UMalloc

This project implements a simulation of the malloc and free functions included in `<stdlib.h>`. These implementations of malloc and free can be used as replacements to the actual malloc and free defined in `<stdlib.h>`. Please refer to the PDF for more details on the directions of this assignment.

Authors

- Hasnain Ali (ha430)
- Rushabh Patel (rsp155)
- Della Maret (dm1379)

Pre-Requisites

This library (and it's corresponding programs) were only intended to be executed on `x86` and `x86_64` architectures. This library and programs assume that the hardware will be consistent with that of `x86` and `x86_64` architecture. This requires the hardware to use 2s complement and be consistent with Little Endian endianness.

Project Structure

- There is one directory: `src`
    - `src` includes all `memgrind.c` and `umalloc`'s necessary files. This is where our implementation of `malloc` lies. `memgrind.c` includes all the tests from the write-up.
        - To run `memgrind`, execute the following in the `src` directory:

        -
          ```
          make
          ./memgrind
          ```

Test Plan

Our test plan includes a series of stress tests that will trigger almost all the error checking we did in our program. `errors.h` contains function names for the errors we are checking for. They are listed below.

Errors

```
void doubleFree(char* file, int line);
void wrongPointer(char* file, int line);
void tooMuchMem(int MEMSIZE, char* file, int line, size_t structSize);
void noMoreMem(char* file, int line);
void mallocZeroError(char *file, int line);
void nullPointerPassed(char *file, int line);
```

```
void notEnoughFreeMemoryForAllocation(char *file, int line);
void enoughFreeMemoryButNoBlockLargeEnough(char *file, int line);
```

- `void doubleFree(char* file, int line);`
  - This error is called in the event that the user is trying to free a pointer they have already freed or if they are trying to free something that may not have been allocated yet and is marked as `AVAILABLE` in the metaData, this error will be triggered with an error message on the screen.
- `void wrongPointer(char* file, int line);`
  - This error is called in the event that the user passes any pointer other than the original one was that given to malloc. In this event, this error will be called and free will fail.
- `void tooMuchMem(int MEMSIZE, char* file, int line);`
  - This error will be called in the event the user tries to allocate too much mem. Say that `MEMSIZE` is 4096. If the user tries to allocate more than 4080 bytes (because malloc needs to store two instances of metaData), then this error will be called indicating that too much mem was requested. `NULL` will be returned, along with an error message.
- `void noMoreMem(char* file, int line);`
  - Contrary to `tooMuchMem`, this function will be called in the event that malloc cannot find a big enough block of space to request the data requested by the user. In this case, this error message will be printed out and `NULL` will be returned.
- `void mallocZeroError(char *file, int line);`
  - Contrary to the actual `malloc` defined in `<stdlib.h>`, in our version of malloc, the user cannot allocate zero bytes (for safety reasons). This dramatically improves the safety of the client's program by preventing them from allocating zero bytes. In such an event, `NULL` will be returned.
- `void nullPointerPassed(char *file, int line);`
  - This means that a `NULL` pointer was passed into the `free` function. This is illegal and the program will crash in this case with a return status of `EXIT_FAILURE`.
- `void notEnoughFreeMemoryForAllocation(char *file, int line);`
  - This error means that there the mem is not full, but there is not enough space for the requested allocation.
- `void enoughFreeMemoryButNoBlockLargeEnough(char *file, int line);`
  - This means that there is enough raw available space for the requested mem, but no large enough block for the requested mem.

Properties Our Library Must Have To Be Correct

In order for the library to be correct, we concluded that the library must have the ability to `malloc`, `free`, and coalesce blocks. The library must also have the ability to call out errors were detailed in the errors section here

How to Check That Our Code Has These Properties

In order to check that our code has these properties, we intend to use various forms of testing. In the `memgrind.c` file, we intend to place the five stress tests from the write-up that was given and add one more stress tests of our own. Furthermore, as we test, we are using a function that prints out our mem array so, we can manually check whether our library is correct.

- `void printMemory(int bytes);`
  - This function will print `bytes` number of `bytes` from the mem array. Each address in the array will be printed, along with the value that is stored in that specific element of the mem array. Please note, since this program is only intended to be ran on `x86` or `x86_64` architecture, the bytes will be printed using 2s compliment with Little Endian endianness.

Memgrind.c Tests ~ Basic Integration / Test Plan

In order to fully test our program we were required to write an additional program `memgrind.c`, the tests that we wrote in `memgrind` were provided in the assignment instructions.

