COMPUTATIONAL PHYSICS: INCLUDES PARALLEL COMPUTING/PARALLEL PROGRAMMING

ERNEST YEUNG ERNESTYALUMNI@GMAIL.COM

Contents

Date: 23 mai 2016.
Key words and phrases. Computational Physics, Parallel Computing, Parallel Programming.

Abstract. Everything about Computational Physics, including Parallel computing/ Parallel programming.

Part 1. Algorithms as needed

1. Hash Tables

Basics of Hash Tables
Load Balancing and Power of Hashing, Math cup Programming,

2. Divide and Conquer

Divide and Conquer
https://classes.soe.ucsc.edu/cmps102/Fall01/solutions4.pdf
http://www3.cs.stonybrook.edu/~skiena/373/hw/hw.pdf

Part 2. Hardware; Memory, C, C++, CUDA C/C++

stackoverflow: Where in memory are my variables stored in C?

3. Pointers in C; Pointers in C categorified (interpreted in Category Theory) and its relation to actual, physical, computer memory and (memory) addresses ((address) bus; pointers, structs, arrays in C

From Shaw (2015) [1], Exercise 15,
e.g. `ages[i]`, You're indexing into array `ages`, and you're using the number that's held in i to do it:

$$\begin{matrix} a : \mathbb{Z} \to \text{Type} \in \textbf{Type} \\ a : i \mapsto a[i] \end{matrix} \quad \text{e.g.} \quad \begin{matrix} a : \mathbb{Z} \to \mathbb{R} \text{ or } \mathbb{Z} \\ a : i \mapsto a[i] \end{matrix}$$

Index $i \in \mathbb{Z}$ is a location *inside* `ages` or a, which can also be called *address*. Thus, $a[i]$.
Indeed, from cppreference for Member access operators,
Built-in *subscript* operator provides access to an object pointed-to by pointer or array operand. And so `E1[E2]` is exactly identical to `*(E1+E2)`.
To C, e.g. `ages`, or a, is a location in computer's memory where, e.g., all these integers (of `ages`) start, i.e. where a starts.

$$\textbf{Memory}, \quad \text{Obj}(\textbf{Memory}) \ni \text{ memory location. Also, to specify CPU,}$$
$$\textbf{Memory}_{CPU}, \quad \text{Obj}(\textbf{Memory}_{CPU}) \ni \text{ computer memory location}$$

It's *also* an address, and C compiler will replace e.g. `ages` or array a, anywhere you type it, with address of very 1st integer (or 1st element) in, e.g. `ages`, or array a.

$$\textbf{Arrays} \overset{\cong}{\longleftrightarrow} \textbf{address}$$
$$\text{Obj}(\textbf{Arrays}) \overset{\cong}{\longleftrightarrow} \text{Obj}(\textbf{address})$$
$$a \overset{\cong}{\longleftrightarrow} \texttt{0x17}$$

"But here's the trick": e.g. "`ages` is an address inside the *entire computer*." (Shaw (2015) [1]).

It's not like i that's just an address inside `ages`. `ages` array name is actually an address in the computer.
"This leads to a certain realization: C thinks your whole computer is 1 massive array of bytes."
"What C does is layer on top of this massive array of bytes the concept of types and sizes of those types." (Shaw (2015) [1]).
Let

$$\textbf{Memory}_{CPU} := 1 \text{ massive array of bytes}$$
$$\text{Obj}(\textbf{Memory}_{CPU})$$

$$\textbf{Type}$$
$$\text{Obj}(\textbf{Type}) \ni \texttt{int}, \texttt{char}, \texttt{float}$$
$$\text{Obj}(\textbf{Type}) \xrightarrow{\texttt{sizeof}} \mathbb{Z}^+$$
$$T \xmapsto{\texttt{sizeof}} \texttt{sizeof}(T)$$
$$\texttt{float} \xmapsto{\texttt{sizeof}} \texttt{sizeof}(\texttt{float})$$

How C is doing the following with your arrays:

- *Create* a block of memory inside your computer:

$$\text{Obj}(\textbf{Memory}_{CPU}) \supset \text{ Memory block}$$

  Let $\text{Obj}(\textbf{Memory}_{CPU})$ be an ordered set. Clearly, then memory can be indexed. Let $b \in \mathbb{Z}^+$ be this index. Then Memory block$(0) = \text{Obj}(\textbf{Memory}_{CPU})(b)$.
- *Pointing* the name `ages`, or a, to beginning of that (memory) block.
  Entertain, possibly, a category of pointers, **Pointers** $\equiv$ **ptrs**.

$$\textbf{ptrs}$$
$$\text{Obj}(\textbf{ptrs}) \ni a, \text{ e.g. } \texttt{ages}$$
$$a \mapsto \text{Memory block}(0)$$
$$\text{Obj}(\textbf{ptrs}) \to \text{Obj}(\textbf{Memory}_{CPU})$$

- *indexing* into the block, by taking the base address of `ages`, or a

$$a \overset{\cong}{\longmapsto} \text{ base address } 0x17$$
$$\text{Obj}((T)\textbf{array}) \overset{\cong}{\longrightarrow} \text{Obj}(\textbf{addresses})$$
$$a[i] \equiv a + i \overset{\cong}{\longmapsto} \text{ base address + i * } \texttt{sizeof}(T) \overset{*}{\mapsto} a[i] \in T \text{ where } T, \text{ e.g. } T = \mathbb{Z}, \mathbb{R}$$
$$\text{Obj}((T)\textbf{array} \overset{\cong}{\longrightarrow} \text{Obj}(\textbf{addresses}) \to T$$

"A pointer is simply an address pointing somewhere inside computer's memory with a type specifier." Shaw (2015) [1]
C knows where pointers are pointing, data type they point at, size of those types, and how to get the data for you.

3.0.1. *Practical Pointer Usage.*

- Ask OS for memory block (chunk of memory) and use pointer to work with it. This includes strings and `structs`.
- Pass by reference - pass large memory blocks (like large structs) to functions with a pointer, so you don't have to pass the entire thing to the function.
- Take the address of a function, for dynamic callback. (function of functions)

"You should go with arrays whenever you can, and then only use pointers as performance optimization if you absolutely have to." Shaw (2015) [1]

3.0.2. *Pointers are not Arrays.* No matter what, pointers and arrays are not the same thing, even though C lets you work with them in many of the same ways.

From Eli Bendersky's website, Are pointers and arrays equivalent in C?

From Eli Bendersky's website, Are pointers and arrays equivalent in C?

He also emphasizes that

3.0.3. *Variable names in C are just labels.* "A variable in C is just a convenient, alphanumeric pseudonym of a memory location." (Bendersky, [2]). What a compiler does is, create label in some memory location, and then access this label instead of always hardcoding the memory value.

"Well, actually the address is not hard-coded in an absolute way because of loading and relocation issues, but for the sake of this discussion we don't have to get into these details." (Bendersky, [2]) (EY : 20171109 so it's on the address bus?)

Compiler assigns label *at compile time*. Thus, the great difference between arrays and pointers in C.

3.0.4. *Arrays passed to functions are converted to pointers.* cf. Bendersky, [2].

Arrays passed into functions are always converted into pointers. The argument declaration `char arr_place[]` in

```
void foo(char arr_arg[], char* ptr_arg)
{
char a = arr_arg[7];
char b = ptr_arg[7];
}
```

is just syntactic sugar to stand for `char* arr_place`.

From Kernighan and Ritchie (1988) [3], pp. 99 of Sec. 5.3, Ch. 5 Pointers and Arrays,

> When an array name is passed to a function, what is passed is the location of the initial element. Within the called function, this argument is a local variable, and so an array name parameter is a pointer, that is, a variable containing an address.

Why?

The C compiler has no choice here, since,

array name is a label the C compiler replaces *at compile time* with the address it represents (which for arrays is the address of the 1st element of the array).

But function isn't called at compile time; it's called *at run time*.

At run time, (where) something should be placed on the stack to be considered as an argument.

Compiler cannot treat array references inside a function as labels and replace them with addresses, because it has no idea what actual array will be passed in at run time.

EY : 20171109 It can't anticipate the exact arguments that'll it be given *at run-time*; at the very least, my guess is, it's given instructions.

Bendersky [2] concludes by saying the difference between arrays and pointers does affect you. One way is how arrays can't be manipulated the way pointers can. Pointer arithmetic isn't allowed for arrays and assignment to an array of a pointer isn't allowed. cf. van der Linden (1994) [4]. Ch. 4, 9, 10.

Bendersky [2] has this one difference example, "actually a common C gotcha":

"Suppose one file contains a global array:"

```
char my_Arr[256]
```

Programmer wants to use it in another file, *mistakingly* declares as

```
extern char* my_arr;
```

When he tries to access some element of the array using this pointer, he'll most likely get a segmentation fault or a fatal exception (nomenclature is OS-dependent).

To understand why, Bendersky [2] gave this hint: look at the assembly listing

```
char a = array_place[7];

0041137E  mov  al,byte ptr [_array_place+7 (417007h)]
00411383  mov  byte ptr [a],al

char b = ptr_place[7];

00411386  mov  eax,dword ptr [_ptr_place (417064h)]
0041138B  mov  cl,byte ptr [eax+7]
0041138E  mov  byte ptr [b],cl
```

or my own, generated from `gdb` on Fedora 25 Workstation Linux:

```
0x00000000004004b1 <main+11>:    movzbl 0x200b8f(%rip),%eax        # 0x601047 <array_place+7>
0x00000000004004b8 <main+18>:    mov    %al,-0x1(%rbp)

0x00000000004004bb <main+21>:    mov    0x200be6(%rip),%rax        # 0x6010a8 <ptr_place>
0x00000000004004c2 <main+28>:    movzbl 0x7(%rax),%eax
0x00000000004004c6 <main+32>:    mov    %al,-0x2(%rbp)
```

"How will the element be accessed via the pointer? What's going to happen if it's not actually a pointer but an array?" Bendersky [2]

EY : 20171106. Instruction-level, the pointer has to

- `mov  0x200be6(%rip),%rax` - 1st., copy value of the pointer (which holds an address), into `%rax` register.
- `movzbl 0x7(%rax),%eax` - off that address in register '
- `mov  %al,-0x2(%rbp)` - `mov` contents `-0x2(%rbp)`into register `%al`

If it's not actually a pointer, but an array, the value is copied into the `%rax` register is an actual `char` (or `float`, some type). *Not* an address that the registers may have been expecting!

3.1. **Structs in C.** From Shaw (2015) [1], Exercise 16,

`struct` in C is a collection of other data types (variables) that are stored in 1 block of memory. You can access each variable independently by name.

- The `struct` you make, i.e.g `struct Person` is now a *compound data type*, meaning you can refer to `struct Person` using the same kinds of expressions you would for other (data) types.
- This lets you pass the whole `struct` to other functions
- You can access individual members of `struct` by their names using `x->y` if dealing with a ptr.

If you didn't have `struct`, you'd have to figure out the size, packing, and location of memory of the contents. In C, you'd let it handle the memory structure and structuring of these compound data types, `structs`. (Shaw (2015) [1])

4. C, Stack and Heap Memory Management, Heap and Stack Memory Allocation

cf. Ex. 17 of Shaw (2015) [1]

Consider chunk of RAM called stack, another chunk of RAM called heap. Difference between heap and stack depends on where you get the storage.

Heap is all the remaining memory on computer. Access it with `malloc` to get more.

Each time you call `malloc`, the OS uses internal functions (EY : 20171110 address bus or overall system bus?) to register that piece of memory to you, then returns ptr to it.

When done, use `free` to return it to OS so OS can use it for other programs. Failing to do so will cause program to *leak* memory. (EY: 20171110, meaning this memory is unavailable to the OS?)

Stack, on a special region of memory, stores temporary variables, which each function creates as locals to that function. How stack works is that each argument to function is *pushed* onto stack and then used inside the function. Stack is really a stack data structure, LIFO (last in, first out). This also happens with all local variables in `main`, such as `char action`, `int id`. The advantage of using stack is when function exits, *C compiler* pops these variables off of stack to clean up.

Shaw's mantra: If you didn't get it from `malloc`, or a function that got it from `malloc`, then it's on the stack.

3 primary problems with stacks and heaps:

- If you get a memory block from `malloc`, and have that ptr on the stack, then when function exits, ptr will get popped off and lost.
- If you put too much data on the stack (like large structs and arrays), then you can cause a *stack overflow* and program will abort. Use the heap with `malloc`.
- If you take a ptr, to something on stack, and then pass or return it from your function, then the function receiving it will *segmentation fault*, because actual data will get popped off and disappear. You'll be pointing at dead space.

cf. Ex. 17 of Shaw (2015) [1]

5. Data segments; Towards Segmentation Fault

cf. Ferres (2010) [5]

When program is loaded into memory, it's organized into 3 *segments* (3 areas of memory): let executable program generated by a compiler (e.g. `gcc`) be organized in memory over a range of addresses (EY : 20171111 assigned to physical RAM memory by address bus?), ordered, from low address to high address.

- *text* segment or code segment - where compiled code of program resides (from lowest address); code segment contains code executable or, i.e. code binary.

  As a memory region, text segment may be placed below heap, or stack, in order to prevent heaps and stack overflows from overwriting it.

  Usually, text segment is sharable so only a single copy needs to be in memory for frequently executed programs, such as text editors, C compiler, shells, etc. Also, text segment is often read-only, to prevent program from accidentally modifying its instructions. cf. Memory Layout of C Programs; GeeksforGeeks
- Data segment: data segment subdivided into 2 parts:
  - initialized data segment - all global, static, constant data stored in data segment. Ferres (2010) [5]. Data segment is a portion of virtual address of a program.
    Note that, data segment not read-only, since values of variables can be altered at run time.
    This segment can also be further classified into initialized read-only area and initialized read-write area.
    e.g. `char s[] = "hello world"` and `int debut = 1` *outside the main (i.e. global)* stored in initialized read-write area.
    `const char *string = "hello world"` in global C statement makes string literal "hello world" stored in initialized read-only area. Character pointer variable string in initialized read-write area. cf. Memory Layout of C Programs
  - uninitialized data stored in BSS. Data in this segment is initialized by kernel (OS?) to arithmetic 0 before program starts executing.
    Uninitialized data starts at end of data segment ("largest" address for data segment) and contains all global and static variables initialized to 0 or don't have explicit initialization in source code.
    e.g. `static int i;` in BSS segment.
    e.g. `int j;` global variable in BSS segment.
    cf. Memory Layout of C Programs
- Heap - "grows upward" (in (larger) address value, begins at end of BSS segment), allocated with `calloc, malloc`, "dynamic memory allocation".

  Heap area shared by all shared libraries and dynamically loaded modules in a process.

  Heap grows when memory allocator invokes `brk()` or `sbrk()` system call, mapping more pages of physical memory into process's virtual address space.
- Stack - store local variables, used for passing arguments to functions along with return address of the instruction which is to be executed after function call is over.

  When a new stack frame needs to be added (resulting from a *newly called function*), stack "grows downward." (Ferres (2010) [5])

  Stack grows automatically when accessed, up to size set by kernel (OS?) (which can be adjusted with setrlimit(`RLIMIT_STACK`,...)).

5.0.1. *Mathematical description of Program Memory (data segments), with **Memory**, **Addresses**.* Let **Address**, with $\mathrm{Obj}(\textbf{Address}) \cong \mathbb{Z}^+$ be an ordered set.

Memory block $\subset \mathrm{Obj}(\textbf{Memory})$, s.t.

$$\text{Memory block} \overset{\cong}{\mapsto} \{ \text{ low address }, \text{ low address} + \texttt{sizeof}(T), \dots \text{ high address } \} \equiv \text{addresses}_{\text{Memory block}} \subset \mathbb{Z}^+$$

where $T \in \mathrm{Obj}(\textbf{Types})$ and $\cong$ assigned by address bus, or the virtual memory table, and $\text{addresses}_{\text{Memory block}} \subset \mathrm{Obj}(\textbf{Addresses})$.

Now,

text segment, (initialized) data segment, (uninitialized) data segment, heap, stack, command-line arguments and environmental variables $\subset \text{addresses}_{\text{Memory block}}$, that these so-called data segments are discrete subsets of the set of all addresses assigned for the memory block assigned for the program.

Now, $\forall i \in$ text segment , $\forall j \in$ (initialized) data segment , $i < j$ and $\forall j \in$ (initialized) data segment , $\forall k \in$ (uninitiaized) data segment , $j < k$, and so on. Let's describe this with the following notation:

$$\begin{aligned} &\text{text segment} < \text{(initialized) data segment} < \text{(uninitialized) data segment} < \\ &< \text{heap} < \text{stack} < \text{ command-line arguments and environmental variables} \end{aligned} \tag{1}$$

Consider stack of variable length $n_{\text{stack}} \in \mathbb{Z}^+$. Index the stack by $i_{\text{stack}} = 0, 1, \dots n_{\text{stack}} - 1$. "Top of the stack" is towards "decreasing" or "low" (memory) address, so that the relation between "top of stack" to beginning of the stack and high address to low address is *reversed*:

$$i_{\text{stack}} \mapsto \text{high addreess} - i_{\text{stack}}$$

Call stack is composed of stack frames (i.e. "activation records"), with each stack frame corresponding to a subroutine call that's not yet termined with a routine (at any time).

The *frame pointer* `FP` points to location where stack pointer was.

Stack pointer usually is a register that contains the "top of the stack", i.e. stack's "low address" currently, Understanding the stack, i.e.

$$\text{eval}(RSP) = \text{high address} - i_{\text{stack}} \tag{2}$$

5.0.2. *Mathematical description of strategy for stack buffer overflow exploitation.* Let $n_{\text{stack}} = n_{\text{stack}}(t)$. Index the stack with i_{stack} (from "bottom" of the stack to the "top" of the stack):

$$0 < i_{\text{stack}} < n_{\text{stack}} - 1$$

Recall that $i_{\text{stack}} \in \mathbb{Z}^+$ and

$$i_{\text{stack}} \mapsto \text{high address } - i_{\text{stack}} \equiv x = x(i_{\text{stack}}) \in \text{Addresses}_{\text{Memory block}} \subset \mathrm{Obj}(\textbf{Address})$$

Let an array of length L (e.g. `char` array) `buf`, with `\&buf` = `\&buf(0)` $\in \mathrm{Obj}(\textbf{Address})$, be s.t. `&buf` $= x(n_{\text{stack}} - 1)$ (starts at "top of the stack and "lowest" address of stack at time t, s.t.

$$\texttt{\&buf}(j) = \texttt{\&buf}(0) + j\texttt{sizeof}(T)$$

with $T \in \textbf{Types}$).

Suppose return address of a function (such as `main`), eval(`RIP`) be

$$\text{eval}(\texttt{RIP}) = \texttt{\&buf} + L \text{or at least eval}(\texttt{RIP}) \geq \texttt{\&buf} + L$$

If we write to `buf` more values than L, we can write over eval(`RIP`), making eval(`RIP`) a different value than before.

5.1. **Stack.** cf. Ferres (2010) [5]

Stack and functions: When a function executes, it may add some of its state data to top of the stack (EY : 20171111, stack grows downward, so "top" is smallest address?); when function exits, stack is responsible for removing that data from stack.

In most modern computer systems, each thread has a reserved region of memory, stack. Thread's stack is used to store location of function calls in order to allow return statements to return to the correct location.

- OS allocates stack for each system-level thread when thread is created.
- Stack is attached to thread, so when thread exits, that stack is reclaimed, vs. heap typically allocated by application at runtime, and is reclaimed when application exits.
- When thread is created, stack size is set.
- Each byte in stack tends to be reused frequently, meaning it tends to be mapped to the processor's cache, making if very fast.
- Stored in computer RAM, *just like* the heap.
- Implemented with an actual stack data structure.
- stores local data, return addresses, used for parameter passing
- Stack overflow, when too much stack is used (mostly from infinite (or too much) recursion, and very large allocation)
- Data created on stack can be used without pointers.

Also note, for **physical location in memory**, because of Virtual Memory, makes your program think that you have access to certain addresses where physical data is somewhere else (even on hard disc!). Addresses you get for stack are in increasing order as your call tree gets deeper.

memory management - What and where are the stack and heap? Stack Overflow, Tom Leys' answer

5.2. **Stack overflow.** If you use heap memory, and you overstep the bounds of your allocated block, you have a decent chance of triggering a segmentation fault (not 100

On stack, since variables created on stack are always contiguous with each other; writing out of bounds can change the value of another variable. e.g. buffer overflow.

5.3. **Heap.** Heap contains a linked list of used and free blocks. New allocations on the heap (by `new` or `malloc`) are satisfied by creating suitable blocks from free blocks.

This requires updating list of blocks on the heap. This meta information about the blocks on the heap is stored *on the heap* often in a small area in front of every block.

- Heap size set on application startup, but can grow as space is needed (allocator requests more memory from OS)
- heap, stored in computer RAM, like stack.

5.3.1. *Memory leaks.* Memory leaks occurs when computer program consumes memory, but memory isn't released back to operating system.

"Typically, a memory leak occurs because dynamically allocated memory becomes unreachable." (Ferres (2010) [5]).

Programs `./Cmemory/heapstack/Memleak.c` deliberately leaks memory by losing the pointer to allocated memory.

Note, generally, the OS delays real memory allocation until something is written into it, so program ends when virtual addresses run out of bounds (per process limits).

5.4. **More Segmentation Faults.** The operating system (OS) is running the program (its instructions). Only from the hardware, with memory protection, with the OS be signaled to a memory access violation, such as writing to read-only memory or writing outside of allotted-to-the-program memory, i.e. data segments. On `x86_64` computers, this general protection fault is initiated by protection mechanisms from the hardware (processor). From there, OS can signal the fault to the (running) process, and stop it (abnormal termination) and sometimes core dump.

For virtual memory, the memory addresses are mapped by program called *virtual addresses* into *physical addresses* and the OS manages virtual addresses space, hardcare in the CPU called memory management unit (*MMU*) translates virtual addresses to physical addresses, and kernel manages memory hierarchy (eliminating possible overlays). In this case, it's the *hardware* that detects an attempt to refer to a non-existent segment, or location outside the bounds of a segment, or to refer to location not allowed by permissions for that segment (e.g. write on read-only memory).

5.4.1. *Dereferencing a ptr to a NULL ptr (in C) at OS, hardware level.* The problem, whether it's for dereferencing a pointer that is a null pointer, or uninitialized pointer, appears to (see the `./Cmemory/` subfolder) be at this instruction at the register level:

```
x000000000040056c <+38>:        movzbl (%rax),%eax

// or

0x00000000004004be <+24>:        movss  %xmm0,(%rax)
```

involving the register RAX, a temporary register and to return a value, upon assignment. And in either case, register RAX has trying to access virtual (memory) address `0x0` (to find this out in `gdb`, do `i r` or `info register`).

Modern OS's run user-level code in a mode, such as *protected mode*, that uses "paging" (using secondary memory source than main memory) to convert virtual addresses into physical addresses.

For each process (thread?), the OS keeps a *page table* dictating how addresses are mapped. Page table is stored in memory (and protected, so user-level code can't modify it). For every memory access, given (memory) address, CPU translates address according to the page table.

When address translation fails, as in the case that *not all addresses are valid*, and so if a memory access generates an invalid address, the processor (hardware!) raises a *page fault exception*. "This triggers a transition from *user mode* (aka *current privilege level (CPL) 3* on x86/x86-64) into *kernel mode* (aka CPL 0) to a specific location in the kernel's code, as defined by the *interrupt descriptor table* (IDT)." cf. [What happens in OS when we dereference a NULL pointer in C?](https://stackoverflow.com/questions/12645647/what-happens-in-os-when-we-dereference-a-null-pointer-in-c)

Kernel regains control and send signal (EY : 20171115 to the OS, I believe).

In modern OS's, page tables are usually set up to make the address 0 an invalid virtual address.

cf. [What happens in OS when we dereference a NULL pointer in C?](https://stackoverflow.com/questions/12645647/what-happens-in-os-when-we-dereference-a-null-pointer-in-c)

Part 3. C++

6. Free Store

GotW #9, Memory Management - Part I

cf. 11.2 Introduction of Ch. 11 Select Operations [6].

7. Initializer list, parentheses vs. brackets

8. Copy vs. Move

cf. 17.1 Introduction of Ch. 17 Construction, Cleanup, Copy, and Move of Stroustrup [6].

Difference between *move* and *copy*: after a copy, 2 objects must have same value; whereas after a move, the source of the move isn't required to have its original value. So moves can be used when source object won't be used again.

Refs.: Sec. 3.2.1.2, Sec. 5.2, notion of moving a resource, Sec. 13.2-Sec.13.3, object lifetime and errors explored further in Stroustrup [6]

5 situations in which an object is copied or moved:

- as source of an *assignment*
- as object initializer
- as function argument
- as function return value
- as an exception

8.1. **Copy constructor.** cf. Copy constructors, cppreference.com

Copy constructor of class T is non-template constructor whose 1st parameter is `T&, const T&, volatile T&`, or `const volatile T&`.

```
class_name ( const class_name & )
class_name ( const class_name & ) = default;
class_name ( const class_name & ) = delete;
```

8.1.2. *Explanation.*

(1) Typical declaration of a copy constructor.
(2) Forcing copy constructor to be generated by the compiler.
(3) Avoiding implicit generation of copy constructor.

Copy constructor called whenever an object is **initialized** (by **direct-initialization** or **copy-initialization**) from another object of same type (unless **overload resolution** selects better match or call is **elided** (???)), which includes

- initialization `T a = b;` or `T a(b);`, where b is of type T;
- function argument passing: `f(a);`, where a is of type T and f is `void f(T t);`
- function return: `return a;` inside function such as `T f()`, where a is of type T, which has no **move constructor**.

```
struct A
{
int n;
A(int n = 1) : n(n) { }
A(const A& a) : n(a.n) { } // user-defined copy ctor
};

struct B : A
{
// implicit default ctor B::B()
// implicit copy ctor B::B(const B&)
};

int main()
{
A a1(7);
A a2(a1); // calls the copy ctor
B b;
B b2 = b;
A a3 = b; // conversion to A& and copy ctor
}
```

i.e. cf. Copy Constructor in C++

Definition 1. ***Copy constructor*** *is a member function which initializes an object using another object of the same class.*

8.1.4. *When is copy constructor called?*

(1) When object of class returned by value
(2) When object of class is passed (to a function) by value as an **argument**.
(3) When object is constructed based on another object of same class (or overloaded)
(4) When compiler generates temporary object

However, it's not guaranteed copy constructor will be called in all cases, because C++ standard allows compiler to optimize the copy away in certain cases.

8.1.5. *When is used defined copy constructor needed? shallow copy, deep copy.* If we don't define our own copy constructor, C++ compiler creates default copy constructor which does member-wise copy between objects.

