”. Note that the apparently similar pattern

```
(?!foo)bar
```

does not find an occurrence of “bar” that is preceded by something other than “foo”; it finds any occurrence of “bar” whatsoever, because the assertion (?!foo) is always true when the next three characters are “bar”. A lookbehind assertion is needed to achieve this effect.

Look-behind assertions start with (?<= for positive assertions and (?<! for negative assertions. For example,

```
(?<!foo)bar
```

does find an occurrence of “bar” that is not preceded by “foo”. The contents of a lookbehind assertion are restricted such that all the strings it matches must have a fixed length. However, if there are several alternatives, they do not all have to have the same fixed length. Thus

```
(?<=bullock|donkey)
```

is permitted, but

```
(?<!dogs?|cats?)
```

causes an error at compile time. Branches that match different length strings are permitted only at the top level of a lookbehind assertion. This is an extension compared with Perl 5.005, which requires all branches to match the same length of string. An assertion such as

```
(?<=ab(c|de))
```

is not permitted, because its single top-level branch can match two different lengths, but it is acceptable if rewritten to use two top-level branches:

```
(?<=abc|abde)
```

The implementation of lookbehind assertions is, for each alternative, to temporarily move the current position back by the fixed width and then try to match. If there are insufficient characters before the current position, the match is deemed to fail. Lookbehinds in conjunction with once-only subpatterns can be particularly useful for matching at the ends of strings; an example is given at the end of the section on once-only subpatterns.

Several assertions (of any sort) may occur in succession. For example,

```
(?<=\d{3})(?<!999)foo
```

matches “foo” preceded by three digits that are not “999”. Notice that each of the assertions is applied independently at the same point in the subject string. First there is a check that the previous three characters are all digits, then there is a check that the same three characters are not “999”. This pattern does not match “foo” preceded by six characters, the first of which are digits and the last three of which are not “999”. For example, it doesn’t match “123abcfoo”. A pattern to do that is

```
(?<=\d{3}...)(?<!999)foo
```

This time the first assertion looks at the preceding six characters, checking that the first three are digits, and then the second assertion checks that the preceding three characters are not “999”.

Assertions can be nested in any combination. For example,

```
(?<=(?<!foo)bar)baz
```

matches an occurrence of “baz” that is preceded by “bar” which in turn is not preceded by “foo”, while

```
(?<=\d{3}(?!999)...)foo
```

is another pattern which matches “foo” preceded by three digits and any three characters that are not “999”.

Assertion subpatterns are not capturing subpatterns, and may not be repeated, because it makes no sense to assert the same thing several times. If any kind of assertion contains capturing subpatterns within it, these are counted for the purposes of numbering the capturing subpatterns in the whole pattern. However, substring capturing is carried out only for positive assertions, because it does not make sense for negative assertions.

Assertions count towards the maximum of 200 parenthesized subpatterns.

ONCE-ONLY SUBPATTERNS

With both maximizing and minimizing repetition, failure of what follows normally causes the repeated item to be re-evaluated to see if a different number of repeats allows the rest of the pattern to match. Sometimes it is useful to prevent this, either to change the nature of the match, or to cause it fail earlier than it otherwise might, when the author of the pattern knows there is no point in carrying on.

Consider, for example, the pattern \d+foo when applied to the subject line

```
123456bar
```

After matching all 6 digits and then failing to match “foo”, the normal action of the matcher is to try again with only 5 digits matching the \d+ item, and then with 4, and so on, before ultimately failing. Once-only subpatterns provide the means for specifying that once a portion of the pattern has matched, it is not to be re-evaluated in this way, so the matcher would give up immediately on failing to match “foo” the first time. The notation is another kind of special parenthesis, starting with (?> as in this example:

```
(?>\d+)bar
```

This kind of parenthesis “locks up” the part of the pattern it contains once it has matched, and a failure further into the pattern is prevented from backtracking into it. Backtracking past it to previous items, however, works as normal.

An alternative description is that a subpattern of this type matches the string of characters that an identical standalone pattern would match, if anchored at the current point in the subject string.

Once-only subpatterns are not capturing subpatterns. Simple cases such as the above example can be thought of as a maximizing repeat that must swallow everything it can. So, while both \d+ and \d+? are prepared to adjust the number of digits they match in order to make the rest of the pattern match, (?>\d+) can only match an entire sequence of digits.

This construction can of course contain arbitrarily complicated subpatterns, and it can be nested.

Once-only subpatterns can be used in conjunction with look-behind assertions to specify efficient matching at the end of the subject string. Consider a simple pattern such as

```
abcd$
```

when applied to a long string which does not match it. Because matching proceeds from left to right, PCRE will look for each “a” in the subject and then see if what follows matches the rest of the pattern. If the pattern is specified as

```
^.*abcd$
```

then the initial .* matches the entire string at first, but when this fails, it backtracks to match all but the last character, then all but the last two characters, and so on. Once again the search for “a” covers the entire string, from right to left, so we are no better off. However, if the pattern is written as

```
^(?>.*)(?<=abcd)
```

then there can be no backtracking for the .* item; it can match only the entire string. The subsequent lookbehind assertion does a single test on the last four characters. If it fails, the match fails immediately. For long strings, this ap-

proach makes a significant difference to the processing time.

