

After this lesson, you should be able to understand the difference between static and dynamic data structures.

• Very simply, a particular arrangement of data • Why do we care? • Each arrangement allows you to do some things efficiently... and other things less efficiently • Solving a given problem efficiently requires • The right algorithm(s) + right data structure(s) • You'll learn more about this in algorithms

Data Structure is an arrangement of data in a computer's memory.

A specific way of data arrangement is necessary when building large scale applications on which even a microsecond difference of the run-time matters.

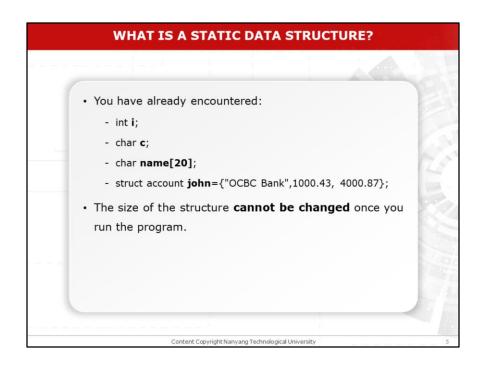
The developer must choose the appropriate data structure and algorithm for better performance.

We'll answer that question by looking at a "static" data structure Data storage elements in C whose structure can't be changed once you write your program Content Copyright Nanyang Technological University

What is a dynamic data structure? Before finding an answer to this question, it is better to understand what a static data structure is.

A static data structure is an organization or collection of data in memory that is fixed in size.

This results in the maximum size needing to be known in advance, as memory cannot be reallocated at run-time.



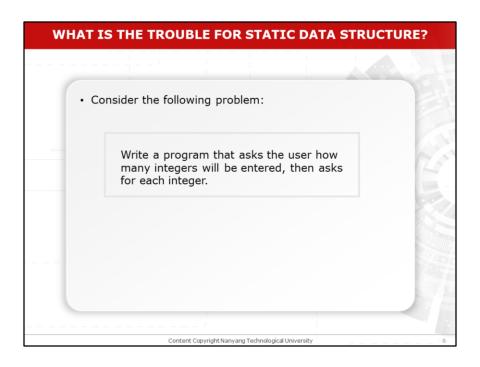
The static data structure is given a fixed area of memory which it can operate within. The content of such data structure can be modified but without changing the memory space allocated to it.

Arrays are a prominent example of a static data structure.

For example, after you define an array of 20 characters, you cannot change the size of the array at run-time.

Even for structs, you have to declare upfront, for instance, struct name and all the member variables inside.

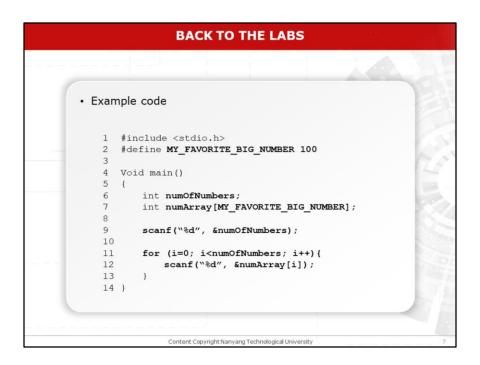
You can't change them once you have compiled the code. You have to make sure that everything is specified before compiling your code.



Static data structures are ideal for storing a fixed number of data items, but they lack the flexibility to consume additional memory if needed or free up memory when possible for improved efficiency.

Let's explain how to overcome such issues through a simple program. Consider the following question;

"Write a program that asks the user, how many integers will be entered and then asks for each integer."



Let's look at the sample code written for the given question.

The main problem that you would encounter when trying this question is that you would not be able to predict how many integers the user will enter.

Will it be 100 as suggested in the given code?

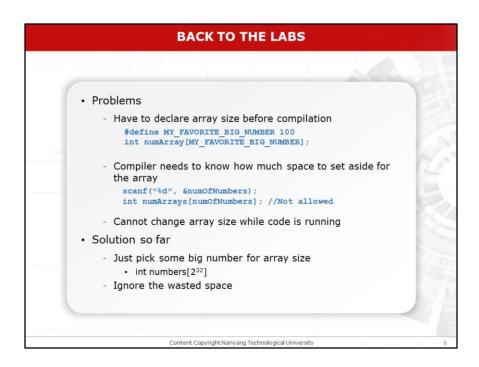
How do you handle when the user enter 200 or 500?