We need to define our own copy constructor only if an object has pointers or any run-time allocation of resource like file handle, network connection, etc.

8.1.6. *Default constructor does only shallow copy.*

8.1.7. *Deep copy is possible only with user-defined copy constructor.* We thus make sure pointers (or references) of copied object point to new memory locations.

```
MyClass t1, t2;
MyClass t3 = t1;          // ----> (1)
t2 = t1;                  // ----> (2)
```

Copy constructor called when new object created from an existing object, as copy of existing object, in (1). Assignment operator called when already initialized object is assigned a new value from another existing object, as assignment operator is called in (2).

8.1.9. *Why argument to a copy constructor should be const?* cf. Why copy constructor argument should be const in C++?, geeksforgeeks.org

(1) Use `const` in C++ whenever possible so objects aren't accidentally modified.
(2) e.g.

```
#include <iostream>

class Test
{
/* Class data members */
public:
Test(Test &t)   { /* Copy data members from t */ }
Test()                  { /* Initialize data members */ }
};

Test fun()
{
Test t;
return t;
};

int main()
{
```

```
Test t1;
Test t2 = fun();  error: invalid initialization of non-const reference of type Test \& from an rvalue of type Test
}
```

`fun()` returns by value, so compiler creates temporary object which is copied to t2 using copy constructor (because this temporary object is passed as argument to copy constructor since compiler generates temp. object). Compiler error is because **compiler-created temporary objects cannot be bound to non-const references**.

8.2. **Move Constructor.** For a class, to control what happens when we move, or move and assign object of this class type, use special member function *move constructor*, *move-assignment operator*, and define these operations. Move constructor and move-assignment operator take a (usually nonconst) rvalue reference, to its type. Typically, move constructor moves data from its parameter into the newly created object. After move, it must be safe to run the destructor on the given argument. cf. Ch. 13 of Lippman, Lajole, and Moo (2012) [15]

9. Design Patterns

cf. C++ Programming/Code/Design Patterns
Each design pattern consists of:

- **Problem/requirement** Go through a mini analysis design that may be coded to test out the solution. State requirements of the problem to be solved. Usually, this is a common problem that'll occur in more than 1 application.
- **Forces** Constraints, usually technological.
- **Solution** This is the design part of the design pattern.

9.1. **Creational Patterns.** Creational Patterns deal with object creation mechanisms.

9.1.1. *Builder.*

- **Problem** - Want: construct complex object; however, don't want to have a complex constructor member or 1 that would need many arguments
- **Solution** - Define intermediate object, whose member functions define desired object part by part before object is available to the client. Builder Pattern defer the construction of object until all options for creation have been specified.

9.1.2. *Factory.*

- **Problem** - Want: decide at run-time what object to be created based on some configuration or application parameter. When writing the code, we don't know what class should be instantiated.
- **Solution** Define an *interface* for creating an object, but let subclasses decide which class to instantiate. Factory method lets a class defer instantiation to subclasses.

9.2. **Structural Patterns.**

9.2.1. *Adapter.* Convert interface of a class into another interface. **Adapter** lets classes work together that couldn't otherwise because of incompatible interfaces.

10. vtable; virtual table

I was given this answer to a question I posed to a 20 year C++ veteran and it was such an important answer (as I did not know a virtual table existed, at all before), that I will copy this, repeat this and explore this extensively:
"The keyword you're looking for is virtual table: " How are virtual functions and vtable implemented?, stackoverflow
Original question, from Brian R. Bondy:

10.1. **How are virtual functions and vtable implemented?** We all know what virtual functions are in C++, but how are they implemented at a deep level?
Can the vtable be modified or even directly accessed at runtime?
Does the vtable exist for all classes, or only those that have at least one virtual function?
Do abstract classes simply have a NULL for the function pointer of at least one entry?
Does having a single virtual function slow down the whole class? Or only the call to the function that is virtual? And does the speed get affected if the virtual function is actually overwritten or not, or does this have no effect so long as it is virtual.
Answer from *community wiki*:

10.1.1. *How are virtual functions implemented at a deep level?* From "Virtual Functions in C++"
Whenever a program has a virtual function declared, a v-table is constructed for the class. The v-table consists of addresses to the virtual functions for classes that contain one or more virtual functions. The object of the class containing the virtual function contains a virtual pointer that points to the base address of the virtual table in memory.
Whenever there is a virtual function call, the v-table is used to resolve to the function address.
An object of the class that contains one or more virtual functions contains a virtual pointer called the vptr at the very beginning of the object in the memory. Hence the size of the object in this case increases by the size of the pointer. This vptr contains the base address of the virtual table in memory.
Note that virtual tables are class specific, i.e., there is only one virtual table for a class irrespective of the number of virtual functions it contains. This virtual table in turn contains the base addresses of one or more virtual functions of the class. At the time when a virtual function is called on an object, the vptr of that object provides the base address of the virtual table for that class in memory. This table is used to resolve the function call as it contains the addresses of all the virtual functions of that class. This is how dynamic binding is resolved during a virtual function call.
cf. "Virtual Functions in C++"

10.1.2. *What is a Virtual Function?* A virtual function is a member function of a class, whose functionality can be over-ridden in its derived classes. It is one that is declared as virtual in the base class using the virtual keyword. The virtual nature is inherited in the subsequent derived classes and the virtual keyword need not be re-stated there. The whole function body can be replaced with a new set of implementation in the derived class.

10.1.3. *What is Binding?* Binding is associating an object or a class with its member. If we call a method `fn()` on an object `o` of a class `c`, we say that object `o` is binded with method `fn()`.
This happens at *compile time* and is known as *static* - or *compile-time* binding. Calls to virtual member functions are resolved during *run-time*. This mechanisms is known as *dynamic-binding*.
The most prominent reason why a virtual function will be used is to have a different functionality in the derived class. The difference between a non-virtual member function and a virtual member function is, the non-virtual member functions are resolved at compile time.

10.1.4. *How does a Virtual Function work?* When a program (code text?) has a virtual function declared, a **v-table** is *constructed* for the class.
The v-table consists of addresses to virtual functions for classes that contain 1 or more virtual functions.
The object of the class containing the virtual function *contains a virtual pointer* that points to the base address of the virtual table in memory. An object of the class that contains 1 or more virtual functions contains a virtual pointer called the **vptr** at the very beginning of the object in the memory. (Hence size of the object in this case increases by the size of the pointer; "memory/size overhead.")

This vptr is added as a hidden member of this object. As such, compiler must generate "hidden" code in the **constructors** of each class to initialize a new object's vptr to the address of its class's vtable.
Whenever there's a virtual function call, vtable is used to resolve to the function address. This vptr contains base address of the virtual table in memory.

Note that virtual tables are class specific, i.e. there's only 1 virtual table for a class, irrespective of number of virtual functions it contains, i.e.

vtable is same for all objects belonging to the same class, and typically is shared between them.

This virtual table in turn contains base addresses of 1 or more virtual functions of the class.

At the time when a virtual function is called on an object, the vptr of that object provides the base address of the virtual table for that class in memory. This table is used to resolve the function call as it contains the addresses of all the virtual functions of that class. This is how dynamic binding is resolved during a virtual function call, i.e.

class (inherited or base/parent) cannot, generally, be determined *statically* (i.e. **compile-time**), so compiler can't decide which function to call at that (compile) time. (Virtual function) call must be dispatched to the right function *dynamically* (i.e. **run-time**).

10.1.5. *Virtual Constructors and Destructors.* A constructor cannot be virtual because at the time when constructor is invoked, the vtable wouldn't be available in memory. Hence, we can't have a virtual constructor.

A virtual destructor is 1 that's declared as virtual in the base class, and is used to ensure that destructors are called in the proper order. Remember that destructors are called in reverse order of inheritance. If a base class pointer points to a derived class object, and we some time later use the delete operator to delete the object, then the derived class destructor is not called.

Finally, the article "Virtual Functions in C++" concludes, saying, "Virtual methods should be used judiciously as they are slow due to the overhead involved in searching the virtual table. They also increase the size of an object of a class by the size of a pointer. The size of a pointer depends on the size of an integer." I will have to check this with other references, because, first of all, how then would class inheritance be otherwise implemented?

cf. How are virtual functions and vtable implemented?, stackoverflow

10.1.6. *Can the vtable be modified or even directly accessed at runtime? No.* "Universally, I believe the answer is "no". You could do some memory mangling to find the vtable but you still wouldn't know what the function signature looks like to call it. Anything that you would want to achieve with this ability (that the language supports) should be possible without access to the vtable directly or modifying it at runtime. Also note, the C++ language spec does not specify that vtables are required - however that is how most compilers implement virtual functions."

10.1.7. *Does the vtable exist for all objects, or only those that have at least one virtual function? Only for class with at least 1 virtual function.* I believe the answer here is "it depends on the implementation" since the spec doesn't require vtables in the first place. However, in practice, I believe all modern compilers only create a vtable if a class has at least 1 virtual function. There is a space overhead associated with the vtable and a time overhead associated with calling a virtual function vs a non-virtual function.

10.1.8. *Do abstract classes simply have a NULL for the function pointer of at least one entry? Some do place NULL pointer in vtable, some place pointer to dummy method; in general, undefined behavior.* The answer is it is unspecified by the language spec so it depends on the implementation. Calling the pure virtual function results in undefined behavior if it is not defined (which it usually isn't) (ISO/IEC 14882:2003 10.4-2). In practice it does allocate a slot in the vtable for the function but does not assign an address to it. This leaves the vtable incomplete which requires the derived classes to implement the function and complete the vtable. Some implementations do simply place a NULL pointer in the vtable entry; other implementations place a pointer to a dummy method that does something similar to an assertion.

Note that an abstract class can define an implementation for a pure virtual function, but that function can only be called with a qualified-id syntax (ie., fully specifying the class in the method name, similar to calling a base class method from a derived class). This is done to provide an easy to use default implementation, while still requiring that a derived class provide an override.

10.2. **pImpl, shallow copy, deep copy.** cf. Item 22: "When using the Pimpl Idiom, define special member functions in the implementation file," pp. 147 of Meyers (2014) [8].

```
class Widget {                    // still in header "widget.h"
public:
Widget();
~Widget();                        // dtor is needed-see below
...

private:
struct Impl;                      // declare implementation struct
Impl *pImpl;                      // and pointer to it
};
```

Because `Widget` no longer mentions types `std::string, std::vector`, and `Gadget, Widget` clients no longer need to `\#include` headers for these types. That speeds compilation.

incomplete type is a type that has been declared, but not defined, e.g. `Widget::Impl`. There are very few things you can do with an incomplete type, but declaring a pointer to it is 1 of them.

`std::unique_ptrs` is advertised as supporting incomplete types. But, when `Widget w;`, `w`, is destroyed (e.g. goes out of scope), destructor is called and if in class definition using `std::unique_ptr`, we didn't declare destructor, compiler generates destructor, and so compiler inserts code to call destructor for `Widget`'s data member `m_Impl` (or `pImpl`).

`m_Impl` (or `pImpl`) is a `std::unique_ptr<Widget::Impl>`, i.e., a `std::unique_ptr` using default deleter. The default deleter is a function that uses `delete` on raw pointer inside the `std::unique_ptr`. Prior to using `delete`, however, implementations typically have default deleter employ C++11's `static_assert` to ensure that raw pointer doesn't point to an incomplete type. When compiler generates code for the destruction of the `Widget w`, then, it generally encounters a `static_assert` that fails, and that's usually what leads to the error message.

To fix the problem, you need to make sure that at point where code to destroy `std::unique_ptr<Widget::Impl>` is generated, `Widget::Impl` is a complete type. The type becomes complete when its definition has been seen, and `Widget::Impl` is defined inside `widget.cpp`. For successful compilation, have compiler see body of `Widget`'s destructor (i.e. place where compiler will generate code to destroy the `std::unique_ptr` data member) only inside `widget.cpp` after `Widget::Impl` has been defined.

For compiler-generated move assignment operator, move assignment operator needs to destroy object pointed to by `m_Impl` (or `pImpl`) before reassigning it, but in the `Widget` header file, `m_Impl` (or `pImpl`) points to an incomplete type. Situation is different for move constructor. Problem there is that compilers typically generate code to destroy `pImpl` in the event that an exception arises inside the move constructor, and destroying `pImpl` requires `Impl` be complete.

Because problem is same as before, so is the fix - *move definition of move operations into the implementation file.*

For copying data members, support copy operations by writing these functions ourselves, because (1) compilers won't generate copy operations for classes with move-only types like `std::unique_ptr` and (2) even if they did, generated functions would copy only the `std::unique_ptr` (i.e. perform a *shallow copy*), and we want to copy what the pointer points to (i.e., perform a *deep copy*).

If we use `std::shared_ptr`, there'd be no need to declare destructor in `Widget`.

Difference stems from differing ways smart pointers support custom deleters. For `std::unique_ptr`, type of deleter is part of type of smart pointer, and this makes it possible for compilers to generate smaller runtime data structures and faster runtime code. A consequence of this greater efficiency is that pointed-to types must be complete when compiler-generated special functions (e.g. destructors or move operations) are used. For `std::shared_ptr`, type of deleter is not part of the type of smart pointer. This necessitates larger runtime data structures and somewhat slower code, but pointed-to types need not be complete when compiler-generated special functions are employed.

Part 4. Parallel Computing

11. Udacity Intro to Parallel Programming : Lesson 1 - The GPU Programming Model

Owens and Luebki pound fists at the end of this video. =)))) Intro to the class.

11.1. **Running CUDA locally.** Also, Intro to the class, in Lesson 1 - The GPU Programming Model, has links to documentation for running CUDA locally; in particular, for Linux: `http://docs.nvidia.com/cuda/cuda-getting-started-guide-for-linux/index.html`. That guide told me to go download the NVIDIA CUDA Toolkit, which is the https://developer.nvidia.com/cuda-downloads.

For *Fedora*, I chose Installer Type `runfile (local)`.

Afterwards, installation of CUDA on Fedora 23 workstation had been nontrivial. Go see either my github repository ML-grabbag (which will be updated) or my wordpress blog (which may not be upgraded frequently).

$P = VI = I^2R$ heating.

11.2. **(faster) clock speed, instruction level parallelism per clock cycle - Digging Holes, make Computers Run Faster, Chickens or Oxen.** cf. 4. Digging Holes, Around minute 1:41; *Methods for Building a faster processor*

- *Faster clock speed.* Faster clock: let T = time period for single computation $\equiv T(1)$.
  Thus
  $$f = \frac{1}{T} = \frac{1}{T(1)} \qquad \text{(frequency)}$$
  Smaller $T(1)$ increases power consumption.
- **instruction level parallelism per clock cycle** $\sim$ more work per step.
  $$v = \frac{d}{t}$$
  For $t = 1$, how much work d gets done in this "clock cycle?"

In summary (in other words),

Using $vt = d$, consider faster **clock speed**; and so $T(1)$ smaller, and so $v(1) = \frac{d}{T(1)} = \frac{1}{T(1)}$ = clock speed is "bigger" (faster).

This is at the expense of power consumption.

instruction level parallelism per clock cycle (more transistors). Looking at $v = \frac{d}{t}$,

let $t = T(1)$, time period for 1 "cycle",

increase $d = d(1)$, work (instructions) done in 1 cycle. $v = \frac{d}{T(1)}$, d bigger, so v "bigger" (faster) for fixed $T(1)$.

5. Quiz: How to Make Computers Run Faster

threads - "parallel pieces of work on the GPU"

6. Chickens or Oxen?

8. Quiz: How are CPUs Getting Faster? We have more transistors available per computation (i.e. *instruction level parallelism per clock cycle*).

9. Why we Cannot keep increasing clock speed? Heat, power! Can't make processors faster and faster.

10. What kind of Processors are we Building Assuming the major design constraint is power, traditionally for CPUs, CPUs have complex control hardware, allowing for more flexibility and performance, but is expensive in terms of power. GPUs have a simple control hardware, devoting more transistors to computation; its simple units are potentially more power efficient (operations/watt).

11.3. **Definitions of Latency and throughput (or bandwidth).** cf. 12. Building a Power Efficient Processor

We can seek to ***minimize* latency**.

For a set of instructions := process, or instructions, the time interval between instruction(s) initiation to completion, or amount of time to complete a task T is latency

Definition 2 (latency)**.** ***latency*** $= T$

Definition 3 (throughput)**.** *throughput - tasks completed per unit time,*

$$= d/T \tag{3}$$

with T fixed.

EY : 20170601: is throughput = bandwidth (???)

CPUs optimize for latency $T(1)$ (minimize latency $T(1)$).

GPUs optimize for throughput $\frac{d}{T_1}$, T_1 fixed (some unit time). (maximize throughput $\frac{d}{T_1}$, T_1 fixed).

bandwidth := bit rate of available or consumed information capacity (bits per second) $v = \frac{d}{t}$.

EY: 20170601 so throughput and bandwidth defined similarly, but *are they the same notion? Same thing???*

13. Quiz, Latency vs Bandwidth

latency [sec]. From the title "Latency vs. bandwidth", I'm thinking that throughput = bandwidth (???). throughput = job/time (of job).

Given total task, velocity v,
total task $/v$ = latency. throughput = latency/(jobs per total task).

Also, in Building a Power Efficient Processor. Owens recommends the article David Patterson, "Latency..."

11.3.1. *Core GPU Design Tenets.* 14. Core GPU Design Tenets

(1) GPUs have lots of simple compute units, more compute power for simpler control complexity (tradeoff)
(2) Explicitly parallel programming model
(3) optimize for throughput, not latency

cf. GPU from the Point of View of the Developer

$n_{\text{core}} \equiv$ number of cores
$n_{\text{vecop}} \equiv (n_{\text{vecop}}-$wide axial vector operations/*core* core)
$n_{\text{thread}} \equiv$ threads/core (hyperthreading)

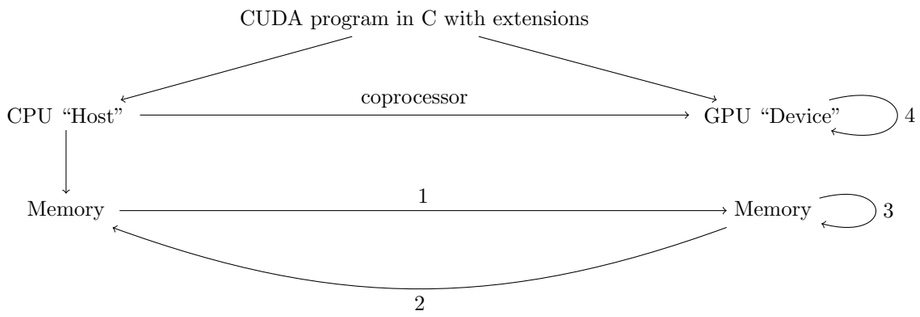
$$n_{\text{core}} \cdot n_{\text{vecop}} \cdot n_{\text{thread}} \text{ parallelism}$$

There were various websites that I looked up to try to find out the capabilities of my video card, but so far, I've only found these commands (and I'll print out the resulting output):

```
$ lspci -vnn | grep VGA -A 12
03:00.0 VGA compatible controller [0300]: NVIDIA Corporation GM200 [GeForce GTX 980 Ti] [10de:17c8] (rev a1) (prog-if 00 [VG
        Subsystem: eVga.com. Corp. Device [3842:3994]
        Physical Slot: 4
        Flags: bus master, fast devsel, latency 0, IRQ 50
        Memory at fa000000 (32-bit, non-prefetchable) [size=16M]
        Memory at e0000000 (64-bit, prefetchable) [size=256M]
        Memory at f0000000 (64-bit, prefetchable) [size=32M]
        I/O ports at e000 [size=128]
        [virtual] Expansion ROM at fb000000 [disabled] [size=512K]
        Capabilities: <access denied>
        Kernel driver in use: nvidia
        Kernel modules: nouveau, nvidia

$ lspci | grep VGA -E
03:00.0 VGA compatible controller: NVIDIA Corporation GM200 [GeForce GTX 980 Ti] (rev a1)

$ grep driver /var/log/Xorg.0.log
[    18.074] Kernel command line: BOOT_IMAGE=/vmlinuz-4.2.3-300.fc23.x86_64 root=/dev/mapper/fedora-root ro rd.lvm.lv=fedora
[    18.087] (WW) Hotplugging is on, devices using drivers 'kbd', 'mouse' or 'vmmouse' will be disabled.
[    18.087]    X.Org XInput driver : 22.1
[    18.192] (II) Loading /usr/lib64/xorg/modules/drivers/nvidia_drv.so
[    19.088] (II) NVIDIA(GPU-0): Found DRM driver nvidia-drm (20150116)
[    19.102] (II) NVIDIA(0):     ACPI event daemon is available, the NVIDIA X driver will
[    19.174] (II) NVIDIA(0): [DRI2]   VDPAU driver: nvidia
[    19.284]    ABI class: X.Org XInput driver, version 22.1
...

$ lspci -k | grep -A 8 VGA
03:00.0 VGA compatible controller: NVIDIA Corporation GM200 [GeForce GTX 980 Ti] (rev a1)
        Subsystem: eVga.com. Corp. Device 3994
        Kernel driver in use: nvidia
        Kernel modules: nouveau, nvidia
03:00.1 Audio device: NVIDIA Corporation GM200 High Definition Audio (rev a1)
        Subsystem: eVga.com. Corp. Device 3994
```

```
        Kernel driver in use: snd_hda_intel
        Kernel modules: snd_hda_intel
05:00.0 USB controller: VIA Technologies, Inc. VL805 USB 3.0 Host Controller (rev 01)
```

CUDA Program Diagram



CPU "host" is the boss (and issues commands) -Owen.

Coprocessor : CPU "host" $\to$ GPU "device"

Coprocessor : CPU process $\mapsto$ (co)-process out to GPU

With

1 data cpu $\to$ gpu

2 data gpu $\to$ cpu (initiated by cpu host)

1., 2., uses `cudaMemcpy`

3 allocate GPU memory: `cudaMalloc`

4 launch kernel on GPU

Remember that for 4., this launching of the kernel, while it's acting on GPU "device" onto itself, it's initiated by the boss, the CPU "host".

Hence, cf. Quiz: What Can GPU Do in CUDA, GPUs can respond to CPU request to receive and send Data CPU $\to$ GPU and Data GPU $\to$ CPU, respectively (1,2, respectively), and compute a kernel launched by the CPU (3).

A CUDA Program A typical GPU program

- `cudaMalloc` - CPU allocates storage on GPU
- `cudaMemcpy` - CPU copies input data from CPU $\to$ GPU
- *kernel launch* - CPU launches kernel(s) on GPU to process the data
- `cudaMemcpy` - CPU copies results back to CPU from GPU

Owens advises minimizing "communication" as much as possible (e.g. the `cudaMemcpy` between CPU and GPU), and do a lot of computation in the CPU and GPU, each separately.

Defining the GPU Computation

Owens circled this

BIG IDEA | This is Important

Kernels look like serial programs

Write your program as if it will run on **one** thread

The GPU will run that program on **many** threads

Squaring A Number on the CPU

Note

(1) Only 1 thread of execution: ("thread" := one independent path of execution through the code) e.g. the `for` loop

(2) no explicit parallelism; it's serial code e.g. the `for` loop through 64 elements in an array

GPU Code A High Level View

CPU:

- Allocate Memory
- Copy Data to/from GPU
- Launch Kernel - species degree of parallelism

GPU:

- Express Out = In $\cdot$ In - says *nothing* about the degree of parallelism

Owens reiterates that in the GPU, everything looks serial, but it's only in the CPU that anything parallel is specified.

pseudocode: CPU code: square kernel <<< 64 >>> (outArray,inArray)

Squaring Numbers Using CUDA Part 3

From the example

```
// launch the kernel
square<<<1, ARRAY_SIZE>>>(d_out , d_in)
```

we're introduced to the "CUDA launch operator", initiating a kernel of 1 block of 64 elements (`ARRAY_SIZE` is 64) on the GPU. Remember that `d_` prefix (this is naming comvention) tells us it's on the device, the GPU, solely.

With CUDA launch operator $\equiv$<<<>>>, then also looking at this explanation on `stackexchange` (so surely others are confused as well, of those who are learning this (cf. CUDA kernel launch parameters explained right?). From Eric's answer,

threads are grouped into blocks. all the threads will execute the invoked kernel function.

Certainly,

$$<<<>>>: (n_{\text{block}}, n_{\text{threads}}) \times \text{kernelfunctions} \mapsto \text{kernelfunction} <<< n_{\text{block}}, n_{\text{threads}} >>> \in \text{End} : \text{Dat}_{\text{GPU}}$$

$$<<<>>>: \mathbb{N}^+ \times \mathbb{N}^+ \times \text{Mor}_{\text{GPU}} \to \text{EndDat}_{\text{GPU}}$$

where I propose that GPU can be modeled as a category containing objects Dat_{GPU}, the collection of all possible data inputs and outputs into the GPU, and Mor_{GPU}, the collection of all kernel functions that run (exclusively, and this *must* be the class, as reiterated by Prof. Owen) on the GPU.

Next,

$$\text{kernelfunction} <<< n_{\text{block}}, n_{\text{threads}} >>>: \text{din} \mapsto \text{dout} \qquad \text{(as given in the "square" example, and so I propose)}$$

$$\text{kernelfunction} <<< n_{\text{block}}, n_{\text{threads}} >>>: (\mathbb{N}^+)^{n_{\text{threads}}} \to (\mathbb{N}^+)^{n_{\text{threads}}}$$

But keep in mind that dout, din are pointers in the C program, pointers to the place in the memory.

`cudaMemcopy` is a functor category, s.t. e.g. $\text{Obj}_{\text{CudaMemcopy}} \ni$ cudaMemcpyDevicetoHost where

$$\text{cudaMemcopy}(-, -, n_{\text{thread}}, \text{cudaMemcpyDeviceToHost}) : \text{Memory}_{\text{GPU}} \to \text{Memory}_{\text{CPU}} \in \text{Hom}(\text{Memory}_{\text{GPU}}, \text{Memory}_{\text{CPU}})$$

Squaring Numbers Using CUDA 4

Note the C language construct *declaration specifier* - denotes that this is a kernel (for the GPU) and not CPU code. Pointers need to be allocated on the GPU (otherwise your program will crash spectacularly -Prof. Owen).

11.3.2. *What are C pointers?* Is $\langle$ type $\rangle *$, a pointer, then a mapping from the category, namely the objects of types, to a mapping from the specified value type to a memory address?

e.g.

$$\langle\rangle * : \text{float} \mapsto \text{float} *$$

$$\text{float } * : \text{din} \mapsto \text{ some memory address}$$

and then we pass in mappings, not values, and so we're actually declaring a square *functor*.