CONDITIONAL SUBPATTERNS

It is possible to cause the matching process to obey a subpattern conditionally or to choose between two alternative subpatterns, depending on the result of an assertion, or whether a previous capturing subpattern matched or not. The two possible forms of conditional subpattern are

```
(?(condition)yes-pattern)

(?(condition)yes-pattern|no-pattern)
```

If the condition is satisfied, the yes-pattern is used; otherwise the no-pattern (if present) is used. If there are more than two alternatives in the subpattern, a compile-time error occurs.

There are two kinds of condition. If the text between the parentheses consists of a sequence of digits, then the condition is satisfied if the capturing subpattern of that number has previously matched. Consider the following pattern, which contains non-significant white space to make it more readable (assume the PCRE_EXTENDED option) and to divide it into three parts for ease of discussion:

```
( \( )?    [^()]+    (?(1) \) )
```

The first part matches an optional opening parenthesis, and if that character is present, sets it as the first captured substring. The second part matches one or more characters that are not parentheses. The third part is a conditional subpattern that tests whether the first set of parentheses matched or not. If they did, that is, if subject started with an opening parenthesis, the condition is true, and so the yes-pattern is executed and a closing parenthesis is required. Otherwise, since no-pattern is not present, the subpattern matches nothing. In other words, this pattern matches a sequence of non-parentheses, optionally enclosed in parentheses.

If the condition is not a sequence of digits, it must be an assertion. This may be a positive or negative lookahead or lookbehind assertion. Consider this pattern, again containing non-significant white space, and with the two alternatives on the second line:

```
(?(?=[^a-z]*[a-z])

\d{2}[a-z]{3}-\d{2} | \d{2}-\d{2}-\d{2} )
```

The condition is a positive lookahead assertion that matches an optional sequence of non-letters followed by a letter. In other words, it tests for the presence of at

least one letter in the subject. If a letter is found, the subject is matched against the first alternative; otherwise it is matched against the second. This pattern matches strings in one of the two forms dd-aaa-dd or dd-dd-dd, where aaa are letters and dd are digits.

COMMENTS

The sequence (?# marks the start of a comment which continues up to the next closing parenthesis. Nested parentheses are not permitted. The characters that make up a comment play no part in the pattern matching at all.

If the PCRE_EXTENDED option is set, an unescaped # character outside a character class introduces a comment that continues up to the next newline character in the pattern.

PERFORMANCE

Certain items that may appear in patterns are more efficient than others. It is more efficient to use a character class like [aeiou] than a set of alternatives such as (a|e|i|o|u). In general, the simplest construction that provides the required behaviour is usually the most efficient. Jeffrey Friedl's book contains a lot of discussion about optimizing regular expressions for efficient performance.

When a pattern begins with .* and the PCRE_DOTALL option is set, the pattern is implicitly anchored by PCRE, since it can match only at the start of a subject string. However, if PCRE_DOTALL is not set, PCRE cannot make this optimization, because the . metacharacter does not then match a newline, and if the subject string contains newlines, the pattern may match from the character immediately following one of them instead of from the very start. For example, the pattern

(.*) second

matches the subject "first\nand second" (where \n stands for a newline character) with the first captured substring being "and". In order to do this, PCRE has to retry the match starting after every newline in the subject.

If you are using such a pattern with subject strings that do not contain newlines, the best performance is obtained by setting PCRE_DOTALL, or starting the pattern with ^.* to indicate explicit anchoring. That saves PCRE from having to scan along the subject looking for a newline to restart at.

Beware of patterns that contain nested indefinite repeats. These can take a long time to run when applied to a string that does not match. Consider the pattern fragment

```
(a+)*
```

This can match “aaaa” in 33 different ways, and this number increases very rapidly as the string gets longer. (The * repeat can match 0, 1, 2, 3, or 4 times, and for each of those cases other than 0, the + repeats can match different numbers of times.) When the remainder of the pattern is such that the entire match is going to fail, PCRE has in principle to try every possible variation, and this can take an extremely long time.

An optimization catches some of the more simple cases such as

```
(a+)*b
```

where a literal character follows. Before embarking on the standard matching procedure, PCRE checks that there is a “b” later in the subject string, and if there is not, it fails the match immediately. However, when there is no following literal this optimization cannot be used. You can see the difference by comparing the behaviour of

```
(a+)*\d
```

with the pattern above. The former gives a failure almost instantly when applied to a whole line of “a” characters, whereas the latter takes an appreciable time with strings longer than about 20 characters.

AUTHOR

Philip Hazel <ph10@cam.ac.uk>
University Computing Service,
New Museums Site,
Cambridge CB2 3QG, England.
Phone: +44 1223 334714
Last updated: 29 July 1999

CHAPTER 8

Interfacing with C and C++

There are several levels at which the ICI interpreter can interface with C and C++ code. This chapter gives a collection of universal rules, then addresses common tasks. Each task can be considered in isolation to alleviate the reader from details beyond their current needs. Finally, a summary of ICI C API elements is given.

Some functions and macros are documented best in the comments in the ICI source code. The reader is expected to be a C/C++ programmer who is able to find and read such comments when this document gives only a general indication of function and example usage.

Universal rules and conventions

Include files and libraries

On most systems ICI is built as a dynamically loading library. It can be linked with statically if required. However for this description I will assume the normal case of dynamic loading.