If you correctly declared the size of the array, it will be big enough to store all the integers and program will run smoothly.

But, if the array size is smaller than the number of integers that the user is going to enter, the program will overflow the bounds of the array.

The problem is that you need to declare the array size before compilation.

Because your compiler needs to know how much memory space to set aside for every single variable that you have before it can complete compiling your code.



Let's look at the sample code which is focused up to 100 elements.

If you declare an array of 100 integer elements, the memory will set aside space for 100 integer elements which are 400 bytes. After compilation, you cannot change the array size.

Therefore, the solution to this problem is to declare the array with a bigger number. You can declare a maximum size for an integer and forget about the wasted space.

Yet, this cannot be considered as a permanent solution since the memory wastage cannot be ignored easily.

• All declared at compile-time • If you want three structs, declare three separate structs with three unique names in code struct mystruct s1, s2, s3; struct mystruct s_arr[3]; • Note that even with the array declaration, each struct has a distinct "name" that you use to access it s_arr[0], s_arr[1], s_arr[2] • What if you want to declare more variables/arrays/structs when your code is already running?

The structures are also very important because structure creates a data type that can be used to group items of possibly different types into a single type.

Let's look at struct declaration more deeply.

Here, I have named 3 structs as s1, s2 and s3. Now, what if I want an extra struct, s4? I cannot declare a struct s4 while the program is running.

I can also declare an array of structs as well. Yet, declaring structs array for 3 elements would limit my program for an array with 3 structs.

That doesn't solve the problem because I'm still limited to a fixed-sized array, I can't grow that array or shrink the array. So we are still back to the same fundamental problem.

What if you want to declare more variables, arrays, or structs when your code is already running?

How to declare an array of integers dynamically?

In the line number five, it says that I need to declare an integer array of size number of numbers.

The idea is that once the user enters a random number, the compiler needs to set an array for the entered number.

The program can then allocate enough space to store the number of integers.

Therefore, you will have an array of integers that is exactly the right size to hold whatever the numbers that the user is going to enter.

Dynamic memory allocation can be done for integers, floats, characters or even structs.

Let's look at how to allocate memory for structs dynamically.

So I've declared the struct in line number one, that struct called my struct which has one member field. There is an integer called theinteger.

The function has written to store input values.

Every time, a new value comes in, the function requests space for it and stores it inside the requested memory space.

This process will continue until the user enters the sentinel value of -1.

Therefore, I can write the code in such a way that when I need more memory I can request space and when it is given, I can store values.

• Static vs. dynamic memory • Elements in "static" memory are allocated on the stack void main() { int i; char c; struct mystruct s_arr[3]; } • Elements in "dynamic" memory are allocated on the heap - This is what we will learn to do in a few more slides - You will be doing a lot of this with data structures

Let's discuss about computer memory layouts to understand the static and dynamic memory allocation.

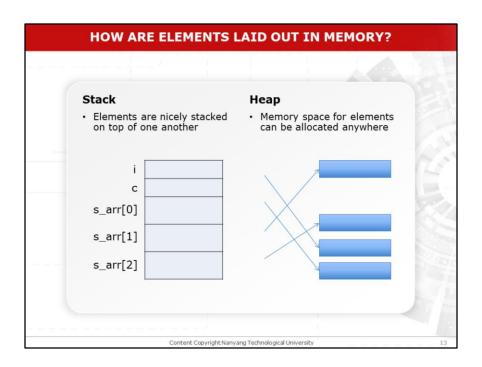
If we have 3 variables as **integer i**, **character c**, and a **struct array of 10 structs**, these variables will be stored in the part of the computer memory called stack.

Stack is used for static memory allocation.

Variables allocated on the stack are stored directly to the memory and access to this memory is very fast, and its allocation is dealt with when the program is compiled.

Variables allocated on the heap have their memory allocated at run time and accessing this memory is a bit slower.

Element of the heap have no dependencies with each other and can always be accessed randomly at any time.



As you can see, in the stack, the declared variables are stored in the respective order contiguously.