What is `threadIdx`? What is it mathematically? Consider that $\exists$ 3 "modules":

$$\text{threadIdx}.x$$
$$\text{threadIdx}.y$$
$$\text{threadIdx}.z$$

And then the line

```
int idx = threadIdx.x;
```

says that idx is an integer, "declares" it to be so, and then assigns idx to threadIdx.x which surely has to also have the same type, integer. So (perhaps)

$$idx \equiv \text{threadIdx}.x \in \mathbb{Z}$$

is the same thing.

Then suppose threadIdx $\subset$ FinSet, a subcategory of the category of all (possible) finite sets, s.t. threadIdx has 3 particular morphisms, $x, y, z \in \text{Mor}\mathit{threadIdx}$,

$$x : \text{threadIdx} \mapsto \text{threadIdx}.x \in \text{Obj}_{\text{FinSet}}$$
$$y : \text{threadIdx} \mapsto \text{threadIdx}.x \in \text{Obj}_{\text{FinSet}}$$
$$z : \text{threadIdx} \mapsto \text{threadIdx}.x \in \text{Obj}_{\text{FinSet}}$$

Configuring the Kernel Launch Parameters Part 1

$n_{\text{blocks}}, n_{\text{threads}}$ with $n_{\text{threads}} \geq 1024$ (this maximum constant is GPU dependent). You should pick the $(n_{\text{blocks}}, n_{\text{threads}})$ that makes sense for your problem, says Prof. Owen.

11.3.3. *More thoughts on Squaring Numbers Using CPU, and then using CUDA.* Note that this squaring of numbers is really element-wise multiplication of a vector.

I sought an isomorphism between abstract algebra and computer code.

Consider

$$\mathbb{R}^N \ni x \qquad N \in \mathbb{Z}^+$$
$$\mathbb{R} \ni x[i] \qquad i = 1 \dots N \to i = 0, \dots N-1$$

Then the element-wise squaring of numbers is

$$(x[i])^2 = x[i] \cdot x[i]$$

In general,

$$(x[i])^p = \underbrace{x[i] \cdot x[i] \dots x[i]}_{p \text{ times}}$$

11.3.4. *Memory layout of blocks and threads.* $\forall (n_{\text{blocks}}, n_{\text{threads}}) \in \mathbb{Z} \times \{1 \dots 1024\}$, $\{1 \dots n_{\text{block}} \times \{1 \dots n_{\text{threads}}\}$ is now an ordered index (with lexicographical ordering). This is just 1-dimensional (so possibly there's a 1-to-1 mapping to a finite subset of $\mathbb{Z}$).

I propose that "adding another dimension" or the 2-dimension, that Prof. Owen mentions is being able to do the Cartesian product, up to 3 Cartesian products, of the block-thread index.

Quiz: Configuring the Kernel Launch Parameters 2

Most general syntax:

Configuring the kernel launhc

```
kernel<<<grid of blocks, block of threads >>>(...)

// for example

square<<<dim3(bx,by,bz), dim3(tx,ty,tz), shmem>>>(...)
```

where `dim3(tx,ty,tz)` is the grid of blocks $bx \cdot by \cdot bz$
`{dim3}(tx,ty,tz)` is the block of threads $tx \cdot ty \cdot tz$
`shmem` is the shared memory per block in bytes

Quiz: Map

I wanted to try to mathematically formulate the idea of `map`.

$$\texttt{MAP(ELEMENTS,FUNCTION)} \iff \begin{array}{c} \text{given } x \in \mathbb{R}^N \\ x[i] \xrightarrow{f} f(x[i]) \\ \text{or} \\ \text{set of elements (finite, so can be indexed)} \qquad \{x_0, \dots, x_{n-1}\}_{\mathcal{A}} \in \text{ObjFin} \\ x_i \xrightarrow{f} f(x_i), \qquad \forall\, i \in \mathcal{A} \end{array}$$

Problem Set 1 "Also, the image is represented as an 1D array in the kernel, not a 2D array like I mentioned in the video." Here's part of that code for squaring numbers:

```
__global__ void square(float *d_out, float *d_in) {
  int idx = threadIdx.x;
  float f = d_in[idx];
  d_out[idx] = f*f;
  }
```

11.3.5. *Problem Set 1, Udacity CS344.* Let $L_x \equiv$ total number of pixels in x-direction of image $\in \mathbb{Z}^+$
$L_y \equiv$ total number of pixels in y-direction of image $\in \mathbb{Z}^+$
and so $L_x L_y =$ total number of pixels in image.

The formula for ensuring that all threads will be computed, given an arbitrary choice of the number of threads in a (single) block, is the following:

$$\frac{L_x + (M_x - 1)}{M_x} = N_x \in \mathbb{N} \qquad N_x = \text{ number of (thread) blocks in } x\text{-direction}$$
$$\frac{L_y + (M_y - 1)}{M_y} = N_y \in \mathbb{N} \qquad N_y = \text{ number of (thread) blocks in } y\text{-direction}$$

Then

$$(M_x, M_y, 1) \in \mathbb{N}^3 \iff \texttt{dim3}$$

needs to be determined manually, empirically, and in consideration of the actual GPU hardware architecture (look up number of CUDA cores, and allowed maximum threads), where

$$M_x \equiv \text{ number of threads per block in } x\text{-direction}$$
$$M_y \equiv \text{ number of threads per block in } y\text{-direction}$$

Consider that we want to go from the indices on each thread per block, on each block on the grid, in each of the 2 dimensions, to a global 2-dimensional position, and then "flatten" these coordinates to a 1-dimensional array that CUDA C can load onto global memory. In other words, for

$$i_x \in \{0, \dots, M_x - 1\} \iff \texttt{threadIdx.x}$$
$$i_y \in \{0, \dots, M_y - 1\} \iff \texttt{threadIdx.y}$$
$$j_x \in \{0, \dots, N_x - 1\} \iff \texttt{blockIdx.x}$$
$$j_y \in \{0, \dots, N_y - 1\} \iff \texttt{blockIdx.y}$$

and so for

$$(k_x, k_y)$$
$$k_X = i_x + j_x M_x$$
$$k_y = i_y + j_y M_y$$

then we sought the following operations:

$$(j_x, j_y) \times (i_x, i_y) \in \{0, \dots, N_x - 1\} \times \{0 \dots N_y - 1\} \times \{0 \dots M_x - 1\} \times \{0 \dots M_y - 1\} \in \texttt{dim3} \times \texttt{dim3}$$
$$\mapsto (k_x, k_y) \in \{0 \dots L_x - 1\} \times \{0 \dots L - y - 1\}$$
$$\mapsto k = k_x + L_x k_y \in \{0 \dots L_x L_y - 1\}$$

11.3.6. *Grid of blocks, block of threads, thread that's indexed; (mathematical) structure of it all.* Let

$$\text{grid} = \prod_{I=1}^{N} (\text{block})^{n_I^{\text{block}}}$$

where $N = 1, 2, 3$ (for CUDA) and by naming convention $I = 1 \equiv x$, $I = 2 \equiv y$, $I = 3 \equiv z$

Let's try to make it explicitly (as others had difficulty understanding the grid, block, thread model, cf. colored image to greyscale image using CUDA parallel processing, Cuda gridDim and blockDim) through commutative diagrams and categories (from math):

$$\begin{array}{ccc} \prod_{I=1}^{N} \mathbb{Z}^+ & & \ni (N_x^{\text{blocks}}, N_y^{\text{blocks}}, N_z^{\text{blocks}}) \\ \text{gridDim} \circlearrowright \downarrow \text{dim3} & & \text{dim3} \downarrow \circlearrowright (\text{gridDim.}x,\ \text{gridDim.}y,\ \text{gridDim.}z) \\ \text{grid} & & \ni \text{gridSize}(N_x^{\text{blocks}}, N_y^{\text{blocks}}, N_z^{\text{blocks}}) \end{array}$$

$$\begin{array}{ccc} \text{grid} & & \ni \texttt{d\_rgbaImage} \\ \downarrow \text{blockIdx} & & \downarrow (\text{blockIdx.}x,\text{blockIdx.}y,\text{blockIdx.}z) \\ \prod_{I=1}^{N} \mathbb{Z} \supset \prod_{I=1}^{N} \{1 \dots N_I^{\text{blocks}}\} & & \ni (i^{\text{blocks}}, j^{\text{blocks}}, k^{\text{blocks}}) \end{array}$$

and then similar relations (i.e. arrows, i.e. relations) go for a block of threads:

$$\begin{array}{ccc} \prod_{I=1}^{N} \mathbb{Z}^+ & & \ni (N_x^{\text{threads}}, N_y^{\text{threads}}, N_z^{\text{threads}}) \\ \text{blockDim} \circlearrowright \downarrow \text{dim3} & & \text{dim3} \downarrow \circlearrowright (\text{blockDim.}x,\ \text{blockDim.}y,\ \text{blockDim.}z) \\ \text{block} & & \ni \text{blockSize}(N_x^{\text{threads}}, N_y^{\text{threads}}, N_z^{\text{threads}}) \end{array}$$

$$\begin{array}{ccc} \text{block} & & \ni \text{block} \\ \downarrow \text{threadIdx} & & \downarrow (\text{threadIdx.}x,\text{threadIdx.}y,\text{threadIdx.}z) \\ \prod_{I=1}^{N} \mathbb{Z} \supset \prod_{I=1}^{N} \{1 \dots N_I^{\text{threads}}\} & & \ni (i^{\text{threads}}, j^{\text{threads}}, k^{\text{threads}}) \end{array}$$

gridsize help assignment 1 Pp explains how threads per block is variable, and remember how Owens said Luebki says that a GPU doesn't get up for more than a 1000 threads per block.

11.3.7. *Generalizing the model of an image.* Consider vector space V, e.g. $\dim V = 4$, vector space V over field $\mathbb{K}$, so $V = \mathbb{K}^{\dim V}$.

Each pixel represented by $\forall\, v \in V$.

Consider an image, or space, M. $\dim M = 2$ (image), $\dim M = 3$. Consider a local chart (that happens to be global in our case):

$$\varphi : M \to \mathbb{Z}^{\dim M} \supset \{1 \dots N_1\} \times \{1 \dots N_2\} \times \cdots \times \{1 \dots N_{\dim M}\}$$
$$\varphi : x \mapsto (x^1(x), x^2(x), \dots, x^{\dim M}(x))$$

$$\begin{array}{ccc} E & \xrightarrow{\varphi} & M \times V \\ \pi \downarrow & \swarrow & \\ M & & \end{array} \qquad \begin{array}{ccc} E & \xrightarrow{\varphi} & \text{grid} \times \text{ block of threads} \\ \pi \downarrow & \swarrow & \\ \text{grid} & & \end{array}$$

Consider a "coarsing" of underlying M:

$$\begin{array}{ccc} M \times V & \xrightarrow{\qquad\qquad \text{proj} \qquad\qquad} & \text{proj}(M) \times \text{proj}(V) \\ \pi \downarrow & & \downarrow \text{proj}(\pi) \\ M = \{1 \dots N_1\} \times \{1 \dots N_2\} \times \cdots \times \{1 \dots N_{\dim M}\} & \xrightarrow{\text{proj}} & \text{proj}(M) = \{1 \dots \frac{N_1}{N_1^{\text{threads}}}\} \times \{1 \dots \frac{N_2}{N_2^{\text{threads}}}\} \times \cdots \times \{1 \dots \frac{N_{\dim M}}{N_{\dim M}^{\text{threads}}}\} \end{array}$$

e.g. $N_1^{\text{thread}} = 12$

$N_2^{\text{thread}} = 12$

Just note that in terms of syntax, you have the "block" model, in which you allocate blocks along each dimension. So in

$$const\ dim3\ blockSize(n_x^b, n_y^b, n_z^b)$$
$$const\ dim3\ gridSize(n_x^{\text{gr}}, n_y^{\text{gr}}, n_z^{\text{gr}})$$

Then the condition is $n_x^b/\dim V, n_y^b/\dim V, n_z^b/\dim V \in \mathbb{Z}$ (condition), $\qquad (n_x^{\text{gr}} - 1)/\dim V, n_y^{\text{gr}}/\dim V, n_z^{\text{gr}}/\dim V \in \mathbb{Z}$

11.4. **Unit 2, Lesson 2 GPU Hardware and Parallel Communication Patterns.** Transpose Part 1

Now

$$\text{Mat}_{\mathbb{F}}(n, n) \xrightarrow{T} \text{Mat}_{\mathbb{F}}(n, n)$$
$$A \mapsto A^T \text{ s.t. } (A^T)_{ij} = A_{ji}$$
$$\text{Mat}_{\mathbb{F}} \xrightarrow{T} \mathbb{F}^{n^2}$$
$$A_{ij} \mapsto A_{ij} = A_{in+j}$$

$$\begin{array}{ccc} \mathrm{Mat}_{\mathbb{F}}(n,n) & \longrightarrow & \mathbb{F}^{n^2} \\ T\downarrow & & \downarrow T \\ \mathrm{Mat}_{\mathbb{F}}(n,n) & \longrightarrow & \mathbb{F}^{n^2} \end{array} \qquad \begin{array}{ccc} A_{ij} & \longmapsto & A_{in+j} \\ T\downarrow & & \downarrow T \\ (A^T)_{ij} = A_{ji} & \longmapsto & A_{jn+i} \end{array}$$

Transpose Part 2

Possibly, transpose is a functor.

Consider struct as a category. In this special case, Objstruct = {arrays} (a struct of arrays). Now this struct already has a hash table for indexing upon declaration (i.e. "creation"): so this category struct will need to be equipped with a "diagram" from the category of indices J to struct: $J \to$ struct.

So possibly

$$\begin{array}{rcl} \text{struct} & \xrightarrow{T} & \text{array} \\ \text{ObjStruct} = \{ \text{ arrays } \} & \xrightarrow{T} & \text{Objarray} = \{ \text{ struct } \} \\ J \to \text{ struct} & \xrightarrow{T} & J \to \text{ array} \end{array}$$

Quiz: What Kind Of Communication Pattern This quiz made a few points that clarified the characteristics of these so-called communication patterns (amongst the memory?)

- map is bijective, and map : Idx → Idx
- gather - not necessarily surjective
- scatter - not necessarily surjective
- stencil - surjective
- transpose (see before)

Parallel Communication Patterns Recap

- map - bijective
- transpose - bijective
- gather - not necessarily surjective, and is many-to-one (by def.)
- scatter - one-to-many (by def.) and is not necessarily surjective
- stencil - several-to-one (not injective, by definition), and is surjective
- reduce - all-to-one
- scan/sort - all-to-all

Programmer View of the GPU

thread blocks: group of threads that cooperate to solve a (sub)problem

Thread Blocks And GPU Hardware

CUDA GPU is a bunch of SMs:

Streaming Multiprocessors (SM)s

SMs have a bunch of simple processors and memory.

Dr. Luebki:

> Let me say that again because it's really important
> GPU is responsible for allocating blocks to SMs

Programmer only gives GPU a pile of blocks.

Quiz: What Can The Programmer Specify

I myself thought this was a revelation and was not intuitive at first:

Given a single kernel that's launched on many thread blocks include X, Y, the programmer cannot specify the sequence the blocks, e.g. block X, block Y, run (same time, or run one after the other), and which SM the block will run on (GPU does all this).

Quiz: A Thread Block Programming Example

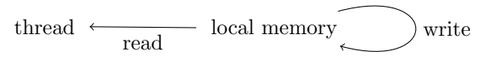
Open up `hello blockIdx.cu` in Lesson 2 Code Snippets (I got the repository from github, repo name is cs344).

At first, I thought you can do a single file compile and run in Eclipse without creating a new project. No. cf. Eclipse creating projects every time to run a single file?.

I ended up creating a new CUDA C/C++ project from File -¿ New project, and then chose project type Executable, Empty Project, making sure to include Toolchain CUDA Toolkit (my version is 7.5), and chose an arbitrary project name (I chose cs344single). Then, as suggested by Kenny Nguyen, I dragged and dropped files into the folder, from my file directory program.

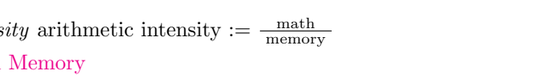
I ran the program with the "Play" triangle button, clicking on the green triangle button, and it ran as expected. I also turned off Build Automatically by deselecting the option (no checkmark).

GPU Memory Model



Then consider threadblock ≡ thread block

$$\begin{array}{l} \text{Objthreadblock} \supset \{ \text{ threads } \} \\ \text{FinSet} \xrightarrow{\text{threadIdx}} \text{ thread } \in \text{Morthreadblock} \end{array}$$



$\forall$ thread,

Synchronization - Barrier

Danger: what if a thread reads a result before another thread writes it?

Threads need to *synchronize*.

one of the most fundamental problems in parallel computing

Quiz: The Need For Barriers

3 barriers were needed (wasn't obvious to me at first). All threads need to finish the write, or initialization, so it'll need a barrier.

While

```
array[idx] = array[idx+1];
```

is 1 line, it'll actually need 2 barriers; first read. Then write.

So *actually* we'll need to *rewrite* this code:

```
int temp = array[idx+1];
__syncthreads();
array[idx] = temp;
__syncthreads();
```

Make sure each *read* and *write* operation is completed.

kernels have implicit barrier for each.

Writing Efficient Programs

(1) Maximize *arithmetic intensity* arithmetic intensity := $\frac{\text{math}}{\text{memory}}$

video: Minimize Time Spent On Memory

local memory is fastest; global memory is slower

$$\text{local} > \text{ shared} >> \text{global} >> \text{CPU}$$

kernel we know (in the code) is tagged with `__global__`

11.4.1. *Coalesce global memory accesses.* 31. Coalesce Memory Access, from Unit 2/Lesson 2 - GPU Hardware and Parallel Communication Patterns

Whenever a thread on the GPU reads or writes global memory, it always acceses a large chunk of memory at once.
Even if the thread needs to only access a smaller subset of that large chunk.
If other threads are making similar memory access, the GPU can exploit that and reuse that larger chunk.
We saw such access pattern is coalesced; GPU must efficient when threads read or write contiguous memory locations.
quiz: A Quiz on Coalescing Memory Access
Work it out as Dr. Luebki did to figure out if it's coalesced memory access or not.
Atomic Memory Operations
Atomic Memory Operations
atomicadd atomicmin atomicXOR atomicCAS Compare And Swap

11.4.2. *On Problem Set 2.* There is what I call the "naive global memory" scheme, that solves the objective of blurring a photo with a local stencil of the values, using only global memory on the GPU.

Given image of size $L_x \times L_y$, i.e. $(L_x, L_y) \in (\mathbb{Z}^+)^2$; image is really a designated or particular mapping f,

$$f : \{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \to \{0 \dots 255\}^4$$
$$f(x, y) = (f^{(r)}(x, y), f^{(b)}(x, y), f^{(g)}(x, y), f^{(\alpha)}(x, y))$$

Consider "naive global memory scheme" - establishing the following notation:

$$i_x \in \{0 \dots M_x - 1\} \iff \texttt{threadIdx.x}$$
$$i_y \in \{0 \dots M_y - 1\} \iff \texttt{threadIdx.y}$$
$$j_x \in \{0 \dots N_x - 1\} \iff \texttt{blockIdx.x}$$
$$j_y \in \{0 \dots N_y - 1\} \iff \texttt{blockIdx.y}$$
$$M_x \in \{1 \dots 1024\} \iff \texttt{blockDim.x}$$
$$M_y \in \{1 \dots 1024\} \iff \texttt{blockDim.y}$$

with

$$N_x := (L_x + M_x - 1)/M_x \in \mathbb{Z}^+$$
$$N_y := (L_y + M_y - 1)/M_y \in \mathbb{Z}^+$$

There should be a functor called "flatten" such that we end up with the image as a 1-dimensional, contiguous array on the global memory of the GPU; so for

$$k = k_x + k_y L_x \in \{0 \dots L_x L_y - 1\}$$

then

$$(k_x, k_y) \iff (x, y) \in \{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \xmapsto{\text{flatten}} k \in \{0 \dots L_x L_y - 1\}$$
$$f : \{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \xrightarrow{\text{flatten}} f : \{0 \dots L_x L_y - 1\} \to \{0 \dots 255\}^4$$
$$f(x, y) = f(k_x, k_y) \xmapsto{\text{flatten}} f(k)$$

Then there should be a functor called "separateChannels" to represent the `__global__` kernel `separateChannels`.

$$f : \{0 \dots L_x L_y - 1\} \to \{0 \dots 255\}^4 \xrightarrow{\text{separateChannels}} f^{(c)} : \{0 \dots L_x L_y - 1\} \to \{0 \dots 255\},\ c = \{r, g, b\}$$
$$f(k) \xmapsto{\text{separateChannels}} f^{(c)}(k)$$

Then consider a "stencil" of size `filterWidth` $\times$ `filterWidth` $\iff W \times W \in (\mathbb{Z}^+)^2$.

Let $(\nu_x, \nu_y) \in \{0 \dots W - 1\}^2$ and so

$$\left(\nu_x - \frac{W}{2}, \nu_y - \frac{W}{2}\right) \in \{\frac{-W}{2}, \dots \frac{W}{2} - 1\} \subset \mathbb{Z}$$

Now let

$$k_x^{\text{st}} = k_x + \nu_x - \frac{W}{2} \iff \texttt{stencilindex\_x}$$
$$k_y^{\text{st}} = k_y + \nu_y - \frac{W}{2} \iff \texttt{stencilindex\_y}$$

with $k_x^{\text{st}} \in \{0 \dots L_x - 1\}$
$k_y^{\text{st}} \in \{0 \dots L_y - 1\}$

We also have to apply the flatten functor on the stencil:

$$(\nu_x, \nu_y) \in \{0 \dots W - 1\}^2 \xmapsto{\text{flatten}} \nu = \nu_x + W\nu_y \in \{0 \dots W^2 - 1\}$$

And so the gist of the blurring operation is in this equation:

$$g^{(c)}(k) = \sum_{\nu_x=0}^{W-1} \sum_{\nu_y=0}^{W-1} c_{\nu=\nu_x+W\nu_y} f^{(c)}(k_x^{\text{st}} + L_x \cdot k_y^{\text{st}}) \qquad \forall\, c = \{r, g, b\} \tag{4}$$

with $k_x^{\text{st}} = k_x^{\text{st}}(\nu_x) := k_x + \nu_x - \frac{W}{2}$
$k_y^{\text{st}} = k_y^{\text{st}}(\nu_y) := k_y + \nu_y - \frac{W}{2}$

11.4.3. *Problem Set 2, shared memory "tiling" scheme.* I think the `__shared__` memory "tiling" scheme is non-trivial due to accounting for the values "at the edges" of the thread block, including the "corners" the so-called "halo" cells. Storing the value of the "cells" or threads within a thread block into shared memory is *relatively* straightforward - it is a 1-to-1 mapping. But taking care of the corner cases, due to the desired "stencil" for blurring, is nontrivial, I think.

Consider my scheme for "tiling" using shared memory:

Let

$$k_x = i_x + j_x M_x \in \{0 \dots L_x - 1\}$$
$$k_y = i_y + j_y M_y \in \{0 \dots L_y - 1\}$$
$$k_x < L_x \text{ and } k_y < L_y$$
$$0 \le k_x < L_x \text{ and } 0 \le k_y < L_y$$
$$k := k_x + L_x k_y$$

and let

$$S_x := M_x + 2r$$
$$S_y := M_y + 2r$$
$$s_x := i_x + r$$
$$s_y := i_y + r$$
$$0 \le s_x < S_x \text{ and } 0 \le s_y < S_y$$
$$s_k := s_x + S_x s_y$$

where r is the "radius" or essentially the stencil size, out in 1-direction.

Loading the regular cells,

$$s_{\text{in}}[s_k] = f^{(c)}(k)$$

Loading the halo cells,

if $(i_x < r)$,
then requiring

$$\begin{aligned} 0 \le s_x - r < S_x \qquad & 0 \le k_x - r < L_x \\ 0 \le s_y < S_y \qquad & 0 \le k_y < L_y \\ s_{\text{in}}[s_x - r + S_x s_y] &= f^{(c)}[k_x - r + L_x k_y] \\ 0 \le s_x + M_x < S_x \qquad & 0 \le k_x + M_x < L_x \\ 0 \le s_y < S_y \qquad & 0 \le k_y < L_y \\ s_{\text{in}}[s_x + M_x + S_x s_y] &= f^{(c)}[k_x + M_x + L_x k_y] \end{aligned}$$

If $(i_y < r)$,
then requiring

$$\begin{aligned} 0 \le s_x < S_x \qquad & 0 \le k_x < L_x \\ 0 \le s_y - r < S_y \qquad & 0 \le k_y - r < L_y \\ s_{\text{in}}[s_x + S_x(s_y - r)] &= f^{(c)}[k_x + L_x(k_y - r)] \\ 0 \le s_x < S_x \qquad & 0 \le k_x < L_x \\ 0 \le s_y + M_y < S_y \qquad & 0 \le k_y + M_y < L_y \\ s_{\text{in}}[s_x + S_x(s_y + M_y)] &= f^{(c)}[k_x + L_x(k_y + M_y)] \end{aligned}$$

And now the actual stencil calculation:
$\forall \nu_y \in \{\nu_y = 0, 1 \dots W-1 | 0 \le \nu_y < W\}$,
$\quad k_y^{\text{st}} := s_y + \nu_y - r$
$\quad \forall \nu_x \in \{\nu_x = 0, 1 \dots W-1 | 0 \le \nu_x < W\}$,
$\quad\quad k_x^{\text{st}} := s_x + \nu_x - r$

$$\begin{aligned} \texttt{inputvalue} &= s_{\text{in}}[k_x^{\text{st}} + S_x k_y^{\text{st}}] \text{ with } 0 \le k_x^{\text{st}} < S_x \\ & \qquad\qquad\qquad\qquad 0 \le k_y^{\text{st}} < S_y \\ \texttt{filtervalue} &= c(\nu_x + W\nu_y) \\ \texttt{value} \mathrel{+}= \texttt{filtervalue} &\cdot \texttt{inputvalue}, \end{aligned}$$

i.e.

$$g^{(c)}(k) = \sum_{\nu_y=0}^{W-1} \sum_{\nu_x=0}^{W-1} c_{\nu=\nu_x+W\nu_y} s_{\text{in}}[k_x^{\text{st}} + k_y^{\text{st}} S_x]$$

Unfortunately, the (literal) corner cases aren't accounted for correctly, (when $i_x < r$ *and* $i_y < r$), as can be seen by the difference image and output image when it's run.

Samuel Lin or Samuel271828 had both an elegant and *correct* implementation. It's also in Samuel Lin or samuellin3310's github repositories, in student fuction improved share.cu (sic).