To compile and run with ICI you will need, as a minimum:

ici.h	The ICI include file. Include this as necessary. This file is build specifically for each platform from ICI's internal include files.
icistr-setup.h	A utility include file to assist in defining ICI strings. See *Refering to ICI string from C* below.
ici4.{a,lib}	The library file containg the linkage to the dynamicly loading library (or the static library if not using dynamic loading). (Suffixes vary with OS.) Link against this when buildng your program.
ici4.{so,dll}	The dynamic loading library. (Suffixes vary with OS.) Should be somewhere were it will be found at run time.

ici4core.ici, ici4core{1,2,3}.ici Core language features written in ICI. These need to be somewhere ICI will find them at run time.

If you are writting modules that run from ICI, you will also want an ICI top level command:

ici, ici.exe, or *wici.exe* ICI command level executable. These just do argument parseing and invocation of the interpreter on supplied file arguments.

In broad terms ICI is either used as an adjunct to another application, or as the main program with specific custom modules providing special functionality. Most of what is described in this section is common to both.

The nature of ICI objects

ICI objects are structures that have a common 32 bit header (since version 4; in previous versions it was a 64 bit header). Pointers to objects are declared as either *object_t **, which is the type of the header, or a pointer to the particular structure (for example, *string_t **), depending on the circumstances of each piece of code, and depending whether the real type is known at that point. The macro *objof()* is often used to demote to a generic object pointer -- it is just a cast to *object_t **. Most types define a similar macro to promote to their specific type, as well as a macro such as *isstring()* or *isstruct()* to test if a paticular pointer points to an object of the given type. However, there is no particular requirement to use these macros. They are just casts and simple tests.

Garbage collection, *ici_incref()* and *ici_decref()*

ICI objects are garbage collected. Garbage collection can occur on any attempt to allocate memory using ICI's allocation routines. This is fairly often. Garbage collection has the effect of freeing any objects that the garbage collector thinks are not being usefully referenced. Failing to obey the rules associated with garbage collection can be disasterous and hard to debug. But the rules are fairly simple.

The ICI object header includes a *small* reference count field. This is a count of additional references to the object that would otherwise be invisible to the garbage collector. For example, if you keep a global static pointer to an object, the garbage collector would not be aware of that reference, and might free the object leaving you with an invalid pointer. So you must account for the reference by incrementing the reference count. However, references to the object from other ICI objects are visible to the garbage collector, so you do not need to account for these.

The macros *ici_incref()* and *ici_decref()* are used to increment and decrement reference counts. Note that the range of reference counts is quiet small. Their frequency of use if expected to be in relationship to the number of actual instances of C variables that simultaneously refere to the same object.

In practice, many calls to *ici_incref()* and *ici_decref()* can be avoided because objects are known to be referenced by other things. For example, when coding a function called from ICI, the objects that are passed as arguments are known to be referenced from the ICI operand stack, which is referenced by the current execution context, which has a reference count as long as it is running.

The error return convention

When coding functions that are called by the ICI interpreter, a simple mechanism is used for all error returns. Each function will have some return value to indicate an error (commonly 0 is success, 1 is an error, or for functions that returns a pointer, NULL will probably be the error indicator). In any case, the return value will only imply a boolean indicating that an error has occured. The nature of the error *must* have been set before return by setting the global character pointer *ici_error* to a short human readable string revealing what happened.

Note, however, that *only* the originator of an error condition should set *ici_error*. If you call another function that uses the error return convention, and it returns a failure, you must simply clean up you immediate situation (such as any required *ici_decref()* calls) and return your failure indication in turn. For example, the *ici_alloc()* function obeys the convention. Thus we might see the fragment:

```
if ((p = ici_alloc(sizeof(mytype)) == NULL)
     return 1;
```

Now, in setting and returning an error, your code will be losing control and must be concerned about the memory that you set the *ici_error* variable to point to. Simple static strings are, of course, of no concern. For example:

```
if (that_failed)
{
    ici_error = "failed to do that";
    return 1;
}
```

If you need to format a string (say with sprintf) you can avoid the necessity of a permanently allocated buffer by using an generic growable buffer provided by ICI. The buffer is pointed to by *ici_buf* and *always* has room for at least 120 characters. Thus me might see:

```
if (v > 256)
{
    sprintf(ici_buf, "%d set but 256 is limit", v);
    ici_error = ici_buf;
}
```

If the size of the generated message is not so limited, the buffer can be checked (or forced) to have a larger size with *ici_chkbuf()*. For example:

```
if (file == NULL)
{
    if (ici_chkbuf(40 + strlen(fname))
        return 1;
    sprintf(ici_buf, "could not open %s", fname);
    ici_error = ici_buf;
    return 1;
}
```

Notice *ici_chkbuf()* could fail, and if so, we return immediately (the error will inevitably be “ran out of memory”). The *40* above is some number clearly larger than the length of the format string.

One final point, which is not specifically to do with error returns, but commonly associated with them, is how to get a short human readable description of an ICI object suitable for diagnostics. The function *ici_objname()* can be used to get a small (guaranteed to fit within a 30 character buffer) diagnostic description of an object. For example:

```
{
    char        n1[30];
    char        n2[30];

    sprintf(ici_buf, "attempt to read %s keyed by %s",
        ici_objname(n1, o),
        ici_objname(n2, k));
    ici_error = ici_buf;
    return 1;
}
```

ICI's allocation functions

In general ICI uses the native *malloc()* and *free()* functions for its memory allocation. However, because ICI deals in many small objects, and because it needs to track memory usage to control when to run its mark/sweep garbage collector, ICI has a few allocation functions that you may wish to be aware of.