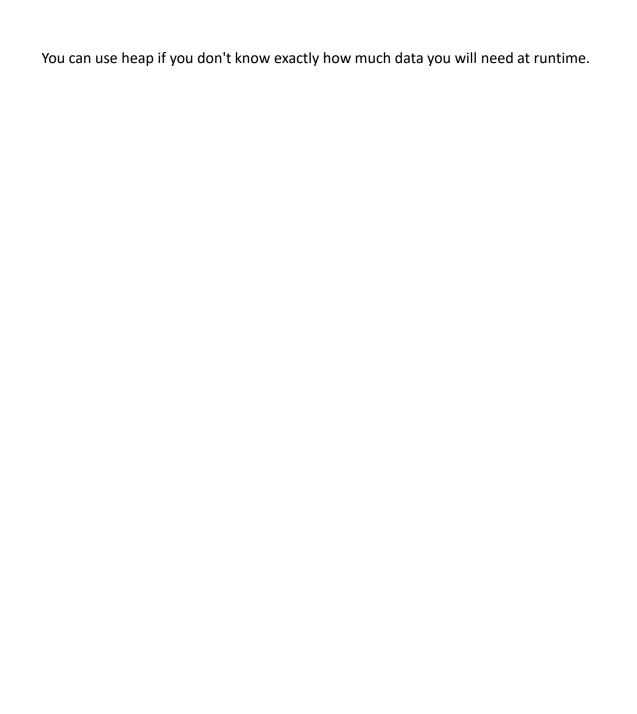
For integer i, the memory allocates 4bytes memory block.

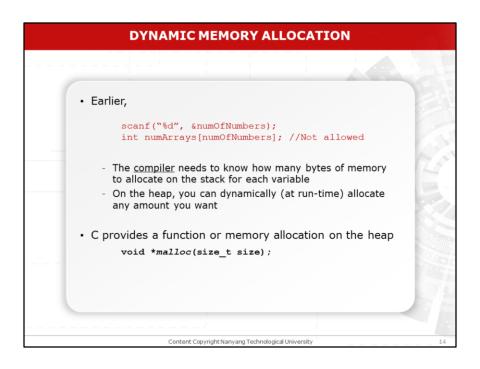
For character c, though it needs only 1byte, there is a fair chance that the stack will allocate 4bytes, and give access to only 1byte, because the computer memory tries to optimise certain things.

You can also see the first three elements of my struct array that I'd declared. So everything is very nicely stacked and nicely arranged.

In heap, you have no idea from where the memory has been allocated. Therefore, for each request that you make, the memory will be allocated from a randomly selected block.

You can use the stack if you know exactly how much data you need to allocate before compile time.





In C, the variable declaration should be done at the top of your code. For an example, you have to declare integer variable at the beginning of your code.

Therefore, to allocate memory in the heap only when needed, C introduced the utility function called 'malloc'.

'malloc' is the short form of memory allocation or memory allocator.

C provides a function or memory allocation on the heap void *malloc(size_t size); Reserves size bytes of memory for your variable/structure Returns the address (a pointer) where the reserved space starts Returns NULL if memory allocation fails

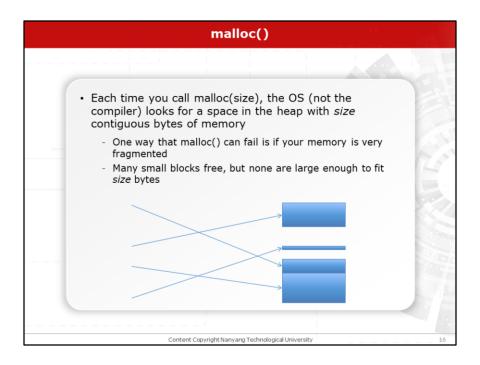
When the malloc() function is called with the required size, it returns a pointer to a newly allocated block *size* bytes long.

If the requested size of the memory is not available on the heap, the function returns a null pointer.

As you can see, the return type for malloc() is void*.

A void pointer is a pointer that has no associated data type with it. A void pointer can hold address of any type and can be typicasted to any type.

For instance, malloc() returns void* which can be typecasted to any type like int*, char*, etc.