Here it is, mathematically:

Let

$$\begin{aligned} s_{\text{in}} &\in \mathbb{R}^{(M_x+2r)(M_y+2r)} \\ k_x = i_x + j_x M_x &\in \mathbb{Z} \\ k_y = i_y + j_y M_y &\in \mathbb{Z} \\ k_x < L_x \text{ and } k_y &< L_y \\ 0 \le k_x < L_x \text{ and } 0 \le k_y &< L_y \\ k &:= k_x + L_x k_y \end{aligned}$$

Then,
$\forall i \in \{i = i_x - r, i_x - r + M_x, i_x - r + 2M_x, \dots | i_x - r \le i < M_x + r\}$,
$\quad \forall j \in \{j = i_y - r, i_y - r + M_y, i_y - r + 2M_y \dots | i_y - r \le j < M_y + r\}$,

$$\begin{aligned} l_x &:= i + M_x j_x \in \mathbb{Z} \text{ with (enforcing) } 0 \le l_x < L_x \\ l_y &:= j + M_y j_y \in \mathbb{Z} \text{ with (enforcing) } 0 \le l_y < L_y \\ s_{\text{in}}[i + r + (j + r)(M_x + 2r)] &= f^{(c)}(l_x + l_y L_x) \end{aligned}$$

Enforce $k_x < L_x$ and $k_y < L_y$, otherwise nothing happens.
And now the actual stencil calculation:
$\forall \nu_y \in \{\nu_y = 0, 1 \dots W-1 | 0 \le \nu_y < W\}$,
$\quad k_y^{\text{st}} := s_y + \nu_y - r$
$\quad \forall \nu_x \in \{\nu_x = 0, 1 \dots W-1 | 0 \le \nu_x < W\}$,
$\quad\quad k_x^{\text{st}} := s_x + \nu_x - r$

$$\begin{aligned} \texttt{inputvalue} &= s_{\text{in}}[k_x^{\text{st}} + S_x k_y^{\text{st}}] \text{ with } 0 \le k_x^{\text{st}} < S_x \\ & \qquad\qquad\qquad\qquad 0 \le k_y^{\text{st}} < S_y \\ \texttt{filtervalue} &= c(\nu_x + W\nu_y) \\ \texttt{value} \mathrel{+}= \texttt{filtervalue} &\cdot \texttt{inputvalue}, \end{aligned}$$

i.e.

$$g^{(c)}(k) = \sum_{\nu_y=0}^{W-1} \sum_{\nu_x=0}^{W-1} c_{\nu=\nu_x+W\nu_y} s_{\text{in}}[k_x^{\text{st}} + k_y^{\text{st}} S_x]$$

It may be non-intuitive, as was in my case, from my personal experience, to have to move the requirement that $k_x < L_x$, and $k_y < L_y$, i.e. the line

```
if ( k_x >= numCols || k_y >= numRows ) {
  return; }
```

after loading all the values from global memory into shared memory, and not in the beginning of the kernel. This can be proven, in general. And again, shout-outs (i.e. credit should go) to Samuel Lin or Samuel271828 for the clarifying discussion here.

For simplicity, consider the 1-dimensional case, e.g. a 1-dimensional pixelated image, represented by a 1-dimensional array. The discussion below can easily be generalized to n-dimensions.

Recall that

$$k_x := i_x + M_x j_x$$

for

$$\begin{aligned} i_x &\in \{0 \dots M_x - 1\} \\ j_x &\in \{0 \dots N_x - 1\} \end{aligned}$$

for

$$i_x \iff \texttt{threadIdx.x}$$
$$j_x \iff \texttt{blockIdx.x}$$
$$M_x \iff \texttt{blockDim.x}$$

with N_x determined by a formula immediately below.

Then

$$k_x \in \{0 \dots N_x M_x - 1\}$$

with N_x being determined by

$$N_x := \frac{L_x + M_x - 1}{M_x} \in \mathbb{Z}$$

By integer division, this formula for N_x guarantees that the number of blocks in the grid is the lowest number that would guarantee that *all* the needed grid points are computed, i.e. N_x is the lowest number such that there are enough (i.e. minimal number of) threads that'll compute all the needed grid points, or pixels, or computations, etc., by how integer division works. Note that

$$L_x \iff \texttt{numCols} \text{ and so } N_x \iff \texttt{gridDim.x}$$

Now

$$N_x M_x \geq L_x$$

meaning, that we could have the case where the very last (thread) block would have more threads than is needed to compute all the pixels.

e.g.

$$L_x = 127$$
$$M_x = 128 \text{ (so } N_x = 1)$$
$$k_x = i_x + 128 \cdot 0 = i_x$$
$$k_x = 127$$

and so $k_x \geq L_x$, namely $127 \geq 127$. If we had, in the beginning, the line that returns nothing for the condition $k_x \geq L_x$, we won't be including this case.

Now recall the shared memory tiling scheme, but in 1-dimension (for the purposes of this present discussion):

$$\forall \{i \in i_x - r, i_x - r + M_x, i_x - r + 2M_x, \dots | i_x - r \leq i < M_x + r\}$$

so $-r \leq i < M_x + r$.

Then

$$l_x := i + M_x j_x$$

and

$$s_{\text{in}}[i + r] = f^{(c)}(l_x)$$

e.g.

$$r = \texttt{filterWidth}/2 = 9/2 = 4$$
$$i_x = 127$$
$$i = 127 - 4 = 123 = l_x$$
$$s_{\text{in}}[127] = f^{(c)}(123)$$

$k_x = 127 = L_x$ so $k_x \geq L_x$

So in this very insightful example, we've seen that for $k_x = 127$, it loads the regular value at index $l_x = 123$ into the appropriate slot in the shared memory, namely $s_{\text{in}}[127]$ correctly, and yet if we placed the code line in question that tests k_x against L_x in the wrong order, this step would've been excluded!

How can we see this in general?

Consider now that $l_x = i + M_x j_x = i_x + M_x j_x - r = k_x - r$.

Also

$$i + r = i_x - r + r = i_x$$

Now then

$$s_{\text{in}}[i_x] = f^{(c)}(l_x) = f^{(c)}(k_x - r)$$

while $N_x M_x - r \geq L_x$ or $N_x M_x - r < L_x$. It's this ambiguity that forces the check of $k_x \geq L_x$ to be moved to after loading all values into shared memory. The $l_x \leq L_x$ check guarantees both that we aren't going "outside" the array indices in global memory *and* that we're "clamping" down on the absolute boundary value if we reach the absolute boundary or "end" of the array.

My big takeaway is that doing the shared memory tiling scheme is much more nuanced and deserves more inspection that the relatively straightforward "naive" global memory scheme.

11.5. **Unit 3, Lesson 3 Fundamental GPU Algorithms (Reduce, Scan, Histogram; Udacity cs344).** More on Udacity forums, in particular, forum for cs344: Please elaborate the micro-optimization techniques discussed by the instructor

cf. Reduce Part 2

Reduce: Inputs.

(1) Set of elements.
    Assume they are in an array.
(2) reduction operator
    (a) binary
    (b) associative

e.g. of binary operators, cf. Binary and Associative Operators

- multiply $*$
- min
- logical or $(a \parallel b)$
- bitwise and $(a \& b)$

Indeed, let f represent the input:

$$f \in K^N; \qquad N \in \mathbb{Z}^+$$
$$f(i) \in K; \qquad i \in \{0, 1 \dots, N-1\}$$

Serial Implementation of Reduce

$$S = 0,$$
$$\forall i \in \{0, 1, \dots N-1\} \iff S = \sum_{i=0}^{N-1} f(i)$$
$$S+ = f(i)$$

Step Complexity of Parallel Reduce

N	steps
2	1
4	2
8	3
2^n	n

The answer is $\log_2 N = n$. Let's prove this with induction.

Consider $N = 2^{n+1}$, number of things to compute.

$\frac{N}{2} = 2^n$. After 1 step (of binary operations), there are 2^n things left to compute.

By induction step, n steps are required.

$$\log_2 N = \log_2 2^{n+1} = n + 1$$

Done.

Reduction Using Global and Shared Memory

For a good explanation of *bitwise left shift* and *bitwise right shift* operators (and also for assignment), see What does a bitwise shift (left or right) do and what is it used for?. Also, a good implementation that converts from integers to bitwise representation is given in Converting integer to a bit representation

Keep in mind these are called *bitwise shift* operations (so we could look them up by this name).

Left shift.

$$\texttt{x << y} \iff x \cdot 2^y$$

Right shift.

$$\texttt{x >> y} \iff x/2^y$$

cf. reduce.cu of Lesson 3 Code Snippet for Udacity cs344

global memory reduce. $M_x \in \mathbb{Z}^+$, e.g. $M_x = 1024$
$k_x := i_x + M_x j_x \in \mathbb{Z}^+$
$t_{id} := i_x \in \{0 \dots M_x - 1\} \subset \mathbb{Z}^+$
$\forall\, s \in \{\frac{M_x}{2}, \frac{M_x}{2^2}, \frac{M_x}{2^3}, \dots\}$,
If $t_{id} < s$,

$$d_{in}[k_x] += d_{in}[k_x + s] \iff d_{\text{out}}[j_x] = \sum_{\substack{t_{\text{id}} < s \\ s \in \{\frac{M_x}{2}, \frac{M_x}{2^2}, \frac{M_x}{2^3}, \dots\}}} d_{\text{in}}[k_x + s] + d_{\text{in}}[k_x]$$

Finally,

$$d_{\text{out}}[j_x] = d_{\text{in}}[k_x]$$

shared memory reduce. $k_x := i_x + M_x j_x \in \mathbb{Z}^+$
$t_{id} := i_x \in \{0 \dots M_x - 1\} \subset \mathbb{Z}^+$
Load values into shared memory:

$$s_{\text{data}}[i_x] = d_{\text{in}}[k_x] \text{ where } s_{\text{data}} \in \mathbb{R}^{M_x}$$

$\forall\, s \in \{\frac{M_x}{2}, \frac{M_x}{2^2}, \frac{M_x}{2^3}, \dots\}$,
If $t_{id} < s$,

$$s_{\text{data}}[t_{\text{id}}] += s_{\text{data}}[t_{\text{id}} + s] \iff d_{\text{out}}[j_x] = \sum_{\substack{t_{\text{id}} < s \\ s \in \{\frac{M_x}{2}, \frac{M_x}{2^2}, \frac{M_x}{2^3}, \dots\}}} s_{\text{data}}[t_{\text{id}} + s] + s_{\text{data}}[t_{\text{id}}]$$

Finally,

$$d_{\text{out}}[j_x] = s_{\text{data}}[0]$$

Scan
Inputs to Scan

Inputs to scan. Like reduce,

- input array, $f \in K^N \qquad N \in \mathbb{Z}^+$
  $f(i) \in K \qquad i \in \{0 \dots, N-1\}$
- binary associative operator

new feature (not in reduce),

- identity element $[I \operatorname{op} a = a\,]$

cf. What Scan Actually Does

$$f \in K^N \xmapsto{\text{scan}} g \in K^N$$

inclusive scan

$$g(i) = \bigoplus_{j=0}^{i} f(j)$$

exclusive scan

$$g(i) = \begin{cases} \bigoplus_{j=0}^{i-1} f(j) & \text{if } i \geq 1 \\ 1 & \text{if } i = 0 \end{cases}$$

These definitions are worth repeating:

Definition 4 (Scan)**.** *Given*

- input array $f \in K^N$, *i.e.*

  $f \in K^N \qquad N \in \mathbb{Z}^+$
  $f[i] \in K \qquad i \in \{0, \dots, N-1\}$

  with K equipped with a
- binary associative operator
- identity element *$[I$ op $a = a]$*

Then for

$$f \xmapsto{scan} g$$

$$K^N \xrightarrow{scan} K^N$$

(5)

with ***inclusive scan*** *defined as*

(6)

$$\boxed{g(i) = \bigoplus_{j=0}^{i} f(j)}$$

and ***exclusive scan*** *defined as*

(7)

$$\boxed{g(i) = \begin{cases} \bigoplus_{j=0}^{i-1} f(j) & \text{if } i \geq 1 \\ 1 & \text{if } i = 0 \end{cases}}$$

cf. Hillis Steele Scan

Hillis/Steele inclusive scan. Let $f \in K^N \qquad N \in \mathbb{Z}^+$
$f(j) \in K \qquad j \in \{0 \dots N-1\}$
Consider step $i = 0$.
For $g_i \in K^N$, $\forall\, i \in \{0, \dots, \log_2 N - 1\}$,
Doing the first 3 steps,

$$g_0(j) = \begin{cases} f(j) + f(j - 2^0) & \text{if } j \geq 1 \\ f(j) & \text{if } j < 1 \end{cases}$$

$$g_1(j) = \begin{cases} g_0(j) + g_0(j - 2^1) & \text{if } j \geq 2^1 \\ g_0(j) & \text{if } j < 2^1 \end{cases}$$

$$g_2(j) = \begin{cases} g_1(j) + g_1(j - 2^2) & \text{if } j \geq 2^2 \\ g_1(j) & \text{if } j < 2^2 \end{cases}$$

Thus

$$g_i(j) = \begin{cases} g_{i-1}(j) + g_{i-1}(j - 2^i) & \text{if } j \geq 2^i \\ g_{i-1}(j) & \text{if } j < 2^i \end{cases}$$

Now do the following induction cases:

$$g_{\log_2 N}(0) = f(0)$$
$$g_{\log_2 N}(1) = g_0(1) = f(1) + f(0)$$
$$g_{\log_2 N}(2) = g_1(2) = g_0(2) + g_0(0) = f(2) + f(1) + f(0)$$
$$g_{\log_2 N}(3) = g_1(3) = g_0(3) + g_0(1) = f(3) + f(2) + f(1) + f(0)$$

So generalize to the induction case:

$$g_{\log_2 N}(j) = \bigoplus_{i=0}^{j} f(i)$$

Then one would check the induction step for a number of cases, whether j was greater or smaller or equal to $2^{\log_2 N}$.

To summarize for the Hillis/Steele inclusive scan,

(8)
$$\boxed{\begin{array}{c} \forall i \in \{0, 1, \dots, \log_2 N | 2^i < N\} \\ g_{i-1}(j) = f(j) \qquad \forall j \in \{0 \dots N-1\} \\ g_i(j) = \begin{cases} g_{i-1}(j) + g_{i-1}(j - 2^i) & \text{if } j \geq 2^i \\ g_{i-1}(j) & \text{if } j < 2^i \end{cases} \end{array}}$$

i.e.

(9)
$$\boxed{\begin{array}{c} \forall i \in \{0, 1, \dots, \lfloor \log_2 N \rfloor | 2^i < N\} \\ g_{-1}(j) = f(j) \qquad \forall j \in \{0 \dots N-1\} \\ g_i(j) = \begin{cases} g_{i-1}(j) + g_{i-1}(j - 2^i) & \text{if } j \geq 2^i \\ g_{i-1}(j) & \text{if } j < 2^i \end{cases} \end{array}}$$

Implementation is CUDA C/C++: consider $k_x := i_x + M_x j_x \in \{0, 1, \dots, L_x\}$, $L_x \in \mathbb{Z}^+$, with the formula

$$N_x := \frac{L_x + M_x - 1}{M_x}$$

that'll give us `blockDim.x`.

$\forall i \in \{0, 1 \dots \log_2 k_x\}$ (or $k_x < 2^i$),

$$g_i(k_x) = g_{i-1}(k_x) + g_{i-1}(k_x - 2^i)$$

For *(inclusive) scan*,
step : $O(\log n)$
work : $O(n^2)$

For *Hillis/Steele* ,
step : $O(\log n)$
work : $O(n \log n)$

(more step efficient)
For B
cf. Inclusive Scan Revisited, Hillis Steele vs Blelloch Scan,Hillis Steele Scan,

11.6. **Blelloch scan.** cf. Blelloch Scan

11.6.1. *Blelloch scan, reduce, 1st part, up sweep.* My development: consider the most basic cases. So for $f \in K^N$,

$$\begin{array}{cc} i \in \{0, 1, \dots N-1\} & N = 8 \qquad i \in \{0, 1, \dots, 8-1 = 7\} \\ 2j, \qquad j \in \{1, 2, \dots, \lfloor N/2 \rfloor\} & 2j \in \{2, 4, 6, 8\},\ j \in \{1, 2, 3, 4\} \\ i = 2j - 1 & i = 2j - 1 \in \{1, 3, 5, 7\} \\ 2^2 j - 1, \qquad j \in \{1, 2, \dots, \lfloor N/2^2 \rfloor\} & j \in \{1, 2\}, \qquad i \in \{3, 7\} \\ 2^3 j - 1, \qquad j \in \{1, 2, \dots, \lfloor N/2^3 \rfloor\} & j \in \{1\}, \qquad i \in \{7\} \end{array}$$

Then $\forall k = \{1, 2, \dots, \lfloor \log_2 N \rfloor\}$,

$$k = 1, \qquad f_{\text{out}}[i] = \begin{cases} f[i] + f[i-1] & \text{if } i = 2j - 1,\ j \in \{1, 2, \dots \lfloor N/2 \rfloor\} \\ f[i] & \text{otherwise} \end{cases}$$

$$k = 2, \qquad f_{\text{out}}[i] = \begin{cases} f[i] + f[i - 2^{(k-1)}] & \text{if } i = 2^2 j - 1,\ j \in \{1, 2, \dots \lfloor N/2^2 \rfloor\} \\ f[i] & \text{otherwise} \end{cases}$$

$$k \quad , \qquad f_{\text{out}}[i] = \begin{cases} f[i] + f[i - 2^{(k-1)}] & \text{if } i = 2^k j - 1,\ j \in \{1, 2, \dots \lfloor N/2^k \rfloor\} \\ f[i] & \text{otherwise} \end{cases}$$

Thus

Definition 5 (Blelloch scan: 1st part that's reduce, i.e. up sweep)**.**

$$\forall k \in \{1, 2, \dots \lfloor \log_2 N \rfloor\}$$

(10)
$$f_{out}[i] = \begin{cases} f[i] + f[i - 2^{(k-1)}] & \textit{if } i = 2^k j - 1,\ j \in \{1, 2, \dots \lfloor N/2^k \rfloor\} \\ f[i] & \textit{otherwise} \end{cases}$$

$$f \longmapsto f_{out}$$

for

$$K^N \longrightarrow K^N$$

11.6.2. *Blelloch scan, down sweep part.* Now consider the down sweep. First, importantly, set the "last" element in the array, entry number $\lfloor \log_2 N \rfloor$, N being the length of the array, to the identity element of K.

I'll present the first few induction steps explicitly, and the general form can be guessed from there.

$$\forall k \in \{\lfloor \log_2 N \rfloor, \lfloor \log_2 N \rfloor - 1, \dots, 2, 1\}$$

If $i = 2^k j - 1, \quad j \in \{1, 2, \dots, \lfloor N/2^k \rfloor\}$, for e.g. $k = \lfloor \log_2 N \rfloor$,

$$f_{\text{out}}[i - 2^{(k-1)}] = f[i]$$
$$f_{\text{out}}[i] = f[i] + f[i - 2^{(k-1)}]$$

For $k = \lfloor \log_2 N \rfloor - 1$, e.g. $k = 3 - 1 = 2$, then for this exmaple,
e.g. $i = 2^k j - 1 = 4j - 1$; $j \in \{1, 2\}$, $i \in \{3, 7\}$,

$$f_{\text{out}}[i] = f[i] + f[i - 2]$$
$$f_{\text{out}}[i - 2] = f[i]$$

Thus, in general,

Definition 6 (Blelloch scan: down sweep step)**.**

$$\forall k \in \{\lfloor \log_2 N \rfloor, \lfloor \log_2 N \rfloor - 1, \dots, 2, 1\}$$

$$(11) \qquad \begin{cases} f_{out}[i] := f[i] + f[i - 2^{(k-1)}] & \text{if } i = 2^k j - 1 \text{ and } j \in \{1, 2, \dots, \lfloor N/2^k \rfloor\} \\ f_{out}[i - 2^{(k-1)}] := f[i] & \\ f_{out}[i] := f[i] & \text{otherwise} \end{cases}$$

cf. Problem Set 3

Histogram Equalization.

(1) Map
(2) Reduce
(3) Scatter
(4) Scan

11.7. **Reduce, Parallel Reduction.** I will expound upon the excellent article from Mark Harris, "Optimizing Parallel Reduction in CUDA". cf. *Optimizing Parallel Reduction in CUDA*, Mark Harris : `http://developer.download.nvidia.com/compute/cuda/1.1-Beta/x86_website/projects/reduction/doc/reduction.pdf`

11.8. **Unit 4.**

11.9. **Unit 5: Lesson 5 - Optimizing GPU Programs.**

11.9.1. *1. Quiz: Optimizing GPU Programs.* Quiz: principals of efficient GPU programming.
We want to

- decrease time spent on memory operations (we want to do more math; $\frac{\text{math}}{\text{memory}}$
- coalesce global memory accesses
- avoid thread divergence

We don't want to necessarily

- decrease arithmetic intensity
- do fewer memory operations per thread
- move all data to shared memory

Reasons; we want to *maximize* arithmetic intensity!

11.9.2. *22. Quiz: Tiling.* Tiling
For the kernel `transpose_parallel_per_element_tiled(float in[], float out[])`, (cf. transpose cu, we desire

$$B = A^T \text{ or } B_{ij} = A_{ji} \quad \forall i = 0, 1, \dots L_x - 1, \forall j = 0, 1 \dots L_y - 1$$

Consider $M_x = M_y = M$, with

$$M_x \equiv \texttt{blockDim.x} \quad i_x \equiv \texttt{threadIdx.x} \quad j_x \equiv \texttt{blockIdx.x}$$
$$M_y \equiv \texttt{blockDim.y} \quad i_y \equiv \texttt{threadIdx.y} \quad j_y \equiv \texttt{blockIdx.y}$$

and so

$$i := i_x + j_x M_x \in \{0, 1, \dots N_x M_x - 1\} \quad j := i_y + j_y M_y \in \{0, 1, \dots N_y M_y - 1\}$$

For the *shared* indices,

$$i_{\text{sh}} = i_x \in \{0, 1 \dots M_x - 1\}$$
$$j_{\text{sh}} = i_y \in \{0, 1 \dots M_y - 1\}$$

$$A_{\text{sh}} \in \mathbb{R}^{K^2} = \mathbb{R}^{K \times K} \in \texttt{__shared__}$$

Note that $\mathbb{R}^{K^2} = \mathbb{R}^{K \times K} \in$ `__shared__` can be a 1-dim. (or 2-dim., multidimensional!) array in shared memory (!!!).
If $A \in \text{Mat}_{\mathbb{R}}(L_x, L_y)$, $B = A^T \in \text{Mat}_{\mathbb{R}}(L_y, L_x)$.
Globally,

$$B(j, i) = B(i_y + j_y M_y, i_x + j_x M_x) = A(i, j) = A(i_x + j_x M_x, i_y + j_y M_y)$$

for

$$A_{\text{sh}} \in \mathbb{R}^{M_x M_y}$$
$$A_{\text{sh}}(j_{\text{sh}}, i_{\text{sh}}) = A_{ij}$$
$$A_{\text{sh}}(i_{\text{sh}}, j_{\text{sh}}) = A_(j_{\text{sh}} + j_x M_x, i_{\text{sh}} + j_y M_y)$$
$$\implies B(i_x + j_y M_y, i_y + j_x M_x) := A_{\text{sh}}(i_{\text{sh}}, j_{\text{sh}}) = A(j_{\text{sh}} + j_x M_x, i_{\text{sh}} + j_y M_y)$$

11.9.3. *Occupancy.* cf. Occupancy Part 1
Each SM (streaming multi-processor) has a limited number of

- thread blocks (so there's a maximum number of thread blocks, e.g. 8)
- threads (so there's a maximum number of threads, e.g. 1536-2048)
- registers for all threads (every thread takes a certain number of registers, and there's a maximum number of registers for all the threads, e.g. 65536)
- bytes of shared memory

cf. Quiz: Occupancy Part 2
Compile and run `deviceQuery_simplified.cpp` (can be found in Lesson 5 Code Snippets).
Look at

- Total amount of shared memory per block
- Maximum number of threads per multiprocessor
- Maximum number of threads per block

For Luebke's laptop, it's 49152 bytes, 2048, 1024 respectively, and 65536 total registers available per block.

11.9.4. *Thread Divergence.* cf. 38. Quiz: Switch Statements and Thread Divergence, of Lesson 5 - Optimizing GPU Programs
Threads in warp is 32 (check hardware, usually this is the case for modern GPUs). Only 32 threads in a warp, i.e. 2^5.
So for something like

```
switch (threadIdx.x % 64) { case 0 ... 63}
kernel<<<1024,1>>>();
```

There are only 32 threads in a warp, and so only a maximum slowdown of "32" (corresponding to 32 threads, each starting and stopping sequentially, i.e. $0- > 1- > 2- > 3- > ...- > 30- > 31$, 1 after the other. The other 32 threads in 64 threads will be assigned to another warp.
For

```
switch (threadIdx.y) { case 0 ... 31}
kernel<<<64x16,1>>>();
```

or, i.e. `kernel<<<(1,1),(64,16)>>>();`, a 64×16 thread block, i.e. $(M_x, M_y) = (64, 16) = (2^6, 2^4)$, CUDA launches thread warps along x first, and then along y.
CUDA assigns thread IDs to warps:

- x varies fastest
- y varies slower
- z varies slowest

warp assignment, consider $(M_x, M_y, M_z) \equiv$ number of threads in a single thread block, in each dimensional direction.
Consider

$$i_x \in \{0, \dots, M_x - 1\}$$
$$i_y \in \{0, \dots, M_y - 1\}$$

If $M_x/2^5 \in \mathbb{Z}^+$, i.e. $M_x/2^5 \geq 1$, then threads in x-direction are in warps that branch (go to same case in if-else, or switch, case statement in kernel) to same case.

11.9.5. *41. Quiz: thread Divergence in the Real World Part 1.* cf. Thread Divergence in the Real World Part 1

Example: Operating on a 1024 ×1024 image, with special handling of pixels on the boundary (boundary condition),

The maximum branch divergence of any warp (32 threads) is 2-way.

That's because, if you consider a 1 pixel deep boundary condition at the (absolute) boundary of the grid, then for most 32-thread warps in the middle, there's no divergence. At most a warp will include 1 boundary pixel, horizontally. Warp that lies completely on a vertical boundary will call the boundary condition, all of those pixels. So there's no divergence there either.

11.9.6. *43. Thread Divergence in the Real World Part 3.* cf. Thread Divergence in the Real World Part 3

- Be aware of branch divergence
- "But don't freak out about it." -Luebke

Reducing branch divergence (general principles)

- Avoid branchy code
  - consider if adjacent threads will likely take different paths
- Beware of large imbalance in thread workloads

11.9.7. *45. Host-GPU Interaction.* cf. Host-GPU Interaction of Lesson 5- Optimizing GPU Programs

-
-

11.10. **Streams.** Stream is a sequence of operations that'll execute in order.