0. Consistency
   - First a small block (1B to 10B) was allocated, it was cast to a type, written to and then freed. Next a block of the same size was allocated, it was cast to a type, written to and then freed. Pointers to both blocks were compared to see if they were the same.
1. Maximization: (simple coalescence)
   - Begin by allocating a 1B block, if the result from malloc is not NULL then free it, double the size of the block and try again. Once malloc returns null, halve the block size and try allocating again. Finally when malloc returns NULL and the size of the block is 1B or 0B then stop and free all the mem. From this the maximal allocation was determined
2. Basic Coalescence
   - Begin by allocating one half of the maximal allocation found from Test 1, then allocate one quarter of the maximal allocation. Next free the first pointer, and then the second pointer. Finally try to allocate the maximal allocation.
3. Saturation
   - Begin by doing 9216 1KB block allocations, then switch to 1B block allocations until malloc returns NULL. After malloc returns NULL, the mem has been successfully saturated.
4. Time Overhead
   - Begin by saturating your mem (continue from Test 3) and then free the last 1B block. Next get the current time (start time), allocate a 1B block and then get the current time again (stop time). Take the difference between the stop and start time, this is your max time overhead.
5. Intermediate Coalescence
   - Begin by saturating your mem (continue from Test 4) and free each block allocation one by one. After all the blocks have been freed, attempt to allocate the maximal allocation (from Test 1). Finally free all mem.
6. Complete Saturation:
   - This test simply allocates the entire mem with 1 byte `char` allocations. We then call `freeAllFast` which frees the entire mem in constant time (we just set the first metadata to free the entire mem).
   - Once the entire mem is full, we free the entire mem.
   - By default, this test is disabled. To run it, please compile using `make long_test` and then run `./memgrind`.

Design Properties

Our library includes several unique and interesting design properties. All design properties for `memgrind.c` were referenced above in Tests. This section will strictly focus on `umalloc.c` and how we actually implemented our versions of `malloc` and `free`.

`umalloc` will allow the user to call `malloc()` and `free()` seamlessly, just as if they had used the `<stdlib.h>` definition of `malloc` and `free`.

Our library also has a cool feature where it can `freeAll` the entire array. This will prevent mem leaks and will provide a hard reset on the entire mem array in case something goes wrong. C does not have such a feature built into any of its standard libraries. This function will also go through the entire array and coalesce all blocks.

We have also included a `freeAllFast()` function which will set the very first metadata in the mem to free, and it will have a data size of the entire mem, thus giving user access to the entire mem again in O(1) time. We effectively set the first metadata to free, freeing the entirety of mem.

Documentation and Design Properties of `umalloc.h`

`umalloc.h` provides function prototypes for these functions that are later defined in `umalloc.c` These prototypes can be seen below:

```
void *umalloc(size_t size, char *file, int line);
void ufree(void *ptr, char *file, int line);
void freeAll();
void printMemory(int bytes);
void freeAllFast();
```

`umalloc.c` also defines a function `void *initializeMemory(size_t size)` and `void coalesceBlocks()` which will be elaborated on later.

- `void *initializeMemory(size_t size)`
    - This function will initialize the publicly defined `mem` array with two sets of metaData. One for information about the requested size from `umalloc` and another that will contain information about the remaining memory in the array (this excludes space for `metaData`, consistent with the methodology in `umalloc`). That is, the first metaData will contain information about the requested size allocation from `umalloc`. Then, to the right of the allocated data, a new, `available` `metaData` is stored. This will contain information about how much mem is left over. This metaData will ensure that it subtracts the space `metaData` will take on the next call to `malloc`. This means, it stores the amount of mem that can actually be allocated by the user. It subtracts space for all the `metaData` overhead. This means, the information stored to the right of the metaData will be the amount of mem that is actually left over to be allocated.
- `void *umalloc(size_t size, char *file, int line);`
    - This function will replace all calls in the client code to `malloc` assuming they have `umalloc.h` included in their files.
    - This function will first check to if the requested mem is greater than `MEMSIZE`. By default, `MEMSIZE` is 10MB.
        - Please note: The requested size must be less than or equal to `MEMSIZE - 16`. This is because of `umalloc` to function correctly it requires that there be enough space to store one `metaData` that will store information about the requested data, and there be