There are three pais of matched alloc/free functions. Memory allocated with one must be freed by the matching one. Common to all three is the property that they (try) to track the amount of memory allocated so that they can trigger garbage collections, and they may run the garbage collector before they return.

Two of them (*ici_talloc* / *ici_tfee* and *ici_nalloc* / *ici_nfree*) are designed to handle small objects efficiently (small meaning up to 64 bytes). They allocate small objects densly (no boundary words) out of larger raw allocations from *malloc()* and maintain fast free lists so that most allocations and frees can avoid function calls completely. However, in order to avoid boundary words they both require that the caller tells the free routines how much memory was asked for on the initial alloc. This is fairly easy 95% of the time, but where it can't be manage, you must use the simpler *ici_alloc* / *ici_free* pair. These are completely *malloc()* equivalent, except they handle the garbage collection and have the ususal error return convention.

The tracking of memory usage is only relevent to memory the garbage collector has some control over, meaning memory associated with ICI objects (that would get freed if the object was collected). So, technically, these routines should be used for memory associated with objects, and not other allocations. But in practice the division is not strict.

Common tasks

Writting a simple function that can be called from ICI

This is sometimes done as part of a dynamically loaded extension module, and at other times done in a program that uses ICI as a sub-module. By *simple function* we mean a function that takes and returns values that are directly equivalent to simple C data types.

Transfer from the ICI execution engine to an "intrinsic" function (as they are called) is designed to have extremely low overhead. Thus, on arrival at a function, the arguments are still ICI objects on the ICI operand stack. If you are dealing with simple argument and return types, you can then use the *ici_typecheck()* function to marshal your arguments into C variables, and a set of similar functions to return simple values.

Intrinsic functions have a return type of *int*, use C calling convention, follow the ususal error convention (1 is failure) and in simple cases declare no arguments. For example, a function that expects to be called with an *int* and a *float* from ICI, prints them, and returns the int plus one, would read:

```
static int
f_myfunc()
{
    long          i;
    double        f;

    if (ici_typecheck("if", &i, &f))
        return 1;
    printf("Got %ld, %lf.\n", i, f);
    return int_ret(i + 1);
}
```

Note that ICI ints are C longs, and ICI floats are C doubles.

The *ici_typecheck()* function allows marshalling of:

- ICI ints to C longs, floats to doubles, and strings to char pointers;
- generic "numeric" (int or float) values to a C double;
- many other ICI types (structs, arrays, generic objects, etc) to generic object pointers (see other tasks on how to deal with them after that);
- ICI pointers to any of the above.

It also supports skipping argument and variable argument lists, but these features are typically used in functions that use a mixture of *ici_typecheck()* calls and explicit argument processing. See *ici_typecheck()* in *cfunc.c*.

To return simple data types, the functions *ici_int_ret()*, *ici_float_ret()*, *ici_str_ret()*, and *ici_null_ret()* can be used. These convert the value to the approriate ICI data object, and replace the arguments on the operand stack with that object, then return 0.

Take care *never* to simply *"return 0;"* from an intrinsic function. Although returning 1 on error is correct, and the non-error return value is zero, before returning the arguments must be replaced with the return value on the ICI operand stack. The functions above, and various others, do this. They should always be used on successful return from any intrinsic function.

It is possible to write a function that is passed pointers to values from ICI, and have those values updated by the intrinsic function, by using a combination of the *ici_typecheck()* function and the *ici_retcheck()* function. For example:

```
static int
f_myotherfunc()
{
    long          i;
```

```
    double        f;

    if (ici_typecheck("fI", &f, &i))
        return 1;
    printf("Got %ld, %lf.\n", i, f);
    if (ici_retcheck("-i", i + 1))
        return 1;
    return float_ret(f * 3.0);
}
```

This function takes a float, and a pointer to an int. It returns three times the passed float value, and increments the pointed-to int. Notice the capitalisation in the *ici_typecheck()* call to indicate a pointer to an int is required. Also note the hyphen in the *ici_retcheck()* call to indicate it should ignore the first argument.

Calling an ICI function from C

Calling an ICI function from C is fairly simple. There are a few slight variations on the same basic call. The simplest is the

Making new ICI primitive objects

Writting and compiling a dynamically loading extension module

The loaded library must contain a function of the following name and declaration:

```
object_t *
ici_var_library_init()
{
    ...
}
```

where *var* is the as yet undefined variable name. This is the initialisation function which is called when the library is loaded. This function should return an ICI object, or NULL on error, in which case the ICI error variable must be set. The returned object will be assigned to *var* as described above.