When we request memory from the heap, it grabs whichever free space that is available.

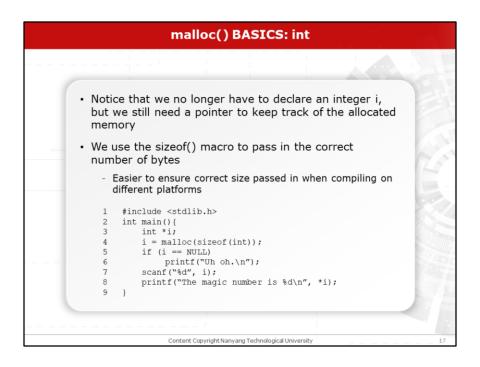
Why would malloc fail?

Let's assume that I have 1GB memory and 750MB is already allocated for different programs, then the heap will be failed to serve a request of 500MB.

If we consider another scenario and request 200MB which is available in the heap but in non-contiguous memory chunks as in 120MB memory block and 80MB memory block, then the malloc() is going to be failed because the memory chunks are fragmented.

Memory allocation failed for both scenarios hence malloc() will return NULL pointer.

Null pointer is a special reserved value of a **pointer**. A **pointer** of any type has such a reserved value. Formally, each specific **pointer** type (int * , char * etc.) has its own dedicated **null-pointer** value. Conceptually, when a **pointer** has that **null** value it is not pointing anywhere.



How to allocate memory for an integer using malloc()?

As you can see, instead of declaring integer i, we declare a pointer i, to keep track of the allocated memory. Once the malloc() function has been called, it allocates space for an integer and brings the address of the allocated memory block. Now we need to store that address because without that address we cannot reach to the allocated memory. Since we are going to store a memory address, we use a pointer (*i).

For an integer, we need 4bytes from the heap. Then why we pass **sizeof(int)**, instead of 4bytes. Because, on different platforms, such as mobile or embedded board, the size of an integer is less than 4bytes. Therefore, instead of passing the number, we pass **sizeof(int)**, so that the compiler allocates the correct value based on the platform.

Let's look at the coding example;

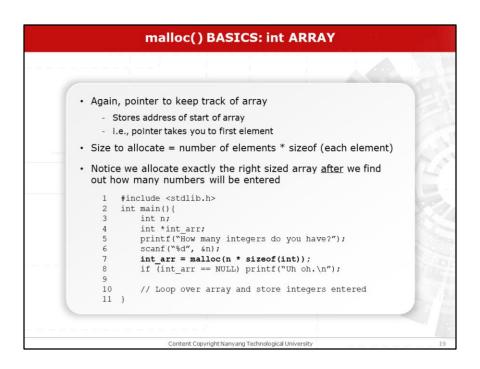
Line 3 is declaring integer * i pointer

Line 4 gets the allocated memory address and assign it to pointer i

Line 5 & 6 are for sanity check just in case if malloc() fails to perform

Line 7 does the scanf() to pass i, instead of ampersand i, to input the address of the variable to go directly to the variable in the memory, and change the value inside the memory.

```
malloc() BASICS: int
  #include <stdio.h>
   #include <stdlib.h>
3
4
  int main()
5
6
       int *i;
7
       i = malloc(sizeof(int));
       if (i == NULL)
8
           printf("Uh oh.\n");
9
10
       printf("Enter an integer: ");
11
12
       scanf("%d", i);
       printf("The magic number is %d\n", *i);
13
14
       return 0;
15 }
               Content Copyright Nanyang Technological University
```



When we allocate memory for an array of 10 integers, we need to pass 10*4bytes or 10*sizeof(int) because now we need space to store 10 integers.

Therefore, as in Line number 7, we pass **n*sizeof(int)** for n number of integers. And from here, you should be able to write the corresponding code from lines ten onwards, to make use of this array that you declared.