        enough space to store a second `metaData` AFTER the allocated space to store information about how much space is left over. This `metaData` is set to be available.
        - Also, there may be a case where the user deallocates all their mem (and then by default, all the blocks will be coalesced). In this case, the metaData that will be left over will contain information about how much space is left, accounting for one more `metaData` allocation and the current `metaData` allocation. This means, that in this case, where all the blocks have been deallocated and coalesced, the mem that will be available to the user will still be `MEMSIZE - 16`.
    - Contrary to the `malloc` defined in `<stdlib.h>`, this version of `malloc` will not allow the user to `malloc` 0 bytes. In this case, `malloc` will fail and return `mallocZeroError`. This will provide more safety and security compared to actual `malloc` defined in `<stdlib>`.
    - If the memory has not been initialized, `umalloc` will initialize it. Please see the `initializeMemory` section for more information.
    - Then, we iterate through the memory array. If we find a block with the right size, we allocate that block. We then check to see if there `metaData` stored to the right of the block we just allocated. If there is no metaData there, we check to see if we are in the bounds of the `mem` array, and if we are, then create a new metaData to the right of the block we previously allocated. This means we are at the right most part of the array where there is no `metaData` yet allocated. We allocate this new `metaData` setting it to `available` by default. We also ensure that we store how much mem is left to be called by the client code. This ensures that we have enough space for the next metaData to be stored, hence why we subtract the sizes of the `metaData` storing from the amount of mem left available to allocate. THIS MEANS that the amount data size left is stored in this `metaData`. In other words, the amount that is stored in this `metaData` is how much the user has left to allocate, accounting for the `metaData` overhead.
        - If there is not enough space for the second metaData stored to the right of the allocated block, then we return `noMoreMem` and return `NULL`.
    - Assuming the above executed without running into any of the described edge cases, we return the address of the allocated chunk back to the client.
        - Note: If we were not able to find an available block in the entire mem array while iterating through each `metaData`, then we return the corresponding error message depending on the context and nature of the user's call.
    - Additionally, if there is a block bigger than what the user requested and there is used `metaData` on the right, but we do not have space to store another `metaData` to the right of this block, then we will simply give the user a bigger block than that of which they requested. If there is space to store an additional `metaData`, then we will break the current block up and store the additional `metaData`.
    - If there is not enough space to store `metaData`, the malloc call will fail.

- `void ufree(void *ptr, char *file, int line);`

    - This function will take a pointer to the BEGINNING of an allocated chunk of mem that was returned by `umalloc`.
        - If `*ptr` is `NULL`, then we print a `nullPointerPassed` error and `EXIT_FAILURE`
        - If the above does not happen, then we take the pointer given and do pointer arithmetic to get the metaData that *should* be stored right behind the pointer that the client passed.

        - If the `metaData` stored at this address is set to `available`, then this means that the user is either trying to double free this pointer OR the user is trying to free a block that has set to `available` already.
    - If the above edge cases fail, then this means the pointer that was passed is valid. In this case, we set that blocks `metaData` to `available` and then we call `coalesceBlocks()`.

- `void coalesceBlocks()`
    - This function loops through the entire mem array looking for two adjacent blocks and combines them into one block (while maintaining the metaData overhead property explained in the `umalloc` function documentation).
    - We store the first `metaData` in the array, and we also find the location of the next metaData in the mem array.
    - If the next `metaData` is within the bounds of the mem array, we can continue. If not, we return because we have reached the end of the array.
    - If the first `metaData` that we stored, and the second `metaData` that we stored are both set to be `available`, then we set the first `metaData` to have enough space for the block that it previously allocated, the size of the second `metaData`'s allocation, and the size of that `metaData`.
        - This means that if the first `metaData` had a four byte allocation, and the second `metaData` also had a 4 byte allocation, then, the coalesced `metaData` (the first one), will store the following size that is allowed to be allocated once again my `malloc`: `firstMetaDataAllocationSize + secondMetaDataAllocationSize + sizeof(secondMetaDataSize)`.
            - Please see the `coalesceBlocks()` function in `umalloc.c` for more information.
        - If we were not able to find two adjacent free blocks, we iterate over to the next metaData.

- `void freeAll();`
    - This function loop through the entire array, looking for any allocated blocks and calling `free` on those blocks manually.
    - After looping through the entire `mem` array, it will call one more `coalesceBlocks` to make sure that everything is entire deallocated and coalesced.

- `void freeAllFast()`
    - Will free the entire memory array in constant time. We do this by setting the first `metaData` to indicate that the rest of the memory is free (excluding metadata).

- `void printMemory(int bytes);`
    - This function was defined and explained HERE. Please refer there.

Design Notes

There are several design notes we believe are worth pointing out. Additionally, we took some extra measures to improve the functionality of our `malloc` compared to the one defined in `<stdlib.h>`. Additionally, here is where we explain what `gcc` flags we are using in the compilation process and why we are using them. Here is where we will just explain some of the decisions we made and why we happened to make them.