The following sample module, *mbox.c*, illistrates a typical form for a simple dynamically loaded ICI module (it is a Windows example, but should be clear anyway):

```
#include <windows.h>
#include <ici.h>

/*
 * mbox_msg => mbox.msg(string) from ICI
 *
 * Pops up a modal message box with the given string in it
 * and waits for the use to hit OK. Returns NULL.
 */
int
mbox_msg()
{
```

```
    char    *msg;

    if (typecheck("s", &msg))
      return 1;
    MessageBox(NULL, msg, (LPCTSTR)"ICI", MB_OK | MB_SETFOREGROUND);
    return ici_null_ret();
}

/*
 * Object stubs for our intrinsic functions.
 */
cfunc_t mbox_cfuncs[] =
{
    {CF_OBJ,"msg",mbox_msg},
    {CF_OBJ}
};

/*
 * ici_mbox_library_init
 *
 * Initialisation routine called on load of this module into the ICI
 * interpreter's address space. Creates and returns a struct which will
 * be assigned to "mbox". This struct contains references to our
 * intrinsic functions.
 */
object_t *
ici_mbox_library_init()
{
    struct_t*s;

    if ((s = new_struct()) == NULL)
      return NULL;
    if (ici_assign_cfuncs(s, mbox_cfuncs))
      return NULL;
    return objof(s);
}
```

The following simple Makefile illustrates forms suitable for compiling this module into a DLL under Windows. Note in particular the use of /export in the link line to make the function *ici_mbox_library_init* externally visible.

```
CFLAGS=
OBJS  = mbox.obj
LIBS  = ici4.lib user32.lib

icimbox.dll: $(OBJS)
            link /dll /out:$@ $(OBJS) /export:ici_mbox_library_init $(LIBS)
```

Note that there is no direct supprt for the /export option in the MS Developer Studio link settings panel, but it can be entered directly in the *Project Options* text box.

The following Makefile achieves an equivalent result under Solaris:

```
CC    = gcc -pipe -g
CFLAGS= -fpic -I..

OBJS  = mbox.o

icimbox.so  : $(OBJS)
            ld -o $@ -dc -dp $(OBJS)
```

Referring to ICI strings from C code

References to short strings that are known at compile time is common in ICI modules for field names and such-like. But ICI strings need to be looked up in the ICI atom pool to find the unique pointer for each particular one (and created if it does not already exist). To assist external modules in obtaining the pointer to names they need (especially when there are lots), some macros are defined in *ici.h* to assist. The following procedure can be used:

1. Make an include file called *icistr.h* with your strings, and what you want to call them by, formatted as in this example:

```
/*
 * Any strings listed in this file may be refered to
 * with ICIS(name) for a (string_t *), and ICISO(name)
 * for an (object_t *).
 *
 * This file is included with variying definitions
 * of ICI_STR() to declare, define, and initialise
 * the strings.
 */
ICI_STR(fred, "fred")
ICI_STR(jane, "jane")
ICI_STR(amp, "&")
```

2. Next, in one of your source files, after an include of *ici.h*, include the special include file *icistr-setup.h*. That is:

```
#include <icistr-setup.h>
```

   This include file both defines variables to hold pointer to the strings (based on the names you gave in *icistr.h*) and defines a function called *init_ici_str()* which initialises those pointers. It does this by including your *icistr.h* file twice, but under the influence of special defines for *ICI_STR()*.

3. Next, call *init_ici_str()* at startup or library load. It returns 1 on error, usual conventions. For example:

```
object_t *
ici_XXX_library_init(void)
{
    if (init_ici_str())
        return NULL;
    ...
```

4. Include your *icistr.h* file in any source files that accesses the named ICI strings. Access them with either *ICIS(fred)* or *ICISO(fred)* which return *string_t* * and *object_t** pointers respectively. For example:

```
#include "ici.h"
#include "icistr.h"

 ...

    o = ici_fetch(s, ICIS(fred));
```

Accessing ICI array objects from C

Using ICI independently from multiple threads

Summary of ICI's C API

The following table summarises function and public data that form ICI's C API. The use of some of these functions has been illustrated above. The full specification of each is given in a comment in the ICI source. This summary is only intended to document the limits of the public interface, allow you to select the relevant functions, and direct you to the source with the full description.

The final column gives a hint about upgrading from ICI 3. If just a name is mentioned, that is the old name of this function (many names have acquired an *ici_* prefix). Other notes indicate what constructs should be checked for possible upgrade to the given function. If it is blank, there is no change. There is an ICI program, *ici3check.ici*, in the ICI source directory that will grep your source for usage of changed constructs and print an note on upgrade action..

Name	Synopsis	Upgrade from ICI3
`array_t`	The ICI array object type. See *array.h*.	
`cfunc_t`	The ICI intrinsic function object types. See *cfunc.h* and *cfunco.c* (not *cfunc.c*).	
`debug_t`	A struct of function pointer for debug functions (like break and watch). See ###.	
`file_t`	The ICI file object type. See *file.h*.	
`float_t`	The ICI float object type. See *float.h*.	
`func_t`	The ICI function object type. See *func.h*.	
`handle_t`	The type of handle objects. See *handle.c*.	
`ici_alloc`	Allocate memory, in particular memory subject to collection (possibly indirectly). Must be freed with *ici_free*. See also *ici_talloc* and *ici_nalloc* which are both preferable. See *alloc.c*.	
`ici_argcount`	Generate an error message indicating that this intrinsic function was given the wrong number of argument. See *cfunc.c*.	
`ici_argerror`	Generate an errror message indicating that this argument of this intrinsic function is wrong. See *cfunc.c*.	