Type is `cudaStream_t`

11.10.1. *Problem Set 5.* 2. One Basic Strategy

A. Sort the Data, into bins, sort the input data, into coarse bins

B. use threads, compute local histogram, use each thread block, to compute local histogram

c. concatenate

e.g.

Consider

$J =$ total number of coarse bins, $J \in \mathbb{Z}^+$.

$j \in \{0 \dots J-1\}$, e.g. $J = 10$.

$K =$ total number of bins,

$k \in \{0 \dots K-1\}$, e.g. $K = 100$

The condition is that $J < K$

For each thread block,

11.10.2. *Parallel radix sort.* cf. look at, in the directory http://www.compsci.hunter.cuny.edu/~sweiss/course_materials/csci360/lecture_notes/ for the parallel radix sort implementation called `radix_sort_cuda.cc`.

Let $M_x \in \mathbb{Z}^+$ represent

$$M_x \leftrightarrow \texttt{blockDim.x} \tag{12}$$

$$j_x \in \{0, 1, \dots M_x - 1\} \leftrightarrow \texttt{blockIdx.x} \tag{13}$$

$\forall\, j_x \in \{0, 1 \dots, M_x - 1\}$, j_x denote which thread block we are in.

j_x also corresponds to the so-called coarse bin id.

11.11. **Lesson 6.1 - Parallel Computing Patterns Part A.**

11.11.1. *parallel All Pairs N-body.* cf. Quiz: All Pairs N-Body

From Arnold, Kozlov, Neishtadt, and Khukhro (2006) [27], given N-bodies $\{1, 2 \dots N\} \subset \mathbb{Z}^+$.

$$(\mathbf{r}_1, M_1), (\mathbf{r}_2, M_2), \dots (\mathbf{r}_N, M_N) \qquad \mathbf{r}_i \in \mathbb{R}^d$$

With the force acting upon the ith body ("destination") due to the jth body ("source"), $\mathbf{F}_{ij}$,

$$\mathbf{F}_{ij} = -\frac{\gamma M_i M_j}{|\mathbf{r}_{ij}|^3}\mathbf{r}_{ij} \qquad \text{with } \mathbf{r}_{ij} := \mathbf{r}_j - \mathbf{r}_i \tag{14}$$

$N(N-1)$ are the number of (unordered) pairs, for

$$\mathbf{F}_i = \sum_{i \neq j} \mathbf{F}_{ij} \tag{15}$$

with $\mathbf{F}_i$ denoting the force acting upon ith body, $N-1$ computations

$$\mathbf{F}_i = \sum_{i \neq j} \frac{-\gamma M_i M_j}{|\mathbf{r}_{ij}|^3}\mathbf{r}_{ij} = -\gamma M_i \sum_{i \neq j} \frac{M_j}{|\mathbf{r}_{ij}|^3}\mathbf{r}_{ij} \tag{16}$$

Consider another pairwise computation:

$$W(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \lambda_i \lambda_j K(X^{(i)}, X^{(j)}) \tag{17}$$

Define, for (writing) convenience, $f_1(\lambda) = -\frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \lambda_i \lambda_j K(X^{(i)}, X^{(j)})$.

cf. Quiz: How To Implement Dense N-Body as Simply As Possible

For a single body, say Ith body $(\mathbf{r}_I, M_I)$, consider $\mathbf{F}_i = \sum_{i \neq j} \mathbf{F}_{ij} \qquad \forall\, i = 1, \dots N$.

$j = I$ once, $\forall\, i = 1, \dots N$ and $i \neq I$ (Ith body as a "source") $N-1$ times.

$i = I$ once, in $\mathbf{F}_I = \sum_{i \neq j} \mathbf{F}_{Ij}$, we'll need $(\mathbf{r}_I, M_I)$, to calculate $\mathbf{F}_{Ij}$, $N-1$ times.

If all N^2 force computations, go to global memory, the number of times we'd fetch each element, as a function of N is $\boxed{2(N-1)}$.

cf. Quiz: Dividing N by N Matrix Into Tiles

Consider $N \times N$ matrix A, $A \in \mathrm{Mat}_{\mathbb{K}}(N_x, N_y)$.

Consider dividing it by tiles, calculate each tile, a tile of $P_x \times P_y$.

So there are $\frac{N_x N_y}{P_x P_y}$ tiles overall.

For this particular problem, and in general, a problem involving pairwise *combination* out of a given set, say N bodies, consider N^2 possible pairs total. Then this all N-body algorithm idea is to transform all these combinations N^2, into a matrix, and so $N_x = N_y = N$, $P_x = P_y = P$.

$$\mathbf{F}_i = \sum_{i \neq j}^{N} \mathbf{F}_{ij} = \sum_{t=0}^{\frac{N}{P}-1} \sum_{\substack{j=0 \\ Pt+j \neq i}}^{P-1} \mathbf{F}_{i(Pt+j)} \equiv \sum_{t=0}^{\frac{N}{P}-1} \mathbf{F}_i^{(t)} \tag{18}$$

$2P$ fetches per tile, instead of $2P^2$, if $\forall$ tile, store the P elements into shared memory.

cf. Using on P threads

Consider this computation:

$$\mathbf{F}_i^{(t)} = \sum_{\substack{j=0 \\ pt+j \neq i}}^{P-1} \mathbf{F}_{i(Pt+j)} \qquad i = 0 \dots N-1, t = 0, 1, \dots \frac{N}{P} - 1 \tag{19}$$

For

$$i \iff i_x + M_x j_x, \qquad \begin{aligned} i_x &\equiv \texttt{threadIdx.x} \in \{0, 1 \dots M_x - 1\} \\ j_x &\equiv \texttt{blockIdx.x} \in \{0, 1, \dots N/M_x - 1\} \\ M_x &= P \end{aligned}$$

So $\forall\, (i_x, j_x)$, (i_x, j_x) representing a single thread, this single thread will be responsible for computing

$$\mathbf{F}_i^{(t)} = \sum_{j=0}^{P-1} \mathbf{F}_{i(Pt+j)}$$

, as opposed to only computing $\mathbf{F}_{i(Pt+j)}$ only.

The pseudocode that Owens gives in Using On P Threads, Lesson 6.1 is the following:

```
__device__ float3
tile_calculation(Params myParams, float3 force) {
  int i;
  extern __shared__ Params[] sourceParams;
  for (i=0; i<blockDim.x; i++) {
    force += bodyBodyInteraction(myParams, sourceParams[i]);
  }
  return force;
}
```

11.12. Lesson 7.1 Additional Parallel Computing.

11.12.1. *4. Quiz: data Layout Transformation.* 1. Data layout transformation

Quiz: Global memory coalescing is important because:
(modern) DRAM systesm transfer large chunks of data per transaction.

11.12.2. *Additional Data Transformation Methods.*

11.12.3. *6. Quiz: Burst Utilization.* Burst Utilization

As a reminder of what the difference between **Array of Structures** and **Structure of Arrays** are illustrated in these examples:

Array of Structures:

```
struct foo {
        float a;
        float b;
        float c;
        float d;
} A[8];
```

Structure of Arrays

```
struct foo {
        float a[8];
        float b[8];
        float c[8];
        float d[8];
} A;
```

Quiz: which layout will perform better on these codes?

```
int i = threadIdx.x;
A[i].a++;
A[i].b += A[i].c * A[i].d;
```

and

```
int i = threadIdx.x;

A.a[i]++;
A.b[i] += A.c[i] + A.d[i];
```

Array of Structures (AoS) Example:

$$\begin{aligned} \{0, 1, \dots M_x - 1\} &\to \mathbb{R}^4 \\ i_x \mapsto (a_{i_x}, b_{i_x}, c_{i_x}, d_{i_x}) &\in \mathbb{R}^4 \equiv (\texttt{float})^4 \end{aligned}$$

In general,

$$\begin{aligned} \{0, 1, \dots L-1\} &\to \mathbb{K}^d \\ i \in \mathbb{Z} &\mapsto A(i) \in \mathbb{K}^d \end{aligned} \qquad (AoS) \tag{20}$$

with field $\mathbb{K} = \mathbb{R}, \mathbb{C}, \mathbb{Z}^+ \dots$ i.e. $\mathbb{K} \in$ **Type** so that $\mathbb{K} =$ `float, int` .

So that

$$(\{0, 1, \dots L-1\} \to \mathbb{K}^d) \to \mathbb{Z} \times \mathbb{K}^d \xrightarrow{\text{flatten}} \prod_{i=0}^{L-1} \mathbb{K}^d$$

$$(i \mapsto A(i)) \mapsto (i, (A(i))^\mu) \mapsto ((A(i))^1, (A(i))^2, \dots (A(i))^d)$$

$$(i, \mu) \mapsto \mu + id$$

So for $t \equiv$ thread index $= i$,
$\mu + td$ address is "strided" by d.

Structure of Arrays (SoA)

Example:

$$\{0, 1, \dots M_x - 1\} \times \{0, 1, \dots M_x - 1\} \times \{0, 1, \dots M_x - 1\} \times \{0, 1, \dots M_x - 1\} = \{0, 1, \dots M_x - 1\}^4$$

In general

$$\prod_{\alpha = \{a, b, \dots\}} \{0, 1, \dots L^{(\alpha)} - 1\} \qquad (SoA) \tag{21}$$

So that

$$\begin{aligned} \prod_{\alpha = \{a, b, \dots\}} \{0, 1, \dots L^{(\alpha)} - 1\} &\overset{\text{flatten}}{\to} \{0, 1 \cdots \prod_\alpha L^{(\alpha)}\} \\ i^{(\alpha)} \xrightarrow{\text{flatten}} &\left(\sum_{\alpha' < \alpha} L^{(\alpha')} \right) + i^{(\alpha)} \end{aligned} \tag{22}$$

So for $\forall\, i_x \equiv$ `threadIdx.x` , or in general $\forall\, i \in \{-, 1, \dots M_x N_x - 1\} = i_x + j_x M_x =$ `threadIDx.x+blockDim.x * blockIdx.x`,
Want to consider,

$$i^{(\alpha)} + \sum_{\alpha' < \alpha} L^{(\alpha')} \xrightarrow{f} i$$

$$f(i^{(\alpha)} + \sum_{\alpha' < \alpha} L^{(\alpha')} = i^{(\alpha)} + \sum_{\alpha' < \alpha} L^{(\alpha')} = i$$

11.13. Final for Udacity cs344.

11.13.1. *Quiz: Final - Question 1; programming exercise for stencil operation, from global memory , to shared memory, reducing global memory transfers.* https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1388615750923

11.13.2. *Quiz: Final - Question 3; strongly Compute-bound, 2x SMs.* cf. https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1385089510923

Performance as function of number of blocks (launched), curve C.

2x SMs, 2 times the performance.

strongly **Compute-bound** is proportional to SMs (cores)

11.13.3. *Quiz: Final - Question 4, Compute-bound problem, but 2x Memory Bandwidth, no effect.* https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1389922470923

Compute-bound, but 2x Memory Bandwidth, no effect on performance.

11.13.4. *Quiz: Final - Question 5; Memory bandwidth-bound, less SM, no effect.* https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1387579670923

Memory bandwidth-bound, but 1/2 SMs, no effect on performance.

11.13.5. *Quiz: Final - Question 6; Memory bandwidth-bound.* https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1388556450923

Memory bandwidth; 2x SMs *and* 1/2 memory bandwidth, because of half of memory bandwidth, performance is 1/2.

11.13.6. *Quiz: Final - Question 7.* Setup: Blurring on 2-dim. image; each pixel is a 32 bit single-precision floating point number. 5x5 kernel. 25 multiplies, 24 adds to compute each pixel output.

cf. https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1421890260923#

Second question:

What is the ratio (expressed as a percentage) between kernel that uses more memory bandwidth and the kernel that uses less?

For both parts of this question, here's how to approach it.

Consider a stencil of radius RAD $\equiv R = 2$. So the filter-width in 1-dimension for this 2-dimensional stencil is

$$\text{filterwidth} \equiv W = 2 \cdot \text{RAD} + 1 = 5$$

W^2 multiples, and $W^2 - 1$ adds.

$\forall\, k \in K$, a (single) pixel of greyscale image, $k \in \{0, 1, \dots N-1\}$,

Consider the threads per block configuration (i.e. a single thread block).

Let $M \equiv M_x M_y = 1024$ in both, either, case, i.e. total number of threads per block is 1024.

Consider $(M_x, M_y) = (1024, 1)$ vs. ,

$(M_x, M_y) = (32, 32)$ vs.

Corresponding to

$$\texttt{dim3} M_i(M_x, M_y)$$

for

$$\text{smooth } \texttt{<<< , M\_i >>>} (\texttt{v\_new}, v);$$

The big idea is that we have a tile of input data we want to load into 2-dimensional shared memory. It is of size

$$(M_x + 2R)(M_y + 2R) \tag{23}$$

Compare this quantity to when $(M_x, M_y) = (1024, 1)$ vs. when it's $(32, 32)$.

$$(1024 + 4)(1 + 4)/(32 + 4) * (32 + 4) = \boxed{3.966}$$

This is the ratio between the kernel with more memory bandwidth and the kernel with less memory bandwidth.

Note, in the Udacity forum, Final: Updated Questions 7 & 8, jlmayfield gave the hint to consider those stencil halo cells or what I call the radius R of the stencil, as this stencil is $5x5$.

11.13.7. *Quiz: Final - Question 8.* cf. Quiz:Final - Question 8

32×32 block, but consider

- 4x storage per SM
- 4x registers
- 4x shared memory

now 64 x 64 since, in each dimension, 2x the global memory available $(2 \times 2 = 4)$.

Express, in decimals, what the speedup is on this new GPU.

Here's how I approached this problem:

so in a single thread block, we can have a much larger tile with inputted data to reside in shared memory and so we can process a much larger tile of inputted data in a single thread block:

$$(M_x + 2R)(M_y + 2R) = (64 + 2 * 2) * (64 + 2 * 2) = 4624 \tag{24}$$

instead of when $(M_x, M_y) = (32, 32)$.

Not only that 4x of shared memory, 4x storage per SM, but also *4x registers*. More registers for the threads, more a thread can be doing something and not have to wait for another thread (or even to hop on to another thread if the thread itself doesn't have enough registers). More throughput, by 4x.

$$(64 + 2R)(64 + 2R) * 4/(32 + 2R)(32 + 2R) = \boxed{\text{14.27 or 14.0}} \tag{25}$$

There was confusion about Final - Question 8 in the Udacity Forums: Final - Question 8

11.13.8. *Quiz: Final - Question 9.* cf. https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1389566540923

If I want to run the maximum number of threads possible that are all resident on an SM (streaming multiprocessor) at the same time, each thread must use no more than, *how many*, registers?

$$\boxed{\text{32 registers}} \tag{26}$$

This is obtained by looking at http://en.wikipedia.org/wiki/CUDA (this was given) and look for

- "Maximum number of resident threads per multiprocessor": 2048 for Compute ability (version) 3.5.
- "Number of 32-bit registers per multiprocessor". For Compute capability 3.5, it's 64K. For 5.0+ it's 64K (again).

Then take number of 32-bit registers per multiprocessor / maximum number of resident threads per multiprocessor = number of registers to get that max. number of resident threads

$$64000/2048 \sim 32$$

Final Exam Question 9 confusion

https://discussions.udacity.com/t/final-exam-question-9-confusion/88637

11.13.9. *Quiz: Final - Question 10, bytes of shared memory per thread.* https://classroom.udacity.com/courses/cs344/lessons/2133758814/concepts/1388977230923

Note that we are asking for the number of bytes of shared memory per thread and not the number of bytes available to each thread.

Given we want max. no. of threads that are all resident on an SM (streaming processor), what is the number of bytes of shared memory, per thread?

I took, cf. https://en.wikipedia.org/wiki/CUDA,

Maximum number of resident threads per multiprocessor = 2048/ Maximum amount of shared memory per multiprocessor $= 48KB$ (for compute capability 3.5) = and taking the inverse and multiplying by 1000, $23 \sim 24$. For compute capability 5.3, 6.2, it's $64kb$.

11.13.10. *Quiz: Final - Question 11.* Maximum number of blocks resident on the same SM, what is the maximum number of threads / block that we can have?

This is obtained by looking at http://en.wikipedia.org/wiki/CUDA (this was given) and look for

- "Maximum number of resident blocks per multiprocessor": 16 for Compute ability (version) 3.5.
- "Maximum number of resident threads per multiprocessor": 2048 for Compute ability (version) 3.5.

The answer is

$$\boxed{2048/16 = 128} \tag{27}$$

Since we take max. number of resident threads per SM, 2048, but the condition is using the max. number of resident blocks per SM, so, divide by 16.

11.13.11. *Quiz: Final - Question 12; Fast "compact" primitive.* Consider the *Fast " compact" primitive.*

(1) sum up the no. of "T" flags

e.g. warp of size 4 (only to make it easy to understand), (4 warps of 4 flags each), e.g.

$$TFTF/FTFT/TTTT/tfff$$

(1)

$$2/2/4/1$$

Let warp size $W = 2^n$; $n \in \mathbb{Z}^+$, e.g. $n = 5$ for $W = 32$.
Given shared memory $s \in (\mathbb{Z}^+)^W$,

$$\forall\, i \in \{W/2, W/4, \dots | i > 0\}$$

For given `ARRAY_SIZE = 32 =` M_x so $M_x = W$ (thread block size is equal to warp size in this case).

Keep in mind, even in a `__device__` function, we've got **access** to

$$i_x \equiv \texttt{threadIdx.x} \in \{0, 1 \dots, M_x - 1\}$$

(and so on).

$i_x < i$, i.e. $\forall i_x < i$ and $i_x \in \{0, 1 \dots M_x - 1\}$

$$s[i_x] = s[i_x] + s[i_x + i]$$

e.g. $i = W/2$

$$i_x \in \{0, 1 \dots, \frac{W}{2} - 1\}$$

$i = W/4$

$$i_x \in \{0, 1 \dots \frac{W}{4} - 1\}$$

$$s[0] = s[0] + s[\frac{W}{2}] + s[0 + \frac{W}{4}]$$

$$s[i_x] = s[i_x] + s[i_x + 1] \Longrightarrow \vdots$$

$$s[\frac{W}{4} - 1] = s[\frac{W}{4} - 1] + s[\frac{3W}{4} - 1] + s[\frac{W}{2} - 1]$$

So by induction, for $i = 1$, $i_x = 0$.

$$s[0] = \sum_{j=0}^{W-1} s[j]$$

which is the desired result (!!!).

Code answer uploaded here: Q12warpreduce shared.cu or here warpreduce.cu

12. Pointers in C; Pointers in C categorified (interpreted in Category Theory)

Suppose $v \in$ ObjData, category of data **Data**,
e.g. $v \in$ Int $\in$ ObjType, category of types Type.

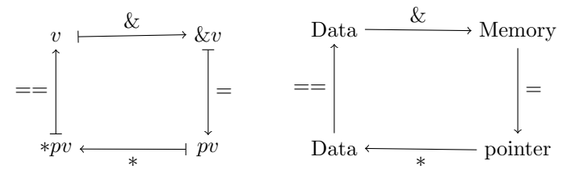
$$\text{Data} \xrightarrow{\&} \text{Memory}$$
$$v \overset{\&}{\mapsto} \&v$$

with address $\&v \in$ Memory.

With
assignment $pv = \&v$,

$pv \in$ Objpointer, category of pointers, pointer
$pv \in$ Memory (i.e. not $pv \in$ Dat, i.e. $pv \notin$ Dat)

$$\text{pointer } \ni pv \overset{*}{\mapsto} *pv \in \text{Dat}$$



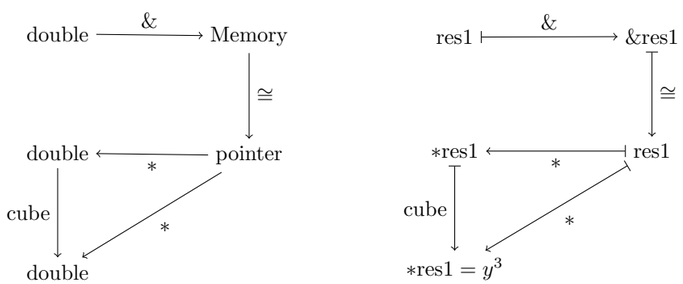
Examples. Consider `passfunction.c` in Fitzpatrick [16].
Consider the type `double`, `double` $\in$ ObjTypes.
fun1, fun2 $\in$ MorTypes namely
fun1, fun2 $\in$ Hom(double, double) $\equiv \text{Hom}_{\text{Types}}$(double, double)
Recall that

$$\text{pointer} \xrightarrow{*} \text{Dat}$$
$$\text{pointer} \xrightarrow{\&} \text{Memory}$$

$*, \&$ are functors with domain on the category pointer.

Pointers to functions is the "extension" of functor $*$ to the codomain of MorTypes:

$$\text{pointer} \xrightarrow{*} \text{MorTypes}$$
$$\text{fun1} \overset{*}{\mapsto} *\text{fun1} \in \text{Hom}_{\text{Types}}(\text{double}, \text{double})$$



It's unclear to me how `void cube` can be represented in terms of category theory, as surely it cannot be represented as a mapping (it acts upon a functor, namely the $*$ functor for pointers). It doesn't return a value, and so one cannot be confident to say there's explicitly a domain and codomain, or range for that matter.

But what is going on is that

$$\text{pointer , double , pointer} \xrightarrow{\text{cube}} \text{pointer , pointer}$$

$$\text{fun1}, x, \text{res1} \overset{\text{cube}}{\mapsto} \text{fun1}, \text{res1}$$

s.t. $*\text{res1} = y^3 = (*\text{fun1}(x))^3$

So I'll speculate that in this case, `cube` is a functor, and in particular, is acting on $*$, the so-called deferencing operator:

$$\begin{array}{ccc} \text{pointer} \xrightarrow{*} \text{float} \in \text{Data} & \xrightarrow{\text{cube}} & \text{pointer} \xrightarrow{\text{cube}(*)} \text{float} \in \text{Data} \\ \text{res1} \overset{*}{\mapsto} *\text{res1} & & \text{res1} \overset{\text{cube}(*)}{\mapsto} \text{cube}(*\text{res1}) = y^3 \end{array}$$

cf. Arrays, from Fitzpatrick [16]

$$\text{Types} \xrightarrow{\text{declaration}} \text{arrays}$$

If $x \in$ Objarrays,

$$\&x[0] \in \text{Memory} \xrightarrow{==} x \in \text{ pointer (to 1st element of array)}$$

cf. Section 2.13 Character Strings from Fitzpatrick [16]

```
char word[20] = ''four''
char *word = ''four''
```

cf. C++ extensions for C according to Fitzpatrick [16]

- simplified syntax to pass by reference pointers into functions
- inline functions
- variable size arrays

```
int n;
double x[n];
```

- complex number class

13. Summary of Udacity cs344 concepts

13.1. **Histogram.** Consider given input values (of observations).
For n observations,
For $i \in \{1, 2, \dots n\} \subset \mathbb{Z}^+$, $\equiv i\{0, 1, \dots n-1\} \subset \mathbb{Z}^+$.

Consider $x[i] \in B \subset \mathbb{K}$, so $x \in B^n \subset \mathbb{K}^n$.

e.g. $\mathbb{K} = \mathbb{R}$, B is a subset that we can make K bins out of, i.e.

$$B = \prod_{j=1}^{K} B_j \equiv \prod_{j=0}^{K-1} B_j$$

e.g. B is bounded interval of $\mathbb{R}$, i.e. $\max B < \infty$
$\min B > -\infty$

Then the *histogram* H is mapping from bins to *number* of observations, i.e.

$$\begin{aligned} &H : \{1, 2, \dots K\} \to \mathbb{N}^K \\ &H(j) \in \mathbb{N} = \{0, 1, \dots\} \end{aligned} \tag{28}$$

s.t.

$$\boxed{n = \sum_{j=1}^{K} H(j)} \tag{29}$$

with n being the total number of observations.

In the implementation of $x : \{1, 2, \dots n\} \to B \subset \mathbb{K}$,
this was simplified to

$$x : \{1, 2, \dots n\} \to \{1, 2, \dots K\} \subset \mathbb{Z}$$

meaning each observation value is *itself* which bin the observation belongs to.

Otherwise, a separate "binning" operation is needed:

$$\begin{array}{lll} x : \{1, 2, \dots n\} \to & B \subset \mathbb{K} \to & \{1, 2, \dots K\} \\ i \mapsto & x[i] \mapsto \text{if } B_{j-1} \le x[i] < B_j, & \text{then return } j \end{array} \tag{30}$$

13.1.1. *Histogram implementations; histogram references.* ernestyalumni/cs344/Problem Sets/Problem Set 5/histogram/ https://github.com/ernestyalumni/cs344/tree/master/Problem%20Sets/Problem%20Set%205/histogram

Part 5. More Parallel Computing

14. Latency, Bandwidth, Throughput

Latency T (μ secs.) - time it takes for operation to start and compute.
Bandwidth - amount $v = d/t$
I tried introducing this notation, following Bader, Pöppl, and Khakhutskyy [18] :

$$\begin{aligned} T(1) \qquad & T : \mathbb{Z}^+ \to \mathbb{R}^+ \\ T(n_{\text{ops}}) \qquad & T : n_{\text{ops}} \mapsto T(n_{\text{ops}}) \end{aligned}$$

with $n_{\text{ops}} :=$ number of operations.

$T(1)$ vs. $T(n_{\text{ops}})$

$T(1)$ = time to complete 1 operation = latency.
$T(n_{\text{ops}})$ = time of complete n_{ops} operations.

$$n_{\text{ops}}/T(n_{\text{ops}}) = U = \text{throughput} \tag{31}$$

15. CUDA Execution model; SIMD, SIMT

cf. Bader, Pöppl, and Khakhutskyy [18]

SIMD (Single Instruction Multiple Data) Registers:

- same operations executed on pairs/triples of 2,4,... operands (ideally in 1 clock cycle)
- vector instructions generated by compiler ("vectorization") or (guided by) programmer ("intinsics", e.g.)

cf. Ch. 3. CUDA Execution Model of Cheng, Grossman, and McKercher (2014) [12]. When kernel grid is launched, thread blocks of that kernel grid are distributed among available SMs for execution. Once scheduled on SM, threads of a thread block execute concurrently only on that assigned SM.

CUDA employs *Single Instruction Multiple Thread (SIMT)* architecture to manage and execute threads in groups of 32 called *warps*.

1 instruction chain is imposed on multiple lightweight cores (via warp scheduler/dispatch unit)[18].

