ECE 353: Systems Software

Lecture 36

Memory Mapping

1.0.0

Jon Eyolfson

April 5, 2023

A 30B LLM on with Only 6.8 GB of RAM, how?

LLaMA is Meta’s large language model, and this C++ implementation is efficient

It’s so efficient people are discussing how this is possible!
`https://github.com/ggerganov/llama.cpp/discussions/638`

We Can Control Our Processes’ Virtual Memory

Memory map, or `mmap` is used to map files to a processes’ virtual address space

The pointer (virtual address) returned will allow you to access the file directly

- There’s no need for `read` and `write` system calls

Let’s See a `mmap` Example

Check out `36-mmap` in the `examples` repository

The `mmap` API

`mmap` takes 6 arguments:

1. **`void *addr`**: suggested starting address (`NULL` means you don’t care)
2. **`size_t length`**: number of bytes to map
3. **`int prot`**: protection flags (read/write/execute)
4. **`int flags`**: mapping flags (shared/private/anonymous)
   anonymous means the mapping isn’t backed by a file
5. **`int fd`**: file descriptor to map (ignored for anonymous)
6. **`off_t offset`**: offset to start the mapping (must be a multiple of page size)

`mmap` Is Lazy

It just sets up the page tables, it doesn't actually read from the file

It would create an invalid PTE during the `mmap` call

The kernel uses the remaining bits of the PTE for book keeping
- Which disk block is associated with this entry

The first access to the page would generate a page fault
- The kernel would then read from disk into memory

This ensures only the used parts of the file get read

Back to the Question

`https://github.com/ggerganov/llama.cpp/discussions/638`

How does an approximately 20 GB file only use 6.8 GB of real memory?

Hint: when you do model inference, the models are sparse (you don’t use all of it)

How Much Space Would the Kernel Need for Page Tables?

Someone posted you'd need 40 MB of page tables:

```
(20*(1024*1024*1024)/4096*8) / (1024*1024)
```

Someone clarified it's:

```
(20GB / 4KB Page size * 8 bytes per PTE) / 1KB
```

(the 1KB at the end should be 1MB)

Is this correct? Why or why not?

How Much Space Do Our Page Tables Need In the Best Case?

$\frac{20\times 2^{30}}{2^12} = 20 \times 2^{18}$ PTEs

However, these are how many PTEs we need across only the L0 page tables!

$\frac{20\times 2^{18}}{2^9} = 20 \times 2^9 = 10240$ full L0 page tables (40 MB)

Each L1 page table can point to 512 L0 page tables

$\frac{10240}{512} = 20$ full L1 page tables

So we'd need **10260** full page tables $= \frac{10260\times 4096}{2^{20}} = 40.078125$ MiB

Finish Lab 6 with the Remaining Time!

I’ll be in class to help!