Name	Synopsis	Upgrade from ICI3
ici_array_gather	Copy a (possibly disjoint) run of elements of an array into contiguous memory.	Use of a_top.
ici_array_get	Fetch an elemet of an array indexed by a C int. See *array.c*.	Use of a_top.
ici_array_nels	Return the number of elements in an array. See *array.c*.	Use of a_top.
ici_array_new	Allocate a new ICI array. See *array.c*.	new_array
ici_array_pop	Pop and return last element of an ICI array. See *array.c*.	Use of a_top.
ici_array_push	Push an object onto an array. See *array.c*.	Use of a_top.
ici_array_rpop	Rpop an obect from an ICI array. See *array.c*.	
ici_array_rpush	Rpush an oject onto an ICI array. See *array.c*.	
ici_assign	Assign to an ICI object (vectors through type). See *object.h*.	assign
ici_assign_cfuncs	Assign of bunch of intrinsic function prototypes into the ICI namespace. See *cfunco.c*.	
ici_assign_fail	Generic function that can be used for types that don't support assignement. See *object.c*.	assign_simple
ici_atexit	Register a function to be called at *ici_uninit*. See *uninit.c*.	
ici_atom	Return the atomic form of an object. See *object.c*.	atom
ici_buf	A general purpose growable character buffer, typically used for error messages. See *ici_chkbuf*. See *buf.c*.	
ici_call	Call an ICI function from C by name. See *call.c*.	Prototype change.
ici_callv	Same as *ici_call* but takes a *va_list*. See *call.c*.	Prototype change.
ici_chkbuf	Veriry or grow *ici_buf* to be big enough. See *buf.c*.	
###copy_simple		
ici_cmp_unique	Generic function that can be used for types that can't be merged through the atom pool. See *object.c*.	
ici_debug	A pointer to the current debug functions. See ###.	
ici_debug_enabled	If compiled with debug, an int giving the current status of debug callbacks. Else defined to 0 by the preprocessor. See *fwd.h*.	

Name	Synopsis	Upgrade from ICI3
ici_decref	Decrement the reference count of an ICI object. See *object.h*.	decref
ici_dont_record_line_nums	A global int which can be set to prevent line number records (which will marginally speed execution, but erorrs won't reveal source location). See *fwd.h*.	
ici_enter	Acquire the global ICI mutex, which is required for access to ICI data and functions. See *thread.c*.	
ici_error	A global char pointer to any current error message.	
ici_exec	A pointer to the current ICI execution context (NB: Don't look at or touch x_os, x_xs or x_vs fields). See *exec.h*.	
ici_fetch	Fetch an element from an object. Vectors by object type. See *object.h*.	fetch
ici_fetch_fail	A generic function that can be used by objects that don't support fetching.	fetch_simple
ici_fetch_int	Fetch an int from an object into a C long. See *mkvar.c*.	
ici_fetch_num	Fetch an int or float from an object into a C double. See *mkvar.c*.	
ici_float_new	Get an ICI float from a C double. See *float.c*.	new_float
ici_float_ret	Return a C double from an intrinsic function as an ICI float. See *cfunc.c*.	float_ret
ici_free	Free memory allocated with *ici_alloc*. See also *ici_tfree* and *ici_nfree*. See *alloc.c*.	
ici_func	Call an ICI function, given you have a *func_t* *. See *ici_call* to call by name. See *call.c*.	Prototype change.
ici_funcv	Same as *ici_func* except it takes a *va_list*.	Prototype change.
ici_get_last_errno	Set *ici_error* (see) based on the last failed system function (i.e. *errno*). See *syserr.c*.	syserr (and semantics change)
ici_get_last_win32_error	Windows only. Set *ici_error* based on the value of *GetLastError()*. See *win32err.c*.	
ici_handle_new	Make a new *handle_t* object.	
ici_hash_unique	A generic function that can be used in a *type_t* struct for objects that can't be merged through the atom pool. See *object.c*.	
ici_incref	Increment reference to an ICI object.	incref
ici_init	Initialise the ICI interpreter. See *init.c*.	
ici_int_new	Get an ICI int from a C long. See *int.c*.	new_int
ici_int_ret	Return a C long from an intrinsic function.See *cfunc.c*.	int_ret

Name	Synopsis	Upgrade from ICI3
ici_leave	Unlock the global ICI mutex to allow thread switching. See *thread.c*.	
ici_main	A wrapper round *ici_init()* that does argc, argv argument processing. See *icimain.c*.	
ici_mark	Mark an object as part of the garbage collection mark phase. See *object.h*.	mark
ici_mem_new	Allocate a new *mem_t* object. See *mem.c*.	new_mem
ici_method	Call an ICI method given an instance and a method name. See *call.c*.	
ici_method_new	Allocate a new *metnod_t* object. See *method.c*.	ici_new_method
ici_nalloc	Allocate memory, in particular memory subject to collection (possibly indirectly). Must be freed with *ici_nfree*. See also *ici_talloc* and *ici_alloc*. See *alloc.c*.	
ici_need_stdin	Return ICI file object that is the current value of stdin..	need_stdin
ici_need_stdout	Return ICI file object that is the current value of stdout.	need_stdout
ici_nfree	Free memory allocated with *ici_nalloc*. See also *ici_tfree* and *ici_free*. See *alloc.c*.	
ici_null_ret	Return an ICI NULL from an intrinsic function.See *cfunc.c*.	null_ret
ici_objname	Get a short human readable representation of any object for diagnostics reports. See *cfunc.c*.	
ici_one	A global pointer to the ICI int 1.	o_one
ici_os	An ICI array which is the operand stack of the current execution context.	
ici_ptr_new	Allocate a new ICI *ptr* object. See *ptr.c*.	new_ptr
ici_reclaim	Run the ICI garbage collector. See *object.c*.	
ici_regexp_new	Make a new ICI regexp object.	new_regexp
ici_rego	Register a new object with the garbage collector. See *object.h* (a macro).	rego
ici_register_type	Register a new *type_t* structure with the interpreter to obtain the small int type code that must be placed in the header of ICI objects.	o_type assignment
ici_ret_no_decref	Return an ICI object from an intrinsic C function, without an *ici_decref*. See *cfunc.c*.	
ici_ret_with_decref	Return an ICI object from an intrinsic C function, but *ici_decref* it in the process. See *cfunc.c*.	

