IM 1003: Programming Design

Pointers

Ling-Chieh Kung

Department of Information Management
National Taiwan University

March 24, 2014

Outline

- **Basics of pointers**
- Call by reference/pointer
- Arrays and pointer arithmetic
- Dynamic memory allocation (DMA)

Pointers

- A **pointer** is a variable which stores a **memory address**.
  – An **array** variable is a pointer.
- To declare a pointer, use `*`.

```
type pointed* pointer name;
```

```
type pointed *pointer name;
```

- Examples:

```
int *ptrInt;
```

```
double* ptrDou;
```

  – These pointers will store addresses.
  – These pointers will point to `int`/`double` variables.
- We may point to **any** type.
- To point to different types, use different types of pointers.

Sizes of pointers

- All pointers have the same size.
  - In a 32-bit computer, a pointer is allocated 4 bytes.
  - In a 64-bit computer, a pointer is allocated 8 bytes.

```
int* p1 = 0;
cout << sizeof(p1) << endl; // 8
double* p2 = 0;
cout << sizeof(p2) << endl; // 8
```

- The length of pointers decides the maximum size of the memory space.
  - 32 bits: 2^{32} bytes = 4GB.
  - 64 bits: 2^{64} bytes = ?

Pointer assignment

- We use the **address-of operator `&`** to obtain a variable's address:

```
pointer name = &variable name
```

- The address-of operator `&` returns the (beginning) **address** of a variable.
- Example:
  - `int a = 5;`
    `int* ptr = &a;`
  - `ptr` points to `a`, i.e., `ptr` stores **the address of `a`**.
- When assigning an address, the two types must **match**.
  - `int a = 5;`
    `double* ptr = &a; // error!`

Variables in memory

- `int a = 5;`
- `double b = 10.5;`
- `int* aPtr = &a;`
- `double* bPtr = &b;`

Address	Identifier	Value
0x20c644	a	5
0x20c64c	bPtr	0x20c660
0x20c650		
0x20c658	aPtr	0x20c644
0x20c65c		
0x20c660	b	10.5
0x20c664		

Memory

Address operators

- There are two address operators.
  - `&`: The **address-of operator**. It returns a variable's address.
  - `*`: The **dereference operator**. It returns the pointed variable (not the value!).
- For `int a = 5`:
  - `a` equals 5.
  - `&a` returns an address (e.g., 0x22ff78).
- For `int* ptrA = &a`:
  - `ptrA` stores an address (e.g., 0x22ff78).
  - `&ptrA` returns the pointer's address (e.g., 0x21aa74). This has nothing to do with the pointed variable `a`.
  - `*ptrA` returns `a`, **the variable** pointed by the pointer.

Address operators

- Example:

```
int a = 10;
int* p1 = &a;
cout << "value of a = " << a << endl;
cout << "value of p1 = " << p1 << endl;
cout << "address of a = " << &a << endl;
cout << "address of p1 = " << &p1 << endl;
cout << "value of the variable pointed by p1 = " << *p1 << endl;
```

Address operators and `NULL`

- `&`: returns a variable's address.
  - We cannot use `&100`, `&(a++)` (which returns the value of `a`).
  - We can only perform `&` on a **variable**.
  - We cannot assign value to `&x` (`&x` is a value!).
  - We can get a usual variable's or a pointer variable's address.
- `*`: returns the pointed variable, **not** its value.
  - We can perform `*` on a pointer variable.
  - We cannot perform `*` on a usual variable.
  - We cannot change a variable's address. No operation can do this.
- A pointer pointing to nothing should be assigned `NULL` or `0`.

Address operators and NULL

- Examples:

```
int a = 10;
int* ptr = NULL;
ptr = &a;
cout << *ptr; // ?
*ptr = 5;
cout << a;    // ?
a = 18;
cout << *ptr; // ?
```

```
int a = 10;
int* ptr1 = NULL;
int* ptr2 = NULL;
ptr1 = ptr2 = &a;
cout << *ptr1; // ?
*ptr2 = 5;
cout << *ptr1; // ?
(*ptr1)++;
cout << a;     // ?
```

Address operators and NULL

- Dereferencing a null pointer shutdowns the program (a run-time error).

```
int* p2 = NULL;
cout << "value of p2 = " << p2 << endl;
cout << "address of p2 = " << &p2 << endl;
cout << "value of the variable pointed by p2 = " << *p2 << endl;
```

& and * cancel each other

- What is `*&x` if `x` is a variable?
  - `&x` is the address of `x`.
  - `*(&x)` is the variable stored in that address.
  - So `*(&x)` is `x`.
- What is `&*x` if `x` is a pointer?
  - If `x` is a pointer, `*x` is the variable stored at `x` (`x` stores an address!).
  - `&*x` is the address of `*x`, which is exactly `x`.
- What is `&*x` if `x` is not a pointer?

Good programming style

- Initialize a pointer variable as `0` or `NULL` if no initial value is available.
  - `0` is the standard in C++, while `NULL` is the standard in C. But they are the same for representing "null pointer".
  - By using `NULL`, everyone knows the variable must be a pointer, and you are not talking about a number or character.
- Without an initialization, a pointer points to **somewhere**... And we do not know where it is!
  - Accessing an unknown address results in unpredictable results.