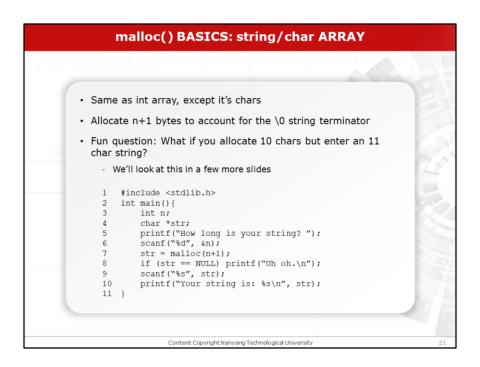
In line number four, you see that I have declared int array as an int star. Here, we declared integer array as a pointer (*int_array). We have to do this because the malloc() function returns the memory address which has to store as a pointer.

Yet, in C, arrays and pointers are interchangeable in a certain situation. When you pass an array into a function, you never pass by value; you never pass a copy of the array and its contents. You are always passing arrays into functions by reference.

I have the pointer to the starting address of the integer array, and I have declared int_array as int * int_array pointer.

Since I have the pointer to the starting address of the integer array, to get to the address of the next element of the array, I should move down 1 integer worth of bytes which is 4bytes. By moving down 4bytes each time, I can reach all the elements in the array.

malloc() BASICS: int ARRAY #include <stdio.h> #include <stdlib.h> 3 int main() 4 { 5 int n; int i; 6 int *int_arr; 7 8 9 printf("How many integers do you have? "); 10 scanf("%d", &n); 11 int_arr = malloc(n * sizeof(int)); 12 13 if (int_arr == NULL) printf("Uh oh.\n"); 14 15 for (i=0; i<n; i++) { printf("Enter an integer: "); 16 17 scanf("%d", &int_arr[i]); 18 19 return 0; 20 } Content Copyright Nanyang Technological University



The same process goes with string or character arrays.

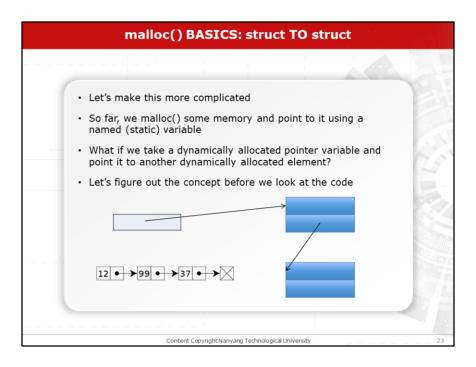
To allocate enough memory space for the string, we pass n+1 for malloc where n is the number of characters.

We do not have to pass sizeof(char) because the size of a character is 1byte.

Yet, for proper and complete code, you can pass sizeof(char), so that it compiles everywhere.

Remember that we use n+1 to leave a space for the slash zero at the end of the array of characters.

malloc() BASICS: string/char ARRAY #include <stdio.h> #include <stdlib.h> 2 3 int main() 5 6 int n; 7 char *str; 8 9 printf("How long is your string? "); scanf("%d", &n); 10 11 12 str = malloc(n+1); if (str == NULL) printf("Uh oh.\n"); 13 14 15 printf("Enter your string: "); scanf("%s", str); printf("Your string is: %s\n", str); 16 17 18 return 0; 19 } Content Copyright Nanyang Technological University

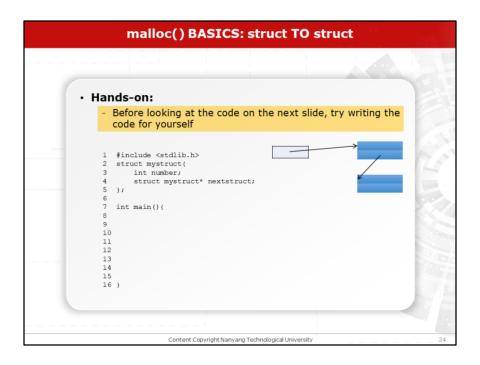


And now we can make things a bit more complicated. This is the start of what we are going to be doing for the next five weeks.

The only thing you start off with is that one tiny gray box on the left, which is one pointer variable. Using that one pointer variable, you will keep track of everything else. You can create new blue nodes, new structs, and you can link them to each other or delete blue chunks and so on.

Now, if you look at the given picture, the first blue block is a C struct that we are going to declare later on. As you can see, struct has two parts. The first part can be an integer variable. The second part of that blue chunk is another pointer.

So, I use malloc() to allocate memory for the first struct and the second variable inside the first struct is a pointer which establishes the link to the second dynamically allocated struct. And all of these will be linked up together.