All threads in warp execute same instruction at the same time. Each thread has its own instruction address counter and register state, and carries out the current instruction on its own data. Each SM partitions the thread blocks assigned to it into 32-thread warps that it then schedules for execution on available hardware resources.

SIMT allows multiple threads in same warp to execute independently. Even though all threads in a warps start together at the same program address, it's possible for individual threads to have different behavior.

16. Performance Evaluation; Speed-Up, Efficiency, Amdahl's law

cf. Bader, Pöppl, and Khakhutskyy [18], Fundamentals

Definition 7. *Let*

$$T(p) := \text{ runtime on } p \text{ processors} \tag{32}$$

speed-up $S(p)$

$$S(p) := \frac{T(1)}{T(p)} \tag{33}$$

Typically, $1 \leq S(p) \leq p$.

16.1. **Absolute vs. Relative Speed-Up (definitions).** *Absolute speed-up* : best sequential algorithm for mono-processor (single processor) $T(1)$, compared to best parallel algorithm for multi-processor system.
relative speed-up : Compare same (parallel) algorithm for mono-(single-) and multi processor system.

16.2. **(Parallel) Efficiency.**

Definition 8. *Efficiency* $E(p)$

$$E(p) = \frac{S(p)}{p} \tag{34}$$

Typically $0 \leq E(p) \leq 1$.
$\exists$ absolute efficiency vs. relative efficiency.

16.3. **Amdahl's law.** Assumptions: program consists of sequential part s, $0 \leq s \leq 1$, which can't be parallelized (synchorniza-tion, data I/O, etc.).

- parallelizable part: $1 - s$
- execution time for parallel program on p processors

$$T(p) \equiv sT(1) + (1-s)T(p) = sT(1) + (1-s)\left(\frac{T(1)}{p}\right) = T(1) = \left[(s)\left(1 - \frac{1}{p}\right) + \frac{1}{p}\right] \tag{35}$$

To derive the relationship

$$T(p) = \frac{T(1)}{p} \tag{36}$$

use the $vt = l$ relation:

$$\sum_{i=1}^{p} v_1 T(p) = v_p T(p) = l = p v_1 T(p) \Longrightarrow T(1) = pT(p) \text{ or } T(p) = \frac{T(1)}{p}$$

Moving onwards, assume perfect-speed-up on arbitrary number of processors.
Resulting speed-up $S(p)$

$$S(p) = \frac{T(1)}{T(p)} = \frac{1}{s + \frac{1-s}{p}}$$
$$\lim_{p\to\infty} S(p) = \frac{1}{s}, \text{so } S(p) \leq \frac{1}{s} \tag{37}$$

Theorem 1 (Amdahl's law)**.** *Amdahl's law: speed-up is bounded by*

$$\boxed{S(p) \leq \frac{1}{s}} \tag{38}$$

16.4. **Gustafson's law.** Assumptions: Amdahl: sequential-part stays for increased problem size vs.
Gustavson: assume that any sufficiently large problem can be efficiently parallelized.
Fixed-time concept:

- parallel execution time normalized to $T(p) = 1$
- this contains a non-parallelizable part σ, $0 \leq \sigma \leq 1$

Execution time on single-processor:

$$T(1) = \sigma + p(1 - \sigma)$$

$$S(p) = \frac{T(1)}{T(p)} = \sigma + p(1-\sigma) = p - \sigma(p-1)$$

Parallel efficiency $E(p)$

$$E(p) = \frac{S(p)}{p} = \frac{\sigma + p(1-\sigma)}{p} = \frac{\sigma}{p} + 1 - \sigma \overset{p\to\infty}{\longmapsto} 1 - \sigma$$

More realistically, the larger problem sizes, if more processors are available, parallelizable parts typically increase.

17. Compute-Bound vs. Memory-Bound Performance; Compute-Bound, Memory-Bound

cf. Bader, Pöppl, and Khakhutskyy [18], Fundamentals
Consider a memory-bandwidth intensive algorithm:

- you can do a lot more flops than can be read from memory

Definition 9 (arithmetic intensity, operational intensity)**.** ***operational intensity*** *(or* ***arithmetic intensity****) of a code: number of performed flops per accessed byte*

17.1. **Memory-Bound Performance:**

- arithmetic intensity smaller than critical ratio
- you could execute additional flops "for free"
- speedup only possible by reducing memory accesses

17.2. **Compute-Bound Performance.**

- enough computational work to "hide" memory latency
- speedup only possibly by reducing operations

17.3. **Compute-Bound, Memory-Bound.** From stackoverflow, "What do the terms "CPU bound" and "I/O bound" mean?", compute bound seems to be **CPU Bound**, means rate at which process progresses is limited by speed of the CPU. Task that performs calculations on a small set of numbers, for example, multiplying small matrices, is likely to be CPU bound.

Memory bound means rate at which a process progresses is limited by amount memory available and speed of that memory access. A task that processes large amounts of in memory data, for example, multiplying large matrices, is likely to be Memory bound.

18. Factorization

Morgan Cooper for CSE710, Fall 2009

Let $n' := \sqrt{n} \in \mathbb{Z}^+$.

We want a_i, b_i, $i = 1, \dots n'$ s.t. $a_i b_i = n$.

e.g. $n = 20$.

$$n' = \sqrt{20} = 5 \in \mathbb{Z}^+$$

(so we want the *ceiling*).

Consider $x_i := i \quad \forall\, i = 1, \dots 2n'$.

Clearly,

if $\frac{n}{i} \in \mathbb{Z}$, i.e. $\gcd(n, i) = i$ for some $i \leq n'$,

then let $x_{i+n'} := q = \frac{n}{i}$ (let $x_{i+n'}$ be the other factor).

Generalize this to $n' \in \mathbb{Z}^+$ ($n' := \sqrt{n}$, usually).

Let $a_0 \in \mathbb{Z}^+$ (most likely $a_0 = 1$ or $a_0 = 2$ is what you'd "start off with", but no hard restrictions).

Then let

$$x_{i_x} := i_x + a_0 \qquad \forall\, i_x = 0, 1, \dots M_x - 1$$

Essentially, looping through entire blocks of threads so to "catch" every element in the array,

$$\forall\, i_x, i_x + M_x, i_x + 2M_x \cdots < n' \quad ,$$
$$i_x + a_0, i_x + a_0 + M_x, i_x + a_0 + 2M_x \dots$$

If $n \bmod x_{i_x} = n \bmod (i_x + a_0) = 0$, i.e. $\frac{0-n}{i_x + a_0} \in \mathbb{Z} \iff \frac{n}{i_x + a_0} \in \mathbb{Z}$, then

$$y(i_x) := x_{i_x}$$
$$y(i_x + n') := \frac{n}{x_{i_x}}$$

19. Matrix Multiplication, tiled, with shared memory

Consider matrix multiplication:

$$A \in \mathrm{Mat}_{\mathbb{K}}(N_i^A, N_j^A)$$
$$B \in \mathrm{Mat}_{\mathbb{K}}(N_i^B, N_j^B)$$
$$C \in \mathrm{Mat}_{\mathbb{K}}(N_i^C, N_j^C)$$

$AB = C$,

$$N_i^A = N_i^C \equiv N^A$$
$$N_j^A = N_i^B \equiv N^B$$
$$N_j^C = N_j^B \equiv N^C$$

$$C_{ij} = A_{ik} B_{kj} = \sum_{k=1}^{N} A_{ik} B_{kj} \qquad \forall\, i = 1, 2, \dots N^A$$
$$j = 1, 2, \dots N^C$$

Consider, given *block size* $M \in \mathbb{Z}^+$,

$$j_y = 0, 1, \dots \frac{N^A}{M} - 1 \equiv \texttt{blockIdx.y} =: j_I$$
$$j_x = 0, 1, \dots \frac{N^C}{M} - 1 \equiv \texttt{blockIdx.x} =: j_J$$

Given $A \in \mathrm{Mat}(N_i^A, N_j^A)$, given $(i, j) \in (\mathbb{Z}^+)^2$,

As C++ works also with pointers as C, "point" the first element of our $A_{\mathrm{sub}} \in \mathrm{Mat}_{\mathbb{K}}(M, M)$ submatrix that we want, to where we "start from" in the "source" matrix A, which is at $A_{iM, jM}$:

$$A_{\mathrm{sub}}(0, 0) := A(iM, jM)$$

and so this means that i.e. we had supposed $i = 0, 1, \dots \frac{N_i^A}{M} - 1$.

$$j = 0, 1 \dots \frac{N_j^A}{M} - 1$$

So consider

$$i = i_y + j_I M = 0, 1, \dots N^A - 1$$
$$j = i_x + j_J M = 0, 1, \dots N^C - 1$$

and so the crucial step is seen as

$$C_{ij} = \sum_{k=0}^{N^B - 1} A_{ik} B_{kj} = \sum_{j_K = 0}^{\frac{N^B}{M} - 1} \sum_{k=0}^{M-1} A_{i(k + j_K M)} B_{(k + j_K M), j}$$

I originally implemented matrix multiplication based on the CUDA C Programming Guide here at github : ernestyalumni/-CompPhys/moreCUDA/matmultShare.cu

Indeed, the idea from the CUDA C Programming Guide is essentially the above - I will reiterate again, since it was important in Pöppl's presentation for TUM in HPC - Algorithms and Applications (cf. Introduction to CUDA): given

$$A \in \mathrm{Mat}_{\mathbb{R}}(N_i^A, N_j^A)$$
$$B \in \mathrm{Mat}_{\mathbb{R}}(N_i^B, N_j^B)$$
$$C \in \mathrm{Mat}_{\mathbb{R}}(N_i^C, N_j^C)$$

Necessarily, for

$$AB = C$$

(39)
$$A_{ik} B_{kj} = C_{ij} \qquad \forall\, i = 1, 2 \dots N_i^A$$
$$\forall\, j = 1, 2, \dots N_j^B$$

we have to have

$$N_i^C = N_i^A$$
$$N_j^C = N_j^B$$
$$N_j^A = N_i^B$$

We sought thread warp coalescing, in conjunction with the so-called "row-major ordering" (also known as "row-major", "order") in the x-direction of thread block. Keep that in mind for the grid, block thread(s) assignment strategy.

19.0.1. *Grid, block thread(s) assignment strategy for Matrix multiplication.* Consider $N_j^B \times N_i^A$ calculations, i.e. $(j, i) \in \{0, 1 \dots N_j^B - 1\} \times \{0, 1 \dots N_i^A - 1\} \subset (\mathbb{Z}^+)^2$.

For (thread block) size $M \times M \equiv$ `dim3 dimBlock(M,M)`.

Then, the number of thread blocks along each dim. of the grid would be $\frac{N_j^B + M - 1}{M} = \lfloor \frac{N_j^B}{M} \rfloor$, $\frac{N_i^A + M - 1}{M} = \lfloor \frac{N_i^A}{M} \rfloor$

Nevertheless, the whole concept of tiling, tiling pattern with shared memory, is encapsulated in this equation:

(40)
$$C_{ij} = \sum_{k=0}^{N^B - 1} A_{ik} B_{kj} = \sum_{j_K = 0}^{\frac{N^B}{M} - 1} \sum_{k=0}^{M-1} A_{i(k + j_K M)} B_{(k + j_K M), j}$$

20. Dense Linear Algebra

cf. Dense Linear Algebra, HPC - Algorithms and Applications, Alexander Pöppl, TUM
Bader, Pöppl, and Khakhutskyy [18]

21. Cooperative Groups

21.1. **Example of reduction (summation).** cf. `thread_sum`
Let

```
cooperative_group::thread_block ctg;
int lane = ctg.thread_rank();
```

Let `lane`$\equiv l = 0, 1, \dots (|\texttt{ctg}| - 1)$, where $|\texttt{ctg}| =$ `ctg.size()`.
Consider,

$$\forall\, i = \frac{|ctg|}{2}, \frac{|ctg|}{2^2}, \dots \frac{|ctg|}{2^{K_{|ctg|}-1}}, 1$$

with $2^{K_{|\texttt{ctg}|}} =$ `ctg` or $\lfloor K_{|\texttt{ctg}|} \rfloor = \log_2 |\texttt{ctg}|$, i.e. $i = \frac{|ctg|}{2^j}$ and $i > 0$, $j = 1, 2, \dots K_{|\texttt{ctg}|}$.
In almost all use cases, since thread group is in multiples of 32 (or 16), then $\log_2 |ctg| \in \mathbb{Z}^+$.
Given

$$X : \{0, 1, \dots N-1\} \to \mathbb{K}$$
$$X : l \mapsto X(l)$$

and given some $s \in \mathbb{K}$, with $\mathbb{K} = \mathbb{Z}$ in this example case, then

$$X(l) := s$$

(which implies $|\texttt{ctg}| = N$; otherwise, it's necessary that $N \leq |\texttt{ctg}|$).
`ctg.sync()` $\iff \forall\, l = 0, 1, \dots (\texttt{ctg} - 1)$, $X(l) := s$

If $l < i$, $s := s + X(l + i)$. Explicitly,
Suppose $i = \frac{|ctg|}{2}$, $l = 0, 1, \dots \frac{|ctg|}{2} - 1$, $l + i = \frac{|ctg|}{2}, \frac{|ctg|}{2} + 1, \dots |ctg| - 1$
Suppose $i = \frac{|ctg|}{2^2}$, $l = 0, 1, \dots \frac{|ctg|}{2^2} - 1$, $l + i = \frac{|ctg|}{2^2}, \frac{|ctg|}{2^2} + 1, \dots \frac{|ctg|}{2} - 1$
Suppose $i = 1$, $l = 0$, $l + i = 1$.
But we should also consider first doing vector loading, and then, as a general strategy, load the result of vector loading into shared memory, *and then* do the reduction directly on shared memory.

22. Vector loads for higher memory access efficiency

cf. `reduce_sum`
Given

$$X : \{0, 1, \dots L-1\} \to \mathbb{K}$$
$$X : i \mapsto X[i] \in \mathbb{K}$$

with $\mathbb{K} = \mathbb{R}, \mathbb{Z}$ or even $\mathbb{C}$, and
Given

$$i_x \in \{0, 1, \dots M_x - 1\} \equiv \texttt{threadIdx.x}$$
$$j_x \in \{0, 1, \dots N_x - 1\} \equiv \texttt{blockIdx.x}$$

Then $\forall\, k_x := i_x + j_x M_x = 0, 1, \dots N_x M_x - 1$,
Consider $\frac{L}{4}$, $\exists\, 0 \leq r_4 < 4$ s.t. $L = 4q_4 + r_4$, so that $\frac{L}{4} = q_4 \in \mathbb{Z}^+$ (Division theorem).
Consider $i = k_x, k_x + N_x M_x, \dots$, with $i < L/4 = q_4$.
if $k_x = 0$, $i = 0, N_x M_x, 2N_x M_x, \dots$, with $i < L/4 = q_4$
if $k_x = 1$, $i = 1, 1 + N_x M_x, 1 + 2N_x M_x, \dots$ with $i < L/4 = q_4$.
if $k_x = N_x M_x - 1$, $i = N_x M_x - 1, 2N_x M_x - 1, 3N_x M_x - 1, \dots$ with $i < L/4 = q_4$.
Then consider $i = k_4 + 4q_4$, $i < L$. Then $i = 4q_4, 4q_4 + 1, \dots 4q_4 + 3$.
So for k_x, consider $q_x \in \{0, 1, \dots \frac{L}{4} - 1\}$, with $k_x = q_x$.
Let

$$v = v(q_x) := \sum_{\nu=0}^{4-1} X[4q_x + \nu]$$

22.1. **Vector loading and reduction on a single thread block in shared memory.** Then consider $i_x + j_x M_x = j_x M_x, j_x M_x + 1, \dots (j_x + 1)M_x - 1$, for fixed j_x.
Let

$$shX : \{0, 1, \dots M_x - 1\} \to \mathbb{K}$$
$$shX : i_x \mapsto shX[i_x]$$

represent the shared memory, upon which we'll directly reduce the sum.
lane $\equiv l := i_x$,

$$shX[l] := v(q_x)$$

`ctg.sync()` $\iff \forall\, l = 0, 1, \dots (\texttt{ctg}01)$, $shX[l] := v(q_x)$ since $q_x = q_x(i_x) = q_x(l)$.
And so, $\forall\, i = \frac{|ctg|}{2}, \frac{|ctg|}{2^2} \dots \frac{|ctg|}{2^{K_{|ctg|}-1}} = 2, 1$,
if $i = \frac{|ctg|}{2}$, $l < i$, $l = 0, 1, \dots \frac{|ctg|}{2} - 1$, $l + i = \frac{|ctg|}{2}, \frac{|ctg|}{2} + 1, \dots |\texttt{ctg}| - 1$,

$$v(q_x) := v(q_x) + shX[l + i]$$

Then $shX[l] := v(q_x)$ and so

$$shX[l] := shX[l] + shX[l + i]$$

So reduce block done entirely on shared memory, which is fast.

Part 6. Notes on Professional CUDA C Programming, Cheng, Grossman, McKercher

cf. Chen, Grossman, and McKercher (2014) [12]

23. CUDA Execution Model

cf. Ch. 3. CUDA Execution Model
When kernel grid is launched, thread blocks of that kernel grid are distributed among available SMs for execution. Once scheduled on SM, threads of a thread block execute concurrently only on that assigned SM.
CUDA employs *Single Instruction Multiple Thread (SIMT)* architecture to manage and execute threads in groups of 32 called *warps*.
All threads in warp execute same instruction at the same time. Each thread has its own instruction address counter and register state, and carries out the current instruction on its own data. Each SM partitions the thread blocks assigned to it into 32-thread warps that it then schedules for execution on available hardware resources.
SIMT allows multiple threads in same warp to execute independently. Even though all threads in a warps start together at the same program address, it's possible for individual threads to have different behavior.

23.0.1. *GPU Core Versus CPU Core.*

- CPU core, relatively heavy-weight, is designed for very complex control logic, optimizing latency T, namely $T(1)$, with $n_{\text{ops}} = 1$.
- GPU core, relatively light-weight, optimized for data-parallel tasks with simpler control logic, optimizing for throughput

cf. pp. 8 of Cheng, Grossman, and McKercher (2014) [12] *GPU capability*, described by

- number of CUDA cores
- Memory size

GPU performance, describe by

- *Peak computational performance* := how many single-precision or double-precision floating point calculations that can be processed per second. (gflops (billion floating-point operations per second) or tflops (trillion floating-point calculations per second)).
  - EY : 20170911 Peak computational performance = (compute) throughput (???)
- **Memory bandwidth** := ratio at which data can be read from or stored to memory; gigabytes per second, GB/s

cf. pp. 8 of Cheng, Grossman, and McKercher (2014) [12] CUDA kernels have no support for *static* variables. Then from What does "static" mean?, Stackoverflow,

(1) A static variable inside a function keeps its value between invocations.
(2) A static global variable or a function is "seen" only in the file it's declared in

cf. pp. 38 of Cheng, Grossman, and McKercher (2014) [12]

23.1. **Understanding the Nature of Warp Execution; Warps and Thread Blocks.** Once a thread block is scheduled to an SM, threads in thread block are further partitioned into warps. A warp consists of 32 consecutive threads and all threads in a warp are executed in SIMT, i.e. all threads execute same instruction, and each thread carries out operation on its own private data.

23.1.1. *Warp Divergence.* For example, consider the following statement:

```
if (cond) {
...
} else {
...
}
```

Suppose for 16 threads in a warp executing this code, `cond` is `true`, but for other 16 `cond` is `false`. Then half of the warp will need to execute the instructions in the `if` block, and the other half will need to execute the instructions in the `else` block. Threads in the same warp executing different instructions is referred to as **warp divergence.**

If threads of a warp diverge, the warp serially executes each branch path, disabling threads that do not take that path.

e.g. All threads within a warp must take both branches of the `if ... then` statement. If condition is `true` for a thread, it executes the `if` clause; otherwise, the thread stalls while waiting for that execution to complete.

24. Streams and Concurrency

cf. Ch. 6 Streams and Concurrency of Chen, Grossman, and McKercher (2014) [12]

24.0.1. *Introducing Streams and Events.*

- Functions in the CUDA API with *synchronous behavior* block the host thread until they complete.
- Functions in the CUDA API with *asynchronous behavior* return control to host immediately after being called.

24.0.2. *CUDA Streams.*

- *NULL stream*, default stream the kernel launches, implicitly declared, and data transfers use if you don't explicitly specify a stream.
- *non-null streams* explicitly created and managed; if you want to overlap different CUDA operations, you must use non-null streams.

Consider

```
cudaMemcpy(..., cudaMemcpyHostToDevice);|
kernel<<<grid, block>>>(...);
cudaMemcpy(..., cudaMemcpyDeviceToHost);|
```

From device perspective, all 3 operations are issued to default stream, and executed in order they were issued. Device has no awareness any other host operations performed.

From host perspective, each data transfer is synchronous and forces idle host time while waiting for them to complete. The kernel launch is *asynchronous*, so host application almost immediately resumes execution afterwards.

When performing an asynchronous data trasfer, you must use pinned (or non-pageable) host memory. Pinned memory can be allocated using either `cudaMallocHost` or `cudaHostAlloc`:

```
cudaError_t cudaMallocHost(void **ptr, size_t size);
cudaError_t cudaHostAlloc(void **pHost, size_t size, unsigned int flags);
```

Part 7. C++ and Computational Physics

cf. 2.1.1 Scientific hello world from Hjorth-Jensen (2015) [17]
in C,

```
int main (int argc, char* argv[])
```

`argc` stands for number of command-line arguments
`argv` is vector of strings containing the command-line arguments with
`argv[0]` containing name of program
`argv[1] , argv[2], ...` are command-line args, i.e. the number of lines of input to the program
"To obtain an executable file for a C++ program" (i.e. compile (???)),

```
gcc -c -Wall myprogram.c
gcc -o myprogram myprogram.o
```

`-Wall` means warning is issued in case of non-standard language
`-c` means compilation only
`-o` links produced object file `myprogram.o` and produces executable `myprogram`

```
# General makefile for c - choose PROG = name of given program

# Here we define compiler option, libraries and the target
CC= c++ -Wall
PROG= myprogram

# Here we make the executable file
${PROG} :       ${PROG}.o
                ${CC} ${PROG}.o -o ${PROG}

# whereas here we create the object file
```

```
#${PROG}.o :        ${PROG}.cpp
                    ${CC} -c ${PROG}.cpp
```

Here's what worked for me:

```
CC= g++ -Wall
PROG= program1

# Here we make the executable file
${PROG} :           ${PROG}.o
        ${CC} ${PROG}.o -o ${PROG}

# whereas here we create the object file

${PROG}.o :         ${PROG}.cpp
        ${CC} -c ${PROG}.cpp

# EY : 20160602 notice the different suffixes, and we see the pattern for the syntax

# (note: the <tab> in the command line is necessary formake towork)
# target: dependency1 dependency2 ...
#         <tab> command
```

cf. 2.3.2 Machine numbers of Hjorth-Jensen (2015) [17]
cf. 2.5.2 Pointers and arrays in C++ of Hjorth-Jensen (2015) [17]
Initialization (diagram):

$$\&\text{var} = \texttt{0x7ffc97efbd8c} \xmapsto{=} \text{pointer} = \&\text{var} = \texttt{0x7ffc97efbd8c}$$

$$\text{Memory} \xrightarrow{=} \text{pointer}$$

$$\text{(memory) addresses} \xrightarrow{=} \text{Obj(pointer)}$$

Referencing and deferencing operations on pointers to variables

$$\begin{array}{ccc} \text{var} & \xmapsto{=} & \&\text{var} = \texttt{0x7ffc97egbd8c} \\ \uparrow & & \\ \text{var}, 421 & \xmapsto{==} & \text{True} \\ \downarrow & & \\ 421 & & \end{array} \qquad \begin{array}{ccccc} \text{pointer} = \texttt{0x7ffc97egbd8c} & \xmapsto{*} & 421 & \xmapsto{\text{typedef}} & \text{int} \\ \downarrow \& & & & & \\ \&\text{pointer} = \texttt{0x7ffc97egbd8c} & & & & \end{array}$$

$$\begin{array}{ccc} \textbf{Types} & \xrightarrow{=} & \textbf{Memory} \\ \uparrow & & \\ \textbf{Types} \otimes \textbf{Dat} & \xrightarrow{==} & \text{Boolean} = \{\text{True}, \text{False}\} \\ \downarrow & & \\ \textbf{Dat} & & \end{array} \qquad \begin{array}{ccccc} \textbf{pointer} & \xrightarrow{*} & \textbf{Dat} & \xrightarrow{\text{typedef}} & \textbf{Types} \\ \downarrow \& & & & & \\ \textbf{Memory} & & & & \end{array}$$

24.1. **Numerical differentiation and interpolation (in C++).** cf. Chapter 3 "Numerical differentiation and interpolation" of Hjorth-Jensen (2015) [17].

This is how I understand it.