Good programming style

- When we use `*` in **declaring** a pointer, that `*` is not a dereference operator.
  - It is just a special syntax for declaring a pointer variable.
- I prefer to treat **`int*`** as a type, which represents an "integer pointer".
- Therefore, I prefer "**`int* p`**" to "**`int *p`**".
- Be careful:

```
int* p, q;   // p is int*, q is int
int *p, *q;  // two pointers
int* p, *q;  // two pointers
int* p, * q; // two pointers
```

- I use multiple statements to declare multiple pointers.

Outline

- The basics of pointers
- **Call by reference/pointer**
- Arrays and pointer arithmetic
- Dynamic memory allocation (DMA)

References and pointers

- Recall this example:

```
void swap (int x, int y);
int main()
{
  int a = 10, b = 20;
  cout << a << " " << b << endl;
  swap(a, b);
  cout << a << " " << b << endl;
}
void swap (int x, int y)
{
  int temp = x;
  x = y;
  y = temp;
}
```

References and pointers

- When invoking a function and passing parameters, the default scheme is to "**call by value**" (or "pass by value").
  - The function declares its own local variables, using a copy of the arguments' values as initial values.
  - Thus we swapped the two local variables declared in the function, not the original two we want to swap.
- To solve this, we can use "**call by reference**" or "call by pointer."
  - They are somewhat different, but the principle is the same.
  - It is enough to know and use only one of them.

Call by reference

- A **reference** is a variable's alias.
- The reference is another variable that refers to the variable.
- Thus, using the reference is the same as using the variable.

```
int c = 10;
int& d = c; // declare d as c's reference
d = 20;
cout << c << endl; // 20
```

- `int& d = c` is to declare `d` as `c`'s reference.
  - This `&` is different from the `&` operator which returns a variable's address.
- `int& d = 10` is an error.
  - A literal cannot have an alias!

Call by reference

- Now we know how to change a parameter's value:
  - Instead of declaring a usual local variable as a parameter, declare a **reference** variable.
- This is to "call by reference".

```
void swap (int& x, int& y);
int main()
{
  int a = 10, b = 20;
  cout << a << " " << b << endl;
  swap(a, b);
  cout << a << " " << b << endl;
}
void swap (int& x, int& y)
{
  int temp = x;
  x = y;
  y = temp;
}
```

Call by reference

- Thus we can call by reference and modify our parameters' value.
- When calling by reference, the only thing you have to do is to add an `&` in the parameter declaration in the function header.
- Mostly people use references only in call by reference.
- View the `&` in declaration as a part of type.
- I use `int& a = b;` instead of `int &a = b;`.

Call by pointers

- To call by pointers:
  - Declare a **pointer** variable as a parameter.
  - Pass a pointer variable or an address returned by **&** when invoking.
- For the **swap()** example:

```
void swap(int* ptrA, int* ptrB)
{
  int temp = *ptrA;
  *ptrA = *ptrB;
  *ptrB = temp;
}
```

- Invocation becomes **swap(&a, &b);**

Call by pointers

- How about the following implementation?

```
void swap(int* ptrA, int* ptrB)
{
  int* temp = ptrA;
  ptrA = ptrB;
  ptrB = temp;
}
```

  - Invocation is still **`swap(&a, &b);`**
- Will the two arguments be swapped? What really happens?

Call by pointers

- The principle behind calling by reference and calling by pointer is the same.
- You can view calling by reference as a special tool made by using pointers.
- Do not mix references and pointers!
  - E.g., we cannot pass a pointer variable or an address to a reference!
- You can use calling by reference in most situations, and it is clearer and more convenient than calling by pointer.
  - When you just want to modify arguments or return several values, call by reference.
  - When you really have to do something by pointers, call by pointer.

Outline

- The basics of pointers
- Call by reference/pointer
- **Arrays and pointer arithmetic**
- Dynamic memory allocation (DMA)

Pointers and arrays

- An array variable **is** a pointer!
  - It records the address of the **first** element of the array.
- When passing an array, we pass a pointer.
- The array indexing operator `[]` indicates **offsetting**.
- To further understand this issue, let's study **pointer arithmetic**.

Pointer arithmetic

- Usually, one arbitrary address returned by performing arithmetic on a pointer variable is useless
- The arithmetic is useful (and should be used) only when you can predict a variable's address.
  - In particular, when variables are stored **consecutively**.

```
int a = 10;
int* ptr = &a;
cout << ptr++;
  // just an address
  // we don't know what's here
cout << *ptr;
  // dangerous!
```

Pointer Arithmetic: ++ and --

- The type a pointer pointing to is used as the unit of measurement.
- **++**: Increment the pointer variable's value by the number of bytes a variable in this type occupies (i.e., point to the **next** variable).
  - e.g., for integer pointers, the value (an address) increases by 4 (bytes).
- **--**: Decrement the pointer variable's value by the number of bytes a variable in this type occupies (i.e., point to the **previous** variable).

```
double a[3] = {10.5, 11.5, 12.5};
double* b = &a[0];
cout << *b << " " << b << endl;  // 10.5
b = b + 2;
cout << *b << " " << b << endl;  // 12.5
b--;
cout << *b << " " << b << endl;  // 11.5
```

Pointer Arithmetic: –

- We cannot add two address.
- However, we can find the difference of two addresses.

```
double a[3] = {10.5, 11.5, 12.5};
double* b = &a[0];
double* c = &a[2];
cout << c - b << endl; // 2, not 16!
```

Pointers and arrays

- Changing the value stored in a pointer is dangerous:

```
int y[3] = {1, 2, 3};
int* x = y;
for(int i = 0; i < 3; i++)
  cout << *(x + i) << " "; // 1 2 3
for(int i = 0; i < 3; i++)
  cout << *(x++) << " "; // 1 2 3
for(int i = 0; i < 3; i++)
  cout << *(x + i) << " "; // unpredictable
```

Indexing and pointer arithmetic

- The array indexing operator `[]` is just an **interface** for doing pointer arithmetic.

```
int x[3] = {1, 2, 3};
for(int i = 0; i < 3; i++)
  cout << x[i] << " "; // x[i] == *(x + i)
for(int i = 0; i < 3; i++)
  cout << *(x++) << " "; // error!
```

- Interface: a (typically safer and easier) way of completing a task.
  - `x[i]` and `*(x + i)` are identical.
  - But using the former is safer and easier.

Outline

- The basics of pointers
- Call by reference/pointer
- Arrays and pointer arithematic
- **Dynamic memory allocation (DMA)**

Static memory allocation

- In C/C++, we declare an array by specifying it's length as a constant variable or a literal.
  - `int a[100];`
- A memory space will be allocated to an array during the compilation time.
  - 400 bytes will be allocated for the above statement.
- This is called "**static memory allocation**".
- We may decide the length of an array "**dynamically**".
  - That is, during the **run** time.
- To do so, we must use a different syntax.
  - All types of variables may also be declared in this way.

Dynamic memory allocation

- The operator **`new`** allocates a memory space **and** returns the address.
  - In C, we use a different keyword **`melloc`**.
- **`new int;`** allocates 4 bytes without recording the address
- **`int* a = new int;`** makes **`a`** record the address of the space
- **`int* a = new int(5);`** makes the space contains 5 as the value
- **`int* a = new int[5];`** allocates 20 bytes (for 5 integers).
  - **`a`** points to the first integer.

Dynamic memory allocation

- All of these spaces are allocated during the **run time**.
- So we may write

```
int len = 0;
cin >> len;
int* a = new int[len];
```

- This allocates a space according to the input from users.

Dynamic memory allocation

- A space allocated during the run time has **no name**!
  - On the other hand, every space allocated during compilation has a name.
- To access a dynamically-allocated space, we use a **pointer** to store its address.

```
int len = 0;
cin >> len; // 3
int* a = new int[len];
for (int i = 0; i < 3; i++)
  a[i] = i + 1;
```

Address	Identifier	Value
0x20c644	N/A	1
0x20c648		2
0x20c64c		3
0x20c650		
0x20c654		
0x20c658	len	0
0x20c65c		
0x20c660	a	0x20c644
0x20c664		

Memory

Memory leak

- For spaces allocated during the **compilation** time, the system will **release these spaces** automatically when the corresponding variables no longer exist.

```
void func (int a)
{
  double b;
} // 4 + 8 bytes are released
```

- For spaces allocated during the **run** time, the system will **NOT** release these spaces unless it is asked to do so.
  – Because the space has no name!

```
void func()
{
  int* b = new int;
}
// 8 bytes for b are released
// 4 bytes for new int are not
```

Memory leak

- Programmers must keep a record for all spaced allocated dynamically.

```
double* b = new double;
*b = 5.2;
double c = 10.6;
b = &c; // now no one can access
        // the space containing 5.2
```

- This problem is called **memory leak**.
  – We lose the control of allocated spaces.
  – These spaces are **wasted**.
  – They will not be released unit the program ends.

Releasing spaces manually

- The **delete** operator will release a dynamically-allocated space.

```
int* a = new int;
delete a; // release 4 bytes
int* b = new int[5];
delete b; // release only 4 bytes!
          // Unpredictable results may happen
delete [] b; // release all 20 bytes
```

- The **delete** operator will do nothing to the pointer. To avoid reuse the released space, set the pointer to **NULL**.

```
int* a = new int;
delete a;   // a is still pointing to the address
a = NULL;   // now a points to nothing
int* b = new int[5];
delete [] b; // b is still pointing to the address
b = NULL;    // now b points to nothing
```

Good programming style

- Use DMA for arrays with **no predetermined** length.
- Whenever you write a `new` statement, add a `delete` statement below immediately (unless you know you really do not need it).
- Whenever you want to change the value of a pointer, check whether memory leak occurs.
- Whenever you write a `delete` statement, set the pointer to `NULL`.