Name	Synopsis	Upgrade from ICI3
`ici_retcheck`	Check and update values returned through pointers.	
`ici_set_new`	Allocate a new ICI set object. See *set.c*.	new_set
`ici_set_unassign`	Remove an element from a set. See *set.c*.	unassign_set
`ici_set_val`	Set a C int, long, double, FILE * or ICI object into the inner-most scope of any object that supports a super. See *mkvar.c*.	
`ici_stk_push_chk`	Ensures there is a certain amount of contiguous room at the end of an array for direct push operations through a_top. See *array.h*.	
`ICI_STR`	Multi-purpose macro used by the *icistr-setup.h* mechanism. See *str.h*.	
`ici_str_get_nul_term`	Get the ICI form of a string of nul terminated chars without an extra reference count. See *string.c*.	get_cname
`ici_str_new`	Get the ICI form of a string of chars, by explicit length. See *string.c*.	new_name
`ici_str_new_nul_term`	Get the ICI form of a string of nul terminated chars. See *string.c*.	new_cname
`ici_str_ret`	Return a nul terminated C string from an intrinsic function as an ICI string. See *cfunc.c*.	str_ret
`ici_struct_new`	Allocate a new ICI struct object. See *struct.c*.	new_struct
`ici_struct_unassign`	Remove an element from an ICI struct. See *struct.c*.	unassign_struct
`ici_talloc`	Allocate memory, in particular memory subject to collection (possibly indirectly) sufficient for a given type. Must be freed with *ici_tfree*. See also *ici_nalloc* and *ici_alloc*. See *alloc.c*.	ici_alloc
`ici_tfree`	Free memory allocated with *ici_tfree*. See *alloc.c*.	
`ici_zero`	The ICI int 0.	o_zero
`ici_typecheck`	Check and marshall ICI arguments to an intrinsic function into C variables. See *cfunc.c*.	
`ici_uninit`	Shutdown and free resources associated with the ICI interpreter. See *uninit.c*.	
`ici_vs`	The scope ("variable") stack of the current ICI execution context.	
`ici_wakeup`	Wake up any ICI threads waiting on a given ICI object. See *thread.c*.	
`ici_xs`	The execution stack of the current ICI execution engine. See *exec.c*.	
`ici_`*`XXX`*`_library_init`	The entry point you must define for dynamically loaded modules.	

Name	Synopsis	Upgrade from ICI3
ici.h	The ICI C API include file. This is generated specifically for each platform.	
icistr.h	The include file you must make to get initialised ICI strings. See *str.h*.	
icistr-setup.h	The include file you must include to initialised ICI strings. See *str.h*.	
###objof		
int_t		
mem_t		
method_t		
null_t		
object_t		
objwsup_t		
ptr_t		
regexp_t		
set_t		
skt_t		
string_t		
struct_t		
type_t	The type that holds information about ICI primitive type. You must declare, initialise, and register one of these to make a new ICI primitive type. See *ici_register_type*. See *object.h*.	
wrap_t	Struct to support *ici_atexit*. See *uninit.c*.	

void *ici_alloc(size_t z)

Allocate a block of size *z* and return a pointer to it. Returns NULL on error, usual conventions. This just maps to a raw ANSI *malloc()* but triggers garbage collection as necessary and attempts to track memory usage to control when the garbage collector runs. Blocks allocated with this must be freed with *ici_free()*.

It is preferable to use *ici_talloc()*, or failing that, or *ici_nalloc()*, instead of this function. But both require that you can match the allocation by calling *ici_tfree()* or *ici_nalloc()* with the original type/size you passed in the allocation call. Those functions use dense fast free lists for small objects, and track memory usage better.

int ici_argcount(int n)

Set the error descriptor (*ici_error*) to a message like:

```
%d arguments given to %s, but it takes %d
```

and then returns 1.

This function may only be called from the implementation of an intrinsic function. It takes the number of actual argument and the function name from the current operand stack, which there-

fore should not have been distured (which is normal for intrincic functions). It takes the number of arguments the function should have been supplied with (or typically is) from *n*. This functionn is typically used from C coded functions that are not using *ici_typecheck()* to process arguments. For example, a function that just takes a single object as an argument might start:

```
static int
myfunc()
{
    object_t   *o;

    if (NARGS() != 1)
        return ici_argcount(1);
    o = ARG(0);
    . . .
```

int ici_register_type(type_t *t)

Register a new type_t structure and return a new small int type code to use in the header of objects of that type. The type_t * passed to this function is retained and assumed to remain valid indefinetly (it is normally a statically initialised structure).