Consider the Taylor expansion for $f(x) \in C^\infty(\mathbb{R})$:

$$f(x) = f(x_0) + \sum_{j=1}^{\infty} \frac{f^{(j)}(x_0)}{j!} h^j$$

For $x = x_0 \pm h$,

$$f(x) = f(x_0 \pm h) = f(x_0) + \sum_{j=1}^{\infty} \frac{f^{(2j)}(x_0)}{(2j)!} h^{2j} \pm \sum_{j=1}^{\infty} \frac{f^{(2j-1)}(x_0)}{(2j-1)!} h^{2j-1}$$

Then

$$f(x_0 + 2^k h) - f(x_0 - 2^k h) = 2 \sum_{j=1}^{\infty} \frac{f^{(2j-1)}}{(2j-1)!}(x_0) 2^{k(2j-1)} h^{2j-1} =$$

$$= 2 \left[f^{(1)}(x_0) 2^k h + \sum_{j=2}^{\infty} \frac{f^{(2j-1)}(x_0)}{(2j-1)!} 2^{k(2j-1)} h^{2j-1} \right] =$$

$$= 2 \left[f^{(1)}(x_0) 2^k h + \frac{f^{(3)}(x_0)}{3!} 2^{k(3)} h^3 + \sum_{j=3}^{\infty} \frac{f^{(2j-1)}(x_0)}{(2j-1)!} 2^{k(2j-1)} h^{2j-1} \right]$$

So for $k = 1$,

$$f(x_0 + h) - f(x_0 - h) = 2 \left[f^{(1)}(x_0) h + \sum_{j=1}^{\infty} \frac{f^{(2j+1)}(x_0)}{(2j+1)!} h^{2j+1} \right]$$

Now

$$f(x_0+2^k h)+f(x_0-2^k h)-2f(x_0)=$$
$$=2\sum_{j=1}^{\infty}\frac{f^{(2j)}(x_0)}{(2j)!}2^{2jk}h^{2j}=$$
$$=2\left[\frac{f^{(2)}(x_0)}{2}2^{2k}h^2+\sum_{j=2}^{\infty}\frac{f^{(2j)}(x_0)}{(2j)!}2^{2jk}h^{2j}\right]=$$
$$=2\left[\frac{f^{(2)}(x_0)}{2}2^{2k}h^2+\frac{f^{(4)}(x_0)}{4!}2^{4k}h^4+\sum_{j=3}^{\infty}\frac{f^{(2j)}(x_0)}{(2j)!}2^{2jk}h^{2j}\right]$$

Thus for the case of $k=1$,

$$f(x_0+h)+f(x_0-h)-2f(x_0)=f^{(2)}(x_0)h^2+2\sum_{j=2}^{\infty}\frac{f^{(2j)}(x_0)}{(2j)!}h^{2j}$$

$$\frac{f(x_0+h)-f(x_0-h)}{2h}=f^{(1)}(x_0)+\sum_{j=1}^{\infty}\frac{f^{(2j+1)}(x_0)}{(2j+1)!}h^{2j}$$
$$\frac{f(x_0+h)+f(x_0-h)-2f(x_0)}{h^2}=f^{(2)}(x_0)+2\sum_{j=2}^{\infty}\frac{f^{(2(j+1))}(x_0)}{(2(j+1))!}h^{2j}$$

A pattern now emerges on how to include more calculations at points $x_0, x_0\pm 2^k h$ so to obtain better accuracy $O(h^l)$. For instance,

Given 5 pts. $\{x_0, x_0\pm h, x_0\pm 2h\}$,

$$f(x_0+2h)-f(x_0-2h)=2[f^{(1)}(x_0)2^1h+\frac{f^{(3)}(x_0)}{3!}2^3h^3+O(h^5)]$$
$$f(x_0+h)-f(x_0-h)=2[f^{(1)}(x_0)h+\frac{f^{(3)}(x_0)}{3!}h^3+O(h^5)]$$
$$\implies f'(x_0)=\frac{f(x_0-2h)-8f(x_0-h)+8f(x_0+h)-f(x_0+2h)}{12h}+O(h^4)$$

Hjorth-Jensen (2015) [17] argues, on pp. 46-47, that the additional evaluations are time consuming, to obtain further accuracy, so it's a balance.

To summarize, for $O(h^2)$ accuracy,

$$\frac{f(x_0+h)-f(x_0-h)}{2h}=f^{(1)}(x_0)+\sum_{j=1}^{\infty}\frac{f^{(2j+1)}(x_0)}{(2j+1)!}h^{2j}\qquad O(h^2)$$
$$\frac{f(x_0+h)+f(x_0-h)-2f(x_0)}{h^2}=f^{(2)}(x_0)+2\sum_{j=2}^{\infty}\frac{f^{(2j+2)}(x_0)}{(2j+2)!}h^{2j}\qquad O(h^2)$$

25. Interpolation

cf. 3.2 Numerical Interpolation and Extrapolation of Hjorth-Jensen (2015) [17]

$$\begin{aligned}y_0&=f(x_0)\\ y_1&=f(x_1)\\ &\vdots\\ y_N&=f(x_N)\end{aligned}$$

Given $N+1$ pts. , x_i's distinct (none of x_i values equal)

We want a polynomial of degree n s.t. $p(x)\in\mathbb{R}[x]$

$$p(x_i)=f(x_i)=y_i\qquad i=0,1\dots N$$

$$p(x)=a_0+a_1(x-x_0)+\cdots+a_i\prod_{j=0}^{i-1}(x-x_j)+\cdots+a_N(x-x_0)\dots(x-x_{N-1})=a_0+\sum_{i=1}^{N}a_i\prod_{j=0}^{i-1}(x-x_j)$$
$$\begin{aligned}a_0&=f(x_0)\\ a_0+a_1(x_1-x_0)&=f(x_1)\\ &\vdots\\ a_0+\sum_{i=1}^{k}a_i\prod_{j=0}^{i-1}(x_k-x_j)&=f(x_k)\end{aligned}$$

Hjorth-Jensen (2015) [17] mentions this Lagrange interpolation formula (I haven't found a good proof for it).

$$\boxed{p_N(x)=\sum_{i=0}^{N}\prod_{k\neq i}\frac{x-x_k}{x_i-x_k}y_i}\tag{41}$$

26. Classes (C++)

cf. C++ Operator Overloading in expression

Take a look at this link: C++ Operator Overloading in expression. This point isn't emphasized enough, as in Hjorth-Jensen (2015) [17]. This makes doing something like

$$d=a*c+d/b$$

work the way we expect. Kudos to user fredoverflow for his answer:

"The expression `(e_x*u_c)` is an rvalue, and references to non-const won't bind to rvalues.

Also, member functions should be marked `const` as well."

26.1. What are lvalues and rvalues in C and C++? C++ Rvalue References Explained

Original definition of *lvalues* and *rvalues* from *C*:

lvalue - expression e that may appear on the left or on the right hand side of an assignment

rvalue - expression that can only appear on right hand side of assignment =.

Examples:

```
int a = 42;
int b = 43;

// a and b are both l-values
a = b; // ok
b = a; // ok
a = a * b; // ok
```

```
// a * b is an rvalue:
int c = a * b; // ok, rvalue on right hand side of assignment
a * b = 42; // error, rvalue on left hand side of assignment
```

In *C++*, this is still useful as a first, intuitive approach, but
lvalue - expression that refers to a memory location and allows us to take the address of that memory location via the & operator.
rvalue - expression that's not a lvalue
So & reference *functor* can't act on rvalue's.

26.2. **Functors (C++); C++ Functors; C++ class templates.** For categories $\mathbf{A}, \mathbf{B}$, consider trying to understand, wrap your mind around C++, especiall C++11/14 style functors. The key *insight* is *composability*: use the mathematical property of **composition**.

(42)
$$\begin{array}{ccc} \mathbf{A} & \xrightarrow{\quad F \quad} & \mathbf{B} \\ \Big\downarrow{\scriptstyle \langle \text{Type} \rangle} & & \Big\downarrow{\scriptstyle \langle \text{Type} \rangle} \\ \text{Type} \circ \mathbf{A} & \xrightarrow{\quad \langle \text{Type} \rangle \circ F \quad} & \text{Type} \circ \mathbf{B} \end{array}$$

This webpage from K Hong helped with understanding C++11/14 style functors: C++ Tutorial - Functors(Function Objects) - 2017, cf. `http://www.bogotobogo.com/cplusplus/functors.php`
I implemented all of that in the webpage here: github functors.cpp , `ernestyalumni/CompPhys/Cpp/Cpp14/functors.cpp`
I will try to write a dictionary between math, i.e. mathematical formulation, and the class templates, structs.
I looked at pp. 213 of Conlon, pp. 513 of Rotman, and looked up keywords "functional."
Consider the bilinear functional that results in a function, i.e. $\mathcal{C}^\infty(\mathbb{R})$.

$$\mathbb{R} \times \mathbb{R} \to C^\infty(\mathbb{R})$$

i.e.

(43)
$$\begin{aligned} \mathbb{R} \times \mathbb{R} \quad &\to \text{Hom}_{\mathbb{R}}(\mathbb{R}, \mathbb{R}) \\ (a, b) \quad &\mapsto f(x) = ax + b \end{aligned}$$

with

$$\text{Hom}_{\mathbb{R}}(\mathbb{R}, \mathbb{R}) \ni \{\mathbb{R} \xrightarrow{f} \mathbb{R}\}$$

Compare this directly to `class Line` in `functor.cpp`. Note that this is *class object working as a functor*:

```
class Line {
        double a;                       // slope
        double b;                       // y-intercept

        public:
                Line(double slope = 1, double yintercept = 1) :
                        a(slope), b(yintercept) { }
                double operator()(double x){
                        return a*x + b;
                }};
```

Now consider the use of C++ function object, but with non-type template, C++ templates:

(44)
$$\begin{aligned} \mathbb{R} \quad &\to \text{Hom}_{\mathbb{R}}(\mathbb{R}, \mathbb{R}) \\ x \quad &\mapsto f(y) = y + x \end{aligned}$$

But suppose $y \in \mathbb{R}^d$, e.g. $y_i \in \mathbb{R}$, $i = 0, 1, \dots d-1$.

(45)
$$\begin{aligned} \mathbb{R} \quad &\to \text{Hom}_{\mathbb{R}}(\mathbb{R}^d, \mathbb{R}^d) \\ x \quad &\mapsto f(y) = y + x \text{ or } (f(y))_i = y_i + x \qquad \forall\, i = 0, 1, \dots d-1 \end{aligned}$$

So for the *class template*, to generalize $\mathbb{R}$ to some choice of field $\mathbb{K}$, generalize $\mathbb{R}^d$ to R-module R.

(46)
$$\begin{aligned} \mathbb{K} \quad &\to \text{Hom}_{\mathbb{K}}(\mathbb{K}, \mathbb{K}) \\ x \quad &\mapsto (f(y))_i = y_i + x \qquad \forall\, i = 0, 1, \dots d-1 \end{aligned}$$

And so the strategy is to generalize type by the class template (declaration), define the Hom from M to M by defining the Hom from $\mathbb{K}$ to $\mathbb{K}$ for each element of M.
Compare this directly to the code for class Add in `functor.cpp`:

```
template <typename T>
class Add
{
        T x;

        public:
                Add(T xx) : x(xx) { }
                void operator()(T& e) const { e += x; }
};
...

std::for_each(v2.begin(), v2.end(), Add<int>(10));

        std::for_each(v2.begin(), v2.end(), Add<int>(*v2.begin()) );
```

Notice how the construction of the Hom needs an input.

27. Numerical Integration

27.0.1. *Trapezoid rule (or trapezoidal rule).* See Integrate.ipynb.
From there, consider integration on $[a, b]$, considering $h := \frac{b-a}{N}$, and N+1 (grid) points, $\{a, a+h, a+2h, \dots, a+jh, \dots, a+Nh = b\}_{j=0\dots N}$.
Then $\frac{N}{2}$ pts. are our "x_0"; x_0's $= \{a+h, a+3h, \dots, a+(2j-1)h, \dots, a+\left(\frac{2N}{2}-1\right)h\}_{j=1\dots\frac{N}{2}}$.
Notice how we really need to care about if N is even or not. If N is not even, then we'd have to deal with the integration at the integration limits and choosing what to do.
Then

$$\int_a^b f(x)dx = \sum_{j=1}^{N/2} \int_{a+(2j-1)h-h}^{a+(2j-1)h+h} f(x)dx = \sum_{j=1}^{N/2} \frac{h}{2}(2f(a+(2j-1)h) + f(a+2(j-1)h) + f(a+2jh)) =$$

$$= h(f(a)/2 + f(a+h) + \cdots + f(b-h) + \frac{f(b)}{2}) = h\left(\frac{f(a)}{2} + \sum_{j=1}^{N-1} f(a+jh) + \frac{f(b)}{2}\right)$$

27.0.2. *Midpoint method or rectangle method.* .

Let $h := \frac{b-a}{N}$ be the step size. The grid is as follows:

$$\{a, a+h, \dots, a+jh, \dots, a+Nh = b\}_{j=0\dots N}$$

The desired midpoint values are at the following N points:

$$\{a + \frac{h}{2}, a + \frac{3}{2}h, \dots, a + \frac{(2j-1)h}{2}, \dots, a + \left(N - \frac{1}{2}\right)h\}_{j=1\dots N}$$

and so

$$\int_a^b f(x)dx \approx \sum_{j=1}^{N} f(x_j)h = \sum_{j=1}^{N} f\left(a + \frac{(2j-1)h}{2}\right)h \tag{47}$$

27.0.3. *Simpson rule.* The idea is to take the next "order" in the Lagrange interpolation formula, the second-order polynomial, and then we can rederive Simpson's rule. The algebra is worked out in Integrate.ipynb.

From there, then we can obtain Simpson's rule,

$$\int_a^b f(x)dx = \sum_{j=1}^{N/2} \int_{a+2(j-1)h}^{a+2jh} f(x)dx = \sum_{j=1}^{N/2} \frac{h}{3}(4f(a+(2j-1)h) + f(a+2(j-1)h) + f(a+2jh)) =$$
$$= \frac{h}{3}\left[f(a) + f(b) + \sum_{j=1}^{N/2} 4f(a+(2j-1)h) + 2\sum_{j=1}^{N/2-1} f(a+2jh)\right]$$

27.1. **Gaussian Quadrature.** cf. Hjorth-Jensen (2015) [17], Section 5.3 Gaussian Quadrature, Chapter 5 Numerical Integration

28. Runge-Kutta methods (RK)

cf. Hjorth-Jensen (2015) [17], Section 8.4 *More on finite difference methods, Runge-Kutta methods* and *wikipedia*, "Runge-Kutta methods," https://en.wikipedia.org/wiki/Runge%E2%80%93Kutta_methods

While Runge-Kutta methods are useful initially for ordinary differential equations, remember that under certain (very general, in fact, for smooth manifolds even) conditions, vector fields admit integral lines and there you simply solve an system of linear ODEs.

29. Partial Differential Equations

29.0.1. *Explicit Scheme.* cf. Hjorth-Jensen (2015) [17], Section 10.2.1 Explicit Scheme

Consider

$$u = u(t,x) \in C^\infty(M) = C^\infty(\mathbb{R} \times N)$$
$$\Delta u = \frac{\partial u}{\partial t}(t,x) \tag{48}$$

with initial conditions

$$u(0,x) = g(x) \qquad \forall\, 0 < x < L_x \text{ or } x \in \Omega \subset N \text{ (in general)}$$

e.g. $L_x = 1$

and boundary conditions

$$u(t,0) = a(t) \quad t \geq 0$$
$$u(t,L) = b(t) \quad t \geq 0$$

Consider the act of discretization as a transformation or a functor:

$$\frac{\partial u}{\partial t} = \frac{u(t+\Delta t, x) - u(t,x)}{\Delta t} + O(\Delta t) \xrightarrow{\text{discretize}} \frac{u(t_j + \Delta t, x_i) - u(t_j, x_i)}{\Delta t}$$
$$\frac{\partial^2 u}{\partial (x^i)^2} \approx \frac{u(t, x+\Delta x) - 2u(t,x) + u(t, x-\Delta x)}{(\Delta x)^2} \xrightarrow{\text{discretize}} \frac{u(t_j, x_i + \Delta x) - 2u(t_j, x_i) + u(t_j, x_i - \Delta x)}{(\Delta x)^2} \tag{49}$$

$$\implies u(t_j + \Delta t, x_i) = \frac{\Delta t}{(\Delta x)^2} u(t_j, x_i + \Delta x) + \left(1 - \frac{2\Delta t}{(\Delta x)^2}\right) u(t_j, x_i) + \frac{\Delta t}{(\Delta x)^2} u(t_j, x_i - \Delta x) \tag{50}$$

Discretize the initial conditions:

$$u(0,x) = g(x) \xrightarrow{\text{discretize}} u(0, x_i) = g(x_i) \tag{51}$$

and so, for the first step,

$$u(\Delta t, x_i) = \frac{\Delta t}{(\Delta x)^2} g(x_i + \Delta x) + \left(1 - \frac{2\Delta t}{(\Delta x)^2}\right) g(x_i) + \frac{\Delta t}{(\Delta x)^2} g(x_i - \Delta x) \tag{52}$$

It would appear to be instructive to show what discretize is doing, as a commutative diagram:

$$\begin{array}{ccc} \mathbb{R} \times N & \xrightarrow{\text{discretize}} & \mathbb{Z}^+ \times (\{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \times \{0 \dots L_z - 1\}) \subset \mathbb{Z}^+ \times \mathbb{Z}^d \\ \downarrow & & \downarrow \\ C^\infty(\mathbb{R} \times N) & \xrightarrow{\text{discretize}} & C^\infty(\mathbb{Z}^+ \times (\{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \times \{0 \dots L_z - 1\})) \subset C^\infty(\mathbb{Z}^+ \times \mathbb{Z}^d) \end{array} \tag{53}$$

Then there's this so-called "sparse" (I think this means that there are a lot more zeros as values for the entries in a matrix than there are nonzero values), tridiagonal (diagonal and next to the diagonal, diagonals) matrices representation of the time-evolution transformation/operator. I'll call this transformation over to this matrix representation, this functor, *matricer*.

$$C^\infty(\mathbb{Z}^+ \times (\{0 \dots L_x - 1\} \times \{0 \dots L_y - 1\} \times \{0 \dots L_z - 1\})) \xrightarrow{\text{matricer}} \mathbb{Z}^+ \times \mathbb{R}^{L_x L_y L_z} = \mathbb{Z}^+ \times \mathrm{Mat}_{\mathbb{R}}(L_x, L_y, L_z) \tag{54}$$

For the boundary conditions,

$$\begin{array}{ll} u(t,0) = a(t) & t \geq 0 \\ u(t,L) = b(t) & t \geq 0 \end{array} \xrightarrow{\text{discretize}} \begin{array}{l} u(t,0) = a(t) \\ u(t, x_{L_x - 1}) = b(t) \end{array}$$

For 1-dim. case, $V_j \in \mathrm{Mat}_{\mathbb{R}}(L_x)$ (vector or "column" matrix)

$$V_j = \begin{bmatrix} u(t_j, x_2) \\ u(t_j, x_3) \\ \vdots \\ u(t_j, x_{L_x - 3}) \end{bmatrix}$$

and so the "time-evolution" operator/transformation is

$$\widehat{A} \in \mathrm{Mat}_{\mathbb{R}}(L_x - 4, L_x - 4) = \mathrm{End}(\mathbb{R}^{L_x - 4}, \mathbb{R}^{L_x - 4})$$
$$\widehat{A} = \begin{bmatrix} \frac{\Delta t}{(\Delta x)^2} & 1 - 2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{\Delta t}{(\Delta x)^2} & 1 - 2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} & 0 & \dots & 0 \\ & & & \ddots & & & \\ 0 & 0 & \dots & 0 & \frac{\Delta t}{(\Delta x)^2} & 1 - 2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} \end{bmatrix} \tag{55}$$

Note that in the specialized 1-dim. case where $a(t) = b(t) = 0$ (boundary conditions for both ends is of value 0), then we can, in this specialized case, define the matrix $\widehat{A}$ to be

$$\widehat{A} \in \mathrm{Mat}_{\mathbb{R}}(L_x - 2, L_x - 2) = \mathrm{End}(\mathbb{R}^{L_x-2}, \mathbb{R}^{L_x-2})$$

$$\widehat{A} = \begin{bmatrix} 1-2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} & 0 & 0 & \dots & 0 & 0 \\ \frac{\Delta t}{(\Delta x)^2} & 1-2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{\Delta t}{(\Delta x)^2} & 1-2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} & 0 & \dots & 0 \\ & & & \ddots & & & \\ 0 & 0 & \dots & 0 & \frac{\Delta t}{(\Delta x)^2} & 1-2\frac{\Delta t}{(\Delta x)^2} & \frac{\Delta t}{(\Delta x)^2} \\ 0 & 0 & 0 & \dots & 0 & \frac{\Delta t}{(\Delta x)^2} & 1-2\frac{\Delta t}{(\Delta x)^2} \end{bmatrix} \tag{56}$$

e.g.

$$g(x) = \sin\left(\frac{\pi}{l_x}x\right)$$

with an analytic solution of

$$u(t,x) = \sin\left(\frac{\pi}{l_x}x\right)\exp\left(-\left(\frac{\pi}{l_x}\right)^2 t\right)$$

It was bizarre to me that in Hjorth-Jensen (2015) [17], Section 10.2.1 Explicit Scheme, on pp. 307, Hjorth-Jensen went through a lengthy and thorough explanation of this "matricer" operation, i.e. doing the time-evolution with a matrix on a vector of values at grid points, and yet in the pseudo-code, essentially, there is no trace of that matrix! It's essentially a local "stencil" operation. What the heck?

I present the "matrix form" code in my github repository: 'diffusion1dexplicit.cpp'. To be explicit, the code follows the previous writeup, with its notation, essentially one-to-one.

29.0.2. *Implicit scheme.* cf. Hjorth-Jensen (2015) [17], Section 10.2.2 Implicit Scheme

Consider

$$\begin{aligned} &\text{backwards formula :} && \frac{\partial u}{\partial t}(t,x) \approx \frac{u(t_j,x_i) - u(t_j - \Delta t, x_i)}{\Delta t} \text{ or even} \\ &\text{midpoint approximations :} && \frac{\partial u}{\partial t}(t,x) \approx \frac{u(t_j + \Delta t,x_i) - u(t_j - \Delta t, x_i)}{2\Delta t} \end{aligned} \tag{57}$$

Consider the same spatial discretization as before for the Laplacian:

$$\frac{\partial^2 u}{\partial (x^i)^2} \approx \frac{u(t,x+\Delta x) - 2u(t,x) + u(t,x-\Delta x)}{(\Delta x)^2} \xrightarrow{\text{discretize}} \frac{u(t_j,x_i+\Delta x) - 2u(t_j,x_i) + u(t_j,x_i-\Delta x)}{(\Delta x)^2}$$

and so for the backwards formula case,

$$\begin{aligned} \frac{u(t_j,x_i) - u(t_j-\Delta t,x_i)}{\Delta t} &= \frac{u(t_j,x_i+\Delta x) - 2u(t_j,x_i) + u(t_j,x_i-\Delta x)}{(\Delta x)^2} \\ u(t_j - \Delta t, x_i) &= -\frac{\Delta t}{(\Delta x)^2}u(t_j,x_i+\Delta x) + \left(1 + \frac{2\Delta t}{(\Delta x)^2}\right)u(t_j,x_i) - \frac{\Delta t}{(\Delta x)^2}u(t_j,x_i-\Delta x) \end{aligned} \tag{58}$$

resulting in the backwards time-evolution matrix $\widehat{A}$ (keeping in mind the special boundary condition of 0 value for u at both ends, for the sake of a simplified discussion):

$$\widehat{A} \in \mathrm{Mat}_{\mathbb{R}}(L_x - 2, L_x - 2) = \mathrm{End}(\mathbb{R}^{L_x-2}, \mathbb{R}^{L_x-2})$$

$$\widehat{A} = \begin{bmatrix} 1+2\frac{\Delta t}{(\Delta x)^2} & -\frac{\Delta t}{(\Delta x)^2} & 0 & 0 & \dots & 0 \\ -\frac{\Delta t}{(\Delta x)^2} & 1+2\frac{\Delta t}{(\Delta x)^2} & -\frac{\Delta t}{(\Delta x)^2} & 0 & \dots & 0 \\ & & & \ddots & & \\ 0 & 0 & \dots & -\frac{\Delta t}{(\Delta x)^2} & 1+2\frac{\Delta t}{(\Delta x)^2} & -\frac{\Delta t}{(\Delta x)^2} \\ 0 & 0 & 0 & \dots & -\frac{\Delta t}{(\Delta x)^2} & 1+2\frac{\Delta t}{(\Delta x)^2} \end{bmatrix} \tag{59}$$

$$\widehat{A}u^j = u^{j-1}$$

and so

$$\widehat{A}^{-1}u^{j-1} = u^j$$

29.1. **Crank-Nicolson method.** Hjorth-Jensen (2015) [17], Section 10.2.3 Crank Nicholson scheme has a write up about the Crank-Nicolson method, but the derivation is unclear (and sloppy, in that after the Taylor expansions, he says that the terms magically add up to the desired result, and the "approximation" notation is vacuous in that nothing new was conveyed). Rather, look at Crank Nicolson Scheme for the Heat Equation for a clearer derivation, that drives home the point of looking at the time between time steps.

Consider the following Taylor expansions about $t^{1/2} := t + \frac{\Delta t}{2}$.

$$u(t+\Delta t,x) = u(t + \frac{\Delta t}{2} + \frac{\Delta t}{2}, x) \equiv u(t^{1/2} + \frac{\Delta t}{2}, x) = u(t^{1/2},x) + \frac{\Delta t}{2}\frac{\partial u}{\partial t}(t^{1/2},x) + \frac{1}{2}\left(\frac{\Delta t}{2}\right)^2\frac{\partial^2 u}{\partial t^2}(t^{1/2},x) + O((\Delta t)^3)$$

$$u(t,x) = u(t + \frac{\Delta t}{2} - \frac{\Delta t}{2}, x) \equiv u(t^{1/2} - \frac{\Delta t}{2}, x) = u(t^{1/2},x) - \frac{\Delta t}{2}\frac{\partial u}{\partial t}(t^{1/2},x) + \frac{1}{2}\left(\frac{\Delta t}{2}\right)^2\frac{\partial^2 u}{\partial t^2}(t^{1/2},x) + O((\Delta t)^3)$$

$$u(t+\Delta t,x) - u(t,x) = \Delta t\frac{\partial u}{\partial t}(t^{1/2},x) + O((\Delta t)^3)$$

$$\implies \frac{\partial u}{\partial t}(t^{1/2},x) = \frac{u(t+\Delta t,x) - u(t,x)}{\Delta t}$$

with $O((\Delta t)^2)$ order of accuracy.