- First off, in our `umalloc()` implementation, we are always storing two instances of `metaData` in the mem array.
    - The first `metaData` marks the block allocated next to it as being `used`. This first `metaData` stores information about the size of the block that the client has allocated.
    - Then, after the first block is allocated, we add a second `metaData` (assuming there is not one there already. If there is, we don't do anything. (Please read the `umalloc` function documentation in Design Properties for more information.))
    - This second `metaData` is marked as `free` and ready to be allocated. This `metaData` stores how much more mem can actually be allocated by the user. In other words, this is the left-over limit that malloc is allowed to call and nothing else. This means that the leftover data that is allowed to be allocated accounts for the current metaData that it is stored in, and the creation of a new metaData.
        - In our implementation, there **ALWAYS** must be a `metaData` stored to the right of any allocated block. Even when the entire mem has been freed (and is currently initialized), there will still always be a metaData left telling the user how much more mem they are allowed to allocate.
    - Thus, when we are allocating the second `metaData` to store how much data is left over to be allocated, we are storing how much mem is left, minus the size of the current `metaData`, minus the size of the next `metaData` that would be allocated in the event another `malloc` call is made. Thus, this is why the `metaData` containing the information about how much mem is left may be less than `MEMSIZE - (what's already allocated)`. It is because there is a `metaData` overhead that we need to consider for first. Thus, anywhere in this document where we mention `metaData` overhead, this is what we mean. We also explain this in our documentation of `umalloc` here.
    - Additionally, if there is a block bigger than what the user requested and there is used `metaData` on the right, but we do not have space to store another `metaData` to the right of this block, then we will simply give the user a bigger block than that of which they requested. If there is space to store an additional `metaData`, then we will break the current block up and store the additional `metaData`.
    - If there is not enough space to store `metaData`, the malloc call will fail.
- Secondly, we included a `freeAll()` function in `umalloc.h` and in `umalloc.c`. This function will free all allocated blocks of mem and coalesce them all together. The `malloc` function defined in `<stdlib.h>` does not include such a function. This includes a "HARD RESET" button on all allocated blocks of mem in case something goes wrong in the client's code. This is meant to be a debugging tool for the client, but it can also be a "lazy way" to deallocate mem if needed.
    - Additionally, this function provides more safety to the client code by having a way to essentially avoid any and all mem leaks they may cause (as long as they are not messing up the `metaData`). This is a key advantage over the `<stdlib.h>`'s implementation of mem management.
- Thirdly, we give the client's code access read-only access to `x` bytes of the array up to `MEMSIZE` in case they want low-level access to their allocations. This is meant for debugging purposes, but `<stdlib.h>`'s implementation of `malloc` does not account for this or provide such access.
- Fourthly, alignment. Because we wanted to not let any mem go to waste, and we wanted to provide the most mem possible (since we were told that this code would only ever run on Intel `x86` or Intel

`x86_64` architectures), we decided not to include any type of alignment to allow for the most amount of mem to be used. Thus, this means that this code is only able to work on any kind of `x86` or `x86_64` architecture. However, based on our code's modularity, this functionality would be easy to add for any kind of data alignment.
  - That is, it is very easy to make sure on our iterations that we are always iterating to `n`-byte aligned addresses to guarantee that our mem is aligned.
  - However, since our use case with this code given the architecture that is being used, we decided to maximize space rather than alignment as that would provide better results in this niche use case.

C Flags That Are Used Throughout This Project

Throughout this project, we use different C Flags when compiling. The custom `tests` that we wrote uses their own C Flags (because some of them) trigger and error on purpose. `memgrind` is compiled using its flags. Here, we explain why we used each one.

- `memgrind` compile flags:
  - `-g`: provides debugging information in case the use for a debugger like `gdb` is needed. It can provide more debugging information.
  - `-std=c99`: This is C standard that we are required to use for this class.
  - `-Wall`: Enable all major warnings.
  - `-Wvla`: Warn if variable length arrays are used. This is to avoid tampering with the `mem` array or any strange behaviors in the client code.
  - `-Werror`: Treat all warnings as errors.
  - `-Wundef`: Warn if any undefined behavior occurs.
  - `-Wpointer-arith`: Warn if any invalid pointer or risky arithmetic occurs.
  - `-O2`: Enable second level optimizations. This will provide a good balance between performance and compile time.
  - `-fsanitize=address`: We include this because this program involves a lot of pointer and address arithmetic. We include this to make sure nothing goes wrong at compile time.