Returns the new type code, or zero on error in which case ici_error has been set.

How it works

These are notes for a new chapter. Cover:

- Implementation of parser/compiler and execution engine using the common data structures.
- Operation of the execution engine.
- Logic behind objects semantics - single pointer, no special cases.
- Garbage collector.
- Lookup-lookaside.

CHAPTER 9

Obsolete features and mistakes

***OBSOLETE: Method* Calls** ###

In addition to the above ICI has a simple mechanism for calling *methods* — functions contained within an object (typically a *struct*) that accept that object as their first parameter. The method call mechanism is enabled via a modification to the *call* operator, "()", to add semantics for calling a pointer object and through the addition of a new operator, binary-@, to form a pointer object from an object and a key. ICI pointers, described below, consist of an object and a key. To indirect though the pointer the object is indexed by the key and the resulting object used as the result. This is the same operation used in dynamic dispatch in languages such as Smalltalk and Objective-C.

The call operator now accepts a pointer as its first operand (we may think of the call operator as a n-ary operator that takes a function or pointer object as its first operand the function parameters as the remaining operands). When a pointer is "called" the key is used to index the pointer's container object and the result, which must be a function object, is called. In addition the container object within the pointer is passed as an implicit first parameter to the function (thus passing the actual object used to invoke the method to the method). Apart from the calling semantics the functions used to implemented methods are in all respects normal ICI functions.

Struct objects are typically used as the "container" for objects used with methods. The super mechanism provides the hierarichal search needed to allow class objects to be shared by multiple instances and provide a natural means of encapsulating information.

A typical way of using methods is,

```
/*
 * Define a "class" object representing our
 * class and containing the class methods.
 */
static MyClass = [struct

    doubleX = [func (self)
    {
        return self.x * 2;
```

```
        }]

    ];

    ...

    static a;
    a = struct(@MyClass);
    a.x = 21;
    printf("%d\n", a@doubleX());
```

We first define a class by using a literal struct to contain our named methods. You could also define class variables in this struct as it is shared by all instances of that class. In our class we've got a single method, doubleX, that doubles the value of an instance variable called x.

Later in the program we create an instance of a MyClass object by making a new struct object and setting its super struct to the class struct. The super is made atomic which ensures all instances share the same object and makes it read-only for them. Then we create an "instance variable" within the object by assigning 21 to a.x and finally invoke the method. We do not pass any parameters to doubleX. The call through the pointer object formed by the binary-@ operator passes "a" implicitly

Mistakes

All too often in language design you realise you made and early mistake and it's too late to fix it. This is a place I can write them down. There are a lot more than are written here or course.

- Indexing a string should never have returned a one character sub-string. It should have returned an integer character code.
- The gsub and smash functions shouldn't have used \ as their escape character.

Symbols
  21
  at start of line 22
' 21
” 21

A
abs 79
acos 80
alloc 80
any 89
argc 102
argv 102
array 80
asin 80
assign 80
atan 80
atan2 80
audible bell 22
auto variable 23

B
back space 22
backslash 21, 22
build 80

C
calendar 81
call 82
carriage return 22
ceil 82
character-code 21
chdir 82
close 82, 83
cmp 83
comments 22
control character 22
copyright 9
cos 83
cputime 83
currentfile 83

D
debug 84
del 84
dir 85
double quote 21, 22

E
eof 85
eq 85
escap 22
eventloop 85
execution engine 21
exit 85
exp 85
explode 86
extern 23

F
fail 86
fetch 86
float 86, 92

floor 86
flush 86
fmod 86
fopen 86
form feed 22

G
getchar 87
getcwd 87
getfile 87
getline 87
gettoken 87
gettokens 87
gotolink calendar 77
gotolink cputime 77
gotolink debug 77
gotolink isa 78
gotolink load 78
gotolink new 78
gotolink now 78
gotolink profile 78
gotolink respondsto 79
gotolink rpop 79
gotolink rpush 79
gotolink signal 79
gotolink signam 79
gotolink version 79
gotolink wakeup 79
gotolinkt build 77
gsub 88

H
hex, character code 22

I
identifier, lexicon 22
implode 89
include 89
integer, lexicon 22
interval 89
isa 89
isatom 89

K
keys 90
keywords 22

L
lexical analyser 21
load 90
log 90
log10 90

M
mem 90
module 22
mopen 90

N
nels 90
new 91
newline 22
now 91

num 91

O
octal, character code 22

P
parse 91
parser 21, 23
pop 92
popen 92
printf 92
profile 93
push 94
put 94

Q
question mark 22

R
rand 94
reclaim 94
regexp 94
regexpi 95
regular-expression 21
remove 95
rename 95
respondsto 95
rpop 95
rpush 95

S
scope 22, 95
seek 96
set 96
signal 96
signam 96
sin 96
single quote 21, 22
sleep 96
smash 97
sopen 97
sort 98
sprintf 98
sqrt 99
static 22
string 21, 99
string-literal 22
struct 99
sub 99
super 100
syntax 23
  notation 23
system 100

T
tab 22
tan 100
thread 100
tochar 100
toint 100
tokens 21
top 100
typeof 101

V
variables 22
version 101
vertical tab 22
vstack 101

W
waitfor 102
wakeup 102
web site 9