To approximate

$$\frac{\partial^2 u}{\partial x^2}(t + \frac{\Delta t}{2}, x) \equiv \frac{\partial^2 u}{\partial x^2}(t^{1/2},x)$$

use average of second, centered differences for $\frac{\partial^2 u}{\partial x^2}(t+\Delta t,x)$ and $\frac{\partial^2 u}{\partial x^2}(t,x)$.

$$\frac{\partial^2 u}{\partial x^2}(t+\frac{\Delta t}{2},x) \approx \frac{1}{2}\left[\frac{u(t+\Delta t,x+\Delta x) - 2u(t+\Delta t,x) + u(t+\Delta t,x-\Delta x)}{(\Delta x)^2} + \frac{u(t,x+\Delta x) - 2u(t,x) + u(t,x-\Delta x)}{(\Delta x)^2}\right]$$

Then for

$$\frac{\partial u}{\partial t}(t^{1/2},x) = C_0\Delta u(t^{1/2},x) \xrightarrow{discretize}$$

$$u(t+\Delta t,x) - u(t,x) =$$

$$= \frac{1}{2}C_0\frac{\Delta t}{(\Delta x)^2}(u(t+\Delta t,x+\Delta x) - 2u(t+\delta t,x) + u(t+\Delta t,x-\Delta x)) + \frac{1}{2}C_0\frac{\Delta t}{(\Delta x)^2}u(t,x+\Delta x) + \frac{-\Delta t}{(\Delta x)^2}C_0u(t,x) + \frac{1}{2}\frac{\Delta t}{(\Delta x)^2}C_0u(t,x-\Delta$$

Let

$$\alpha := \frac{\Delta t}{(\Delta x)^2}$$

Then

$$\frac{-1}{2}\alpha u(t+\Delta t, x+\Delta x) + (1+\alpha)u(t+\Delta t, x) - \frac{\alpha}{2}u(t+\Delta t, x-\Delta x) = \frac{\alpha}{2}u(t, x+\Delta x) + (1-\alpha)u(t,x) + \frac{\alpha}{2}u(t, x-\Delta x) \tag{60}$$

In general

$$\begin{aligned}\frac{-\alpha}{2}C_0 u(t+\Delta t, x+\Delta x) + (1+C_0\alpha)u(t+\Delta t, x) - \frac{C_0\alpha}{2}u(t+\Delta t, x-\Delta x) = \\ = \frac{C_0\alpha}{2}u(t, x+\Delta x) + (1-C_0\alpha)u(t,x) + \frac{C_0\alpha}{2}u(t, x-\Delta x)\end{aligned} \tag{61}$$

This scheme necessitates a matrix representation. In matrix form,

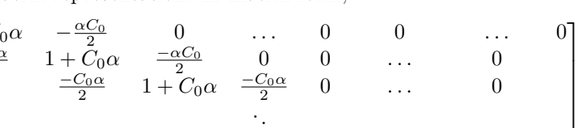
$$\begin{aligned}&\begin{bmatrix} 1+C_0\alpha & -\frac{\alpha C_0}{2} & 0 & \dots & 0 & 0 & \dots & 0 \\ \frac{-C_0\alpha}{2} & 1+C_0\alpha & \frac{-\alpha C_0}{2} & 0 & 0 & \dots & 0 & \\ 0 & \frac{-C_0\alpha}{2} & 1+C_0\alpha & \frac{-C_0\alpha}{2} & 0 & \dots & 0 & \\ & & & \ddots & & & & \\ 0 & 0 & \dots & 0 & \frac{-\alpha}{2} & 1+C_0\alpha & \frac{-\alpha}{2} & \\ 0 & 0 & \dots & 0 & 0 & \frac{-\alpha}{2} & 1+C_0\alpha & \end{bmatrix} u_i^{t+\Delta t} = \\ &= \begin{bmatrix} 1-C_0\alpha & \frac{C_0\alpha}{2} & 0 & 0 & 0 & \dots & 0 \\ \frac{C_0\alpha}{2} & 1-C_0\alpha & \frac{C_0\alpha}{2} & 0 & 0 & \dots 0 & \\ 0 & \frac{C_0\alpha}{2} & 1-C_0\alpha & \frac{C_0\alpha}{2} & 0 & \dots & 0 \\ & & & \ddots & & & \\ 0 & 0 & 0 & \dots & \frac{C_0\alpha}{2} & 1-C_0\alpha & \frac{C_0\alpha}{2} \\ 0 & 0 & 0 & \dots & 0 & \frac{C_0\alpha}{2} & 1-C_0\alpha \end{bmatrix} u_i^t \\ &\implies \widehat{B}u_i^{t+\Delta t} = \widehat{A}u_i^t \text{ or} \\ &u_i^{t+\Delta t} = \widehat{B}^{-1}\widehat{A}u_i^t\end{aligned} \tag{62}$$

29.2. **Jacobi method, SOR method, for the Laplace and Poisson equation.** 3.1 Poisson's Equation and Relaxation Methods of 410-505 Physics had a good, online, clear explanation of Jacobi method and improvements, namely the Successive Over Relaxation (SOR) method, applied to Laplace and Poisson equation, with clearly labelled diagrams: http://www.physics.buffalo.edu/phy410-505/2011/topic3/app1/index.html

30. Call by reference - Call by Value, Call by reference (in C and in C++)

cf. pp. 58, 2.10 Pointers Ch. 2 Scientific Programming in C, Fitzpatrick [16] `printfact3.c`, printfact3.c
pass pointer, pass by reference, call by pointer, call by reference
In C:

- *function prototype* -

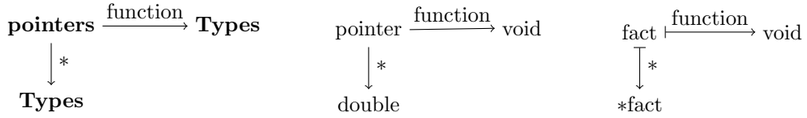


$\implies$

```
void factorial(double *)
```

where for factorial, it's just your choice of name for *function*.

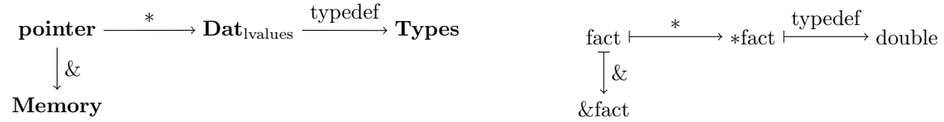
- *function definition* -



$\implies$

```
void function(double *fact) { ... }
```

Inside the function definition,

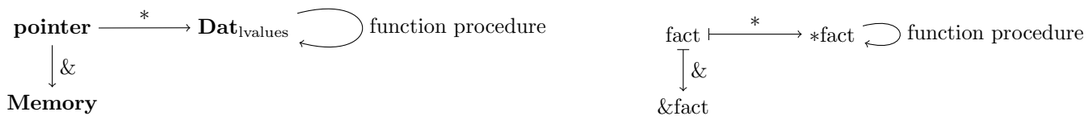


and so, for instance, in the function definition, you can do things like this:

```
*fact = 1
*fact *= (double) n
```

and so notice that from `*fact = 1`, `*fact` is a lvalue.

– *function procedure*



$\implies$

```
*fact *= (double) n
```

- "Using" the function, function "instantiation", "calling" the function, i.e. "running" the function

$$\begin{array}{ccc} \text{function procedure} \circlearrowright \textbf{Types} & \xrightarrow{\&} & \textbf{Memory} \\ & & \downarrow \cong \\ & & \textbf{pointers} \xrightarrow{\text{function}} \textbf{Types} \end{array}$$

$$\begin{array}{ccc} \text{function procedure} \circlearrowright \text{double} & \xrightarrow{\&} & \text{Memory}(\text{Obj}\mathit{Memory}) \\ & & \downarrow \cong \\ & & \text{pointer} \xrightarrow{\text{function}} \text{void} \end{array}$$

$$\begin{array}{ccc} \text{function procedure} \circlearrowright \text{fact} & \xrightarrow{\&} & \&\text{fact} \\ & & \downarrow \cong \\ & & \&\text{fact} \xrightarrow{\text{function}} \text{function(\& fact)} \end{array}$$

where, again simply note the notation, that we're using *function* and *factorial*, *fact* for *nameofpointer*, interchangeably: see printfact3.c for the example I'm referring to.

Again, *in C*, consider *a pointer to a function* passed to another function as an argument. Take a look at passfunction.c simultaneously.

- *function prototype* -

$$\begin{array}{ccc} \textbf{pointer} & \xrightarrow{\text{hostfunction}} & \textbf{Types} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}} & & \end{array} \qquad \begin{array}{ccc} \text{pointer} & \xrightarrow{\text{hostfunction}} & \text{void} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}}(\text{double}, \text{double}) & & \end{array}$$

$\Longrightarrow$

```
void hostfunction(double (*)(double))
```

We could further generalize this syntax, simply for syntax and notation sake, as such:

$$\begin{array}{ccc} \textbf{pointer} & \xrightarrow{\text{hostfunction}} & \textbf{Types} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}} & & \end{array} \qquad \begin{array}{ccc} \text{pointer} & \xrightarrow{\text{hostfunction}} & \text{data-type} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}}(\text{typei}, \text{typef}) & & \end{array}$$

$\Longrightarrow$

```
data-type hostfunction(typef (*)(typei))
```

For practice, consider more than 1 argument in our function, and the other argument, for practice, is a pointer, we're "passing by reference."

$$\begin{array}{ccccc} \textbf{pointers} \times \textbf{pointers} & & \xrightarrow{\text{hostfunction}} & & \textbf{Types} \\ & \overset{\text{pr}_1}{\swarrow} \quad \overset{\text{pr}_2}{\searrow} & & & \\ \textbf{pointers} & & & \textbf{pointers} & \\ \downarrow * & & & \downarrow * & \\ \text{Mor}_{\textbf{Types}} & & & \textbf{Types} & \end{array}$$

$$\begin{array}{ccccc} \mathit{pointer} \times \mathit{pointer} & & \xrightarrow{\text{hostfunction}} & & \text{void} \\ & \overset{\text{pr}_1}{\swarrow} \quad \overset{\text{pr}_2}{\searrow} & & & \\ \mathit{pointer} & & & \mathit{pointers} & \\ \downarrow * & & & \downarrow * & \\ \text{Mor}_{\textbf{Types}}(\text{double}, \text{double}) & & & \text{double} & \end{array}$$

$\Longrightarrow$

```
void hostfunction( double (*)(double), double *)
```

- *function definition*

$$\begin{array}{ccc} \textbf{pointers} & \xrightarrow{\text{hostfunction}} & \textbf{Types} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}} & & \end{array} \qquad \begin{array}{ccc} \text{pointer} & \xrightarrow{\text{hostfunction}} & \text{void} \\ \downarrow * & & \\ \text{Mor}_{\textbf{Types}}(\text{double}, \text{double}) & & \end{array} \qquad \begin{array}{ccc} \text{fun} & \xrightarrow{\text{hostfunction}} & \text{void} \\ \downarrow * & & \\ *\text{fun} & & \end{array}$$

$\Longrightarrow$

```
void hostfunction(double (*fun)(double)) { ... }
```

- *Inside* the function definition,

$$\textbf{Types} \xrightarrow{*\text{fun}} \textbf{Types} \xrightarrow{=} \textbf{Types}$$

$$\text{double} \xrightarrow{*\text{fun}} \text{double} \xrightarrow{=} \text{double}$$

$$x \xmapsto{*\text{fun}} (*\text{fun})(x) \xmapsto{=} y = (*\text{fun})(x)$$

$\Longrightarrow$

```
y = (*fun)(x)
```

- "Using" the function - the *actual* syntax for "passing" a function into a function is interesting (peculiar?): you only need the *name* of the function.

  Let's quickly recall how a function is prototyped, "declared" (or, i.e., defined), and used:

– *function prototype* -

$$\mathbf{Types} \xrightarrow{\text{fun1}} \mathbf{Types}$$

$$\text{double} \xrightarrow{\text{fun1}} \text{double}$$

$\Longrightarrow$

```
double fun1(double)
```

– *function definition* -

$$\mathbf{Types} \xrightarrow{\text{fun1}} \mathbf{Types}$$

$$\text{double} \xrightarrow{\text{fun1}} \text{double}$$

$$z \xmapsto{\text{fun1}} 3.0z * z - z (= 3z^2 - z)$$

$\Longrightarrow$

```
double fun1(double z) { ... }
```

– Using function - `fun1(z)`

and so

$$\text{fun1} \in \text{Mor}_{\mathbf{Types}}(\text{double}, \text{double})$$

And so again, it's interesting in terms of syntax that all you need is the *name* of the function to pass into the arguments of the "host function" when using the host function:

$$\text{Mor}_{\mathbf{Types}} \xrightarrow{\text{hostfunction}} \mathbf{Types}$$

$$\text{Mor}_{\mathbf{Types}}(\text{double}, \text{double}) \xmapsto{\text{hostfunction}} \text{void}$$

$$\text{fun1} \xmapsto{\text{hostfunction}} \text{hostfunction(fun1)}$$

$\Longrightarrow$

```
hostfunction(fun1)
```

30.0.1. *C++ extensions, or how C++ pass by reference (pass a pointer to argument) vs. C.* Recall how C passes by reference, and look at Fitzpatrick [16], pp. 83-84 for the `square` function:

- *function prototype* -

$$\begin{array}{ccc} \mathbf{pointer} & \xrightarrow{\text{square}} & \mathbf{Types} \\ \downarrow * & & \\ \mathbf{Types} & & \end{array} \qquad \begin{array}{ccc} \text{pointer} & \xmapsto{\text{square}} & \text{void} \\ \downarrow * & & \\ \text{double} & & \end{array}$$

$\Longrightarrow$

```
void square(double *)
```

- *function definition* -

$$\begin{array}{ccc} \mathbf{pointers} & \xrightarrow{\text{square}} & \mathbf{Types} \\ \downarrow * & & \\ \mathbf{Types} & & \end{array} \qquad \begin{array}{ccc} \text{pointer} & \xrightarrow{\text{function}} & \text{void} \\ \downarrow * & & \\ \text{double} & & \end{array} \qquad \begin{array}{ccc} y & \xmapsto{\text{square}} & \text{void} \\ \downarrow * & & \\ *y & & \end{array}$$

$\Longrightarrow$

```
void square(double *y) { ... }
```

Inside the function definition,

$$\mathbf{pointer} \xrightarrow{*} \mathbf{Dat}_{\text{lvalues}} \xrightarrow{\text{typedef}} \mathbf{Types} \qquad y \xmapsto{*} *y \xmapsto{\text{typedef}} \text{double}$$

and so, for instance, in the function definition, you can do things like this:

```
*y = x*x
```

- "Using" the function, function "instantiation", "calling" the function, i.e. "running" the function

$$\begin{array}{ccccc} \text{function procedure} & \circlearrowright & \mathbf{Types} & \xrightarrow{\&} & \mathbf{Memory} \\ & & & & \downarrow \cong \\ & & & & \mathbf{pointers} \xrightarrow{\text{square}} \mathbf{Types} \end{array}$$

$$\begin{array}{ccccc} \text{function procedure} & \circlearrowright & \text{double} & \xrightarrow{\&} & \text{Memory}(\text{Obj}\mathit{Memory}) \\ & & & & \downarrow \cong \\ & & & & \text{pointer} \xrightarrow{\text{square}} \text{void} \end{array}$$

$$\begin{array}{ccccc} \text{function procedure} & \circlearrowright & \text{res} & \xmapsto{\&} & \&\text{res} \\ & & & & \downarrow \cong \\ & & & & \&\text{res} \xmapsto{\text{square}} \text{square}(\&\text{res}) \end{array}$$

30.0.2. *C++ syntax for dealing with passing pointers (and arrays) into functions.* However, in *C++*, a lot of the dereferencing $*$ and referencing & is not explicitly said so in the syntax. In this syntax, passing by reference is indicated by prepending the & ampersand to the variable name, in function declaration (prototype and definition). We don't have to explicitly deference the argument in the function (it's done behind the scene) and syntax-wise (it seems), we only have to refer to the argument by regular local name.

Indeed, the syntax appears "shortcutted" greatly:

- *function prototype* -

$$\mathbf{pointer} \times \mathbf{Types} \xrightarrow{\text{function}} \mathbf{Types} \qquad \text{pointer}, \text{double} \xmapsto{\text{function}} \text{void}$$

$\Longrightarrow$

```
void function(double &)
```

- *function definition* -

$$\textbf{pointers} \times \textbf{Types} \xrightarrow{\text{square}} \textbf{Types} \qquad \text{pointer}, \text{double} \xrightarrow{\text{function}} \text{void} \qquad \&, y \xmapsto{\text{function}} \text{function}(\text{double } \&y)$$

$\Longrightarrow$

```
void function(double &y) { ... }
```

Inside the function definition,

$$\begin{array}{c} \text{End}(\text{double}, \text{double}) \\ \text{double} \longrightarrow \text{double} \end{array} \qquad \begin{array}{c} \text{End}(\text{double}, \text{double}) \\ y \longmapsto y = x * x \end{array}$$

and so, for instance, in the function definition, you can do things like this:

```
y = x*x
```

with no deferencing needed.

- "Using" the function, function "instantiation", "calling" the function, i.e. "running" the function

$$\textbf{Types} \xrightarrow{\text{function}} \textbf{Types} \qquad \text{double} \xrightarrow{\text{function}} \text{void} \qquad \text{res} \xmapsto{\text{function}} \text{function}(\text{res})$$

30.0.3. *C++ note on arrays.* For dealing with arrays, Stroustrup (2013) [6], on pp. 12 of Chapter 1 The Basics, Section 1.8 Pointers, Arrays, and References, does the following:

- *array declaration* -

```
type a[n]; // type[n]; array of n type's
```

- "Using" arrays in function prototypes, i.e. passing into arguments of functions for *function prototypes*

```
data-type function( type * arrayname)
```

- "Using" arrays when "using" functions, i.e. passing into arguments when a function is "called" or "executed"

```
function( arrayname )
```

Fitzpatrick [16] mentions using `inline` for short functions, no more than 3 lines long, because of memory cost of calling a function.

30.0.4. *Need a CUDA, C, C++, IDE? Try Eclipse!* This website has a clear, lucid, and pedagogical tutorial for using Eclipse: Creating Your First C++ Program in Eclipse. But it looks like I had to pay. Other than the well-written tips on the webpage, I looked up stackexchange for my Eclipse questions (I had difficulty with the Eclipse documentation).

Others, like myself, had questions on how to use an IDE like Eclipse when learning CUDA, and "building" (is that the same as compiling?) and running only single files.

My workflow: I have a separate, in my file directory, folder with my github repository clone that's local.

I start a New Project, CUDA Project, in Eclipse. I type up my single file (I right click on the `src` folder and add a 'Source File'). I build it (with the Hammer, Hammer looking icon; yes there are a lot of new icons near the top) and it runs. I can then run it again with the Play, triangle, icon.

I found that if I have more than 1 (2 or more) file in the `src` folder, that requires the `main` function, it won't build right.

So once a file builds and it's good, I, in Terminal, `cp` the file into my local github repository. Note that from there, I could use the `nvcc` compiler to build, from there, if I wanted to.

Now with my file saved (for example, `helloworldkernel.cu`), then I can delete it, without fear, from my, say, `cuda-workplace`, from the right side, "C/C++ Projects" window in Eclipse.

31. On CUDA By Example

Take a look at 3.2.2 A Kernel Call, a Hello World in CUDA C, with a simple kernel, on pp. 23 of Sanders and Kandrot (2010) [19] and on github, [helloworldkernel.cu](https://github.com/ernestyalumni/CompPhys/blob/master/CUDA-By-Example/helloworldkernel.cu). Let's work out the functor interpretation for practice.

- *function definition* -

$$\textbf{Types} \xrightarrow{\text{kernel}} \textbf{Types}$$
$$\text{void} \xrightarrow{\text{kernel}} \text{void}$$

where `kernel` $\in$ `__global__`

$\Longrightarrow$

```
__global__ void kernel(void) { }
```

CUDA C adds the `__global__` qualifier to standard C to *alert the compiler that the function*, `kernelfunction`, should be compiled to run on the *device*, not the host (pp. 24 [19]).

- "Using", "calling", "running" function -

$<<<>>>: (n_{\text{block}}, n_{\text{threads}}) \times \text{kernelfunction} \mapsto \text{kernelfunction} <<< n_{\text{block}}, n_{\text{threads}} >>> \in \text{End}(\text{Dat}_{\textbf{Types}})$

$<<<>>>: \mathbb{N}^+ \times \mathbb{N}^+ \times \text{Mor}_{\text{GPU}} \to \text{End}(\text{Dat}_{GPU})$

$\Longrightarrow$

```
kernel<<<1,1>>>();
```

cf. 3.2.3 Passing Parameters of Sanders and Kandrot (2010) [19]

Taking a look at [add-passb.cu](https://github.com/ernestyalumni/CompPhys/blob/master/CUDA-By-Example/add-passb.cu), let's work out the functor interpretation of `cudaMalloc`, `cudaMemcpy`.

In `main`, "declaring" a pointer:

```
int *dev_c
```

$\Longleftarrow$

$$\textbf{pointers} \xrightarrow{*} \textbf{Dat}_{\text{lvalues}} \xrightarrow{\text{typedef}} \textbf{Types}$$

$$\texttt{dev\_c} \xmapsto{*} \texttt{*dev\_c} \xmapsto{\text{typedef}} \text{int}$$

We can also do, note, the `sizeof` function (which is a well-defined mapping, for once) on Obj**Types**:

$$\textbf{pointers} \xrightarrow{*} \textbf{Dat}_{\text{lvalues}} \xrightarrow{\text{typedef}} \textbf{Types} \xrightarrow{\text{sizeof}} \mathbb{N}^+$$

$$\texttt{dev\_c} \xmapsto{*} \texttt{*dev\_c} \xmapsto{\text{typedef}} \text{int} \xmapsto{\text{sizeof}} \text{sizeof(int)}$$

Consider what Sanders and Kandrot says about the pointer to the pointer that (you want to) holds the address of the newly allocated memory. [19] Consider this diagram:

$$\textbf{pointers} \xrightarrow{*} \textbf{pointers} \xrightarrow{*} \textbf{Types}$$

$$\textbf{pointer} \xrightarrow{*} \textbf{pointer} \xrightarrow{*} \text{void}$$

$$\&\texttt{dev\_c} \xrightarrow{*} *(\&\texttt{dev\_c}) \xrightarrow{*} (\text{void}**)(\&\texttt{dev\_c})$$

I propose that what `cudaMalloc` does (actually) is the following:

$$\begin{array}{ccccccc} \textbf{Memory}_{\text{GPU}} & \xrightarrow{\text{cudaMalloc}} & \textbf{pointers} & \xrightarrow{*} & \textbf{pointers} & \xrightarrow{*} & \textbf{Types} \\ & & \downarrow{\scriptstyle *} & & & & \\ & & \textbf{pointers}_{\text{GPU}} & \xrightarrow{*} & \textbf{Types} & & \end{array} \tag{63}$$

$$\begin{array}{ccccccc} \text{Memory address}_{\text{GPU}} & \xrightarrow{\text{cudaMalloc}} & \&\texttt{dev\_c} & \xmapsto{*} & *(\&\texttt{dev\_c}) & \xmapsto{*} & (\text{void}**)(\&\texttt{dev\_c}) \\ & & \downarrow{\scriptstyle *} & & & & \\ & & \texttt{dev\_c} & \xmapsto{*} & *\texttt{dev\_c} & & \end{array}$$

`dev_c` is now a *device pointer*, available to kernel functions on the GPU.
Syntax-wise, we can relate this diagram to the corresponding function "usage":

$$\textbf{pointers} \times \mathbb{N}^+ \xrightarrow{\text{cudaMalloc}} \texttt{cudaError\_r}$$

$$((\text{void}**)(\&\texttt{dev\_c}), (\text{sizeof}(\text{int}))) \xmapsto{\text{cudaMalloc}} \text{cudaSuccess (for example)}$$

$\Longrightarrow$

```
cudaMalloc((void**)&dev_c, sizeof(int))
```

For practice, consider now `cudaMemcpy` in the functor interpretation, and its definition as such:

`cudaMemcpy` is a "functor category", s.t. we equip the functor `cudaMemcpy` with a collection of objects $\text{Obj}_{\text{cudaMemcpy}}$, s.t., for example, `cudaMemcpyDevicetoHost` $\in \text{Obj}_{\text{cudaMemcpy}}$, where

$$(\text{cudaMemcpy}(-,-,n_{\text{thread}}, \text{cudaMemcpyDevicetoHost}) : \textbf{Memory}_{GPU} \to \textbf{Memory}_{CPU}) \in \text{Hom}(\textbf{Memory}_{GPU}, \textbf{Memory}_{CPU})$$

where $\text{Obj}\textbf{Memory}_{GPU} \equiv$ collection of all possible memory (addresses) on GPU.

It should be noted that, syntax-wise, $\&c \in \text{Obj}\textbf{Memory}_{CPU}$ and $\&c$ belongs in the "first slot" of the arguments for cudaMemcpy, whereas `dev_c` $\in \textbf{pointers}_{GPU}$ a *device pointer*, is "passed in" to the "second slot" of the arguments for cudaMemcpy.

32. Threads, Blocks, Grids

cf. Chapter 5 Thread Cooperation, Section 5.2. Splitting Parallel Blocks of Sanders and Kandrot (2010) [19].

Consider first a 1-dimensional block.

- `threadIdx.x` $\Longleftarrow M_x \equiv$ number of threads per block in x-direction. Let $j_x = 0 \dots M_x - 1$ be the index for the thread. Note that $1 \leq M_x \leq M_x^{\max}$, e.g. $M_x^{\max} = 1024$, max. threads per block
- `blockIdx.x` $\Longleftarrow N_x \equiv$ number of blocks in x-direction. Let $i_x = 0 \dots N_x - 1$
- `blockDim` stores number of threads along each dimension of the block M_x.

Then if we were to "linearize" or "flatten" in this x-direction,

$$k = j_x + i_x M_x$$

where k is the kth thread. $k = 0 \dots N_x M_x - 1$.

Take a look at heattexture1.cu which uses the GPU texture memory. Look at how `threadIdx`/`blockIdx` is mapped to pixel position.

As an exercise, let's again rewrite the code in mathematical notation:

- `threadIdx.x` $\Longleftarrow j_x$, $0 \leq j_x \leq M_x - 1$
- `blockIdx.x` $\Longleftarrow i_x$, $0 \leq i_x \leq N_x - 1$
- `blockDim.x` $\Longleftarrow M_x$, number of threads along each dimension (here dimension x) of a block, $1 \leq M_x \leq M_x^{\max} = 1024$
- `gridDim.x` $\Longleftarrow N_x$, $1 \leq N_x$

resulting in

- $k_x = j_x + i_x M_x \Longrightarrow$

```
int x = threadIdx.x + blockIdx.x * blockDim.x ;
```

- $k_y = j_y + i_y M_y \Longrightarrow$

```
int y = threadIdx.y + blockIdx.y * blockDim.y ;
```

and so for a "flattened" thread index $J \in \mathbb{N}$,

$$J = k_x + N_x \cdot M_x \cdot k_y$$

$\Longrightarrow$

```
offset = x + y * blockDim.x * gridDim.x ;
```

Suppose vector is of length N. So we *need* N parallel threads to launch, in total.
e.g. if $M_x = 128$ threads per block, $N/128 = N/M_x$ blocks to get our total of N threads running.
Wrinkle: integer division! e.g. if $N = 127$, $\frac{N}{128} = 0$.
Solution: consider $\frac{N+127}{128}$ blocks. If $N = l \cdot 128 + r$, $l \in \mathbb{N}$, $r = 0 \dots 127$.

$$\frac{N+127}{128} = \frac{l \cdot 128 + r + 127}{128} = \frac{(l+1)128 + r - 1}{128} =$$
$$= l + 1 + \frac{r-1}{128} = \begin{cases} l & \text{if } r = 0 \\ l+1 & \text{if } r = 1 \dots 127 \end{cases}$$

$$\frac{N + (M_x - 1)}{M_x} = \frac{l \cdot M_x + r + M_x - 1}{M_x} = \frac{(l+1)M_x + r - 1}{M_x} =$$
$$= l + 1 + \frac{r-1}{M_x} = \begin{cases} l & \text{if } r = 0 \\ l+1 & \text{if