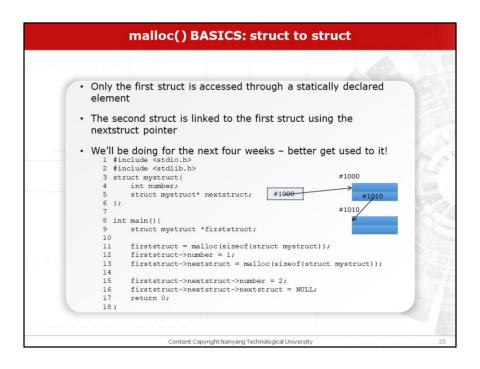
If you look at those two blue chunks over there, each one of those is supposed to be its own struct, an instance of that struct.

The struct is declared up there, struct my struct from line numbers two to five.

It has two members inside, the first member is an integer, so just the number that we are trying to store.

And the second part, which is important because it creates the links between different chunks of data, comes from that pointer.

The type has to be a struct my struct star. Because that link is going to take you to another instance of that struct, so it has to be a pointer to that same type.



Let's discuss the coding example in details:

In the code, I have two pointer variables. They correspond to the arrows that I've drawn out on the screen.

The first pointer is **firststruct** which is the pointer to the **struct mystruct**. This pointer has been declared in line 8 is statistically allocated. Therefore, it comes from the stack portion of the memory. Therefore, once you compile your code, it knows that it needs to allocate 4bytes of memory to hold the address, the pointer variable. That takes me to my very first structure.

After that, from line 10 onwards, everything is being dynamically created. Line number 10 says that firststruct equals to malloc size of struct mystruct, that is the point where it requests enough memory from the heap to store that structure, to store those two member variables inside.

Once malloc() returns with the address in the heap, where you've been given a certain amount of memory to store that structure. And that address is going to be stored into the first struct pointer variable, and that's going to create your very first link. As you can see in the diagram, the firststruct is at #1000, and that value #1000 gets stored into the firststruct pointer variable.

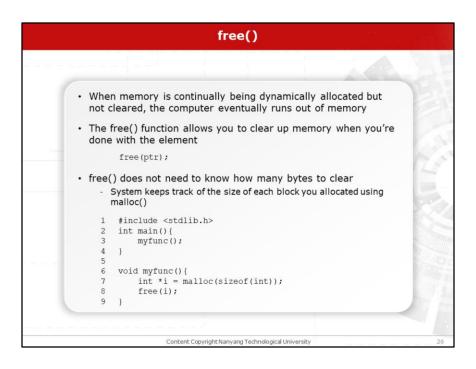
Now we need to connect the first dynamically allocated struct to the second structure. To do so, we need to call malloc() function again. But now look at where this new address gets stored into. It's now being stored into, first struct,

arrow, next struct as you can see in line number twelve.

Let me explain again. I have a pointer variable inside that first struct that I'm accessing. I'm setting the address of my second structure from the heap, at address one zero one zero. I'm storing that address into the next struct pointer, which is inside the first struct.

For instance, once the malloc() returns with the memory address which is #1010 then, we will store that value to the nextstruct pointer variable which is inside the firststruct.

So basically, we can keep adding new structs. The basic idea here is that we can access every struct because they are all linked together by this chain of pointers. And that's a fundamental idea behind a linked list.



When memory is continually being dynamically allocated but not cleared, the computer eventually runs out of memory.

Using malloc() function, you can request a certain number of bytes, and it returns with the address of the allocated memory block.

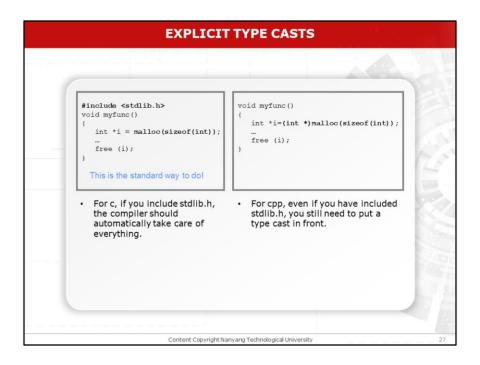
Using free() function, you can pass the address given by malloc() function and request to clear the memory block which was allocated using malloc() function.

You do not have to pass the number of bytes that need to be cleared because the operating system keeps track of those data.

So now look at this updated function, which is written from line numbers six to eight. In line number seven, malloc() function requests four bytes stores the address where those four bytes start into pointer variable i.

In line number eight, it returns allocated memory back to the system using free() function.

I call malloc() function, so I request four bytes of memory, and then I call free() functions, so I give it back to the system. The total amount of memory that I've held on to is going to remain constant, at any time it's at most four bytes.

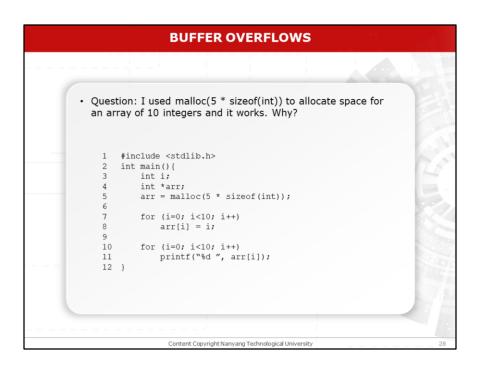


Never forget to include stdlib.h when you write standard C code. If you are writing standard C code, use standard C syntax and not C++ syntax.

For C language, if you include stdlib.h, the compiler should automatically take care of rest of the stuff.

For C++, even if you have included stdlib.h, you still need to put a type cast in front.

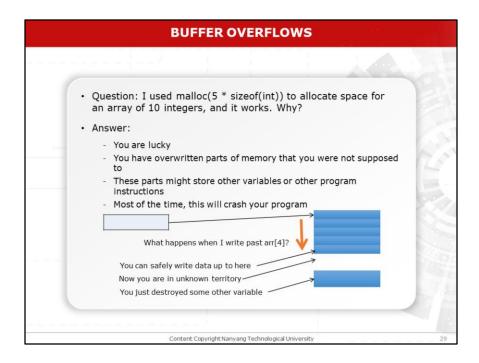
If you explicitly introduce a type cast, the compiler will assume you know what you're doing, even if you make a mistake with the pointer type.



Now, this question is quite interesting because it tests you on your understanding of arrays, how much space you have, and what happens when you start to overflow the bounds of an array.

As you have learnt, when I pass malloc (5 * sizeof(int)), you know that malloc() will allocate enough space for an array of five integers.

Then, by using line numbers 6 and 7, I am trying to store 10 integers, and in line numbers 9 and 10, I am trying to print the numbers I have assigned. What will this piece of code print?



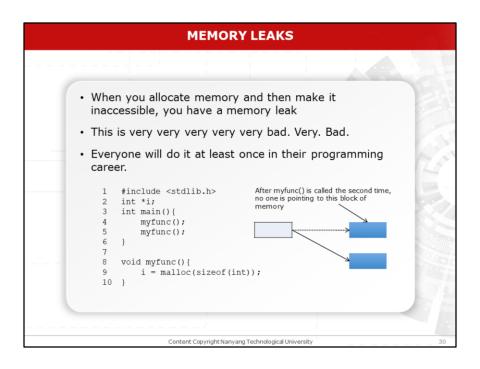
And if you look at the picture over here, we are looking at the heap on the right.

As you can see that I have five spaces, each one representing four bytes, one integer's worth of memory, that I can write integers into.

Now, as I'm writing into index zero one two three four, I'm writing into space that actually has been reserved for me. I'm not using somebody else's variable.

But what happens when I start going to array index five, six, seven, eight, and nine?

What am I actually doing? I can still keep going; I can still keep writing to memory. But at some point, I have no idea whether that memory will stay unused. I might be writing into some segment of memory that is already reserved for someone else's variable. I might actually end up destroying other values that were stored there previously. Hence this will end up with unexpected results.



When you call malloc() function and forget to call free, that block of memory has already been allocated to you; the system is holding on to it, it assumes that you're making use of it.

Once you finished the function, you will lose the address of the memory block which you previously had.

That memory address is the only way that you can find a way to the block of memory that was given to you. Once you lose that value, that's it; you can't get back to it. This is called memory leak, and it will affect the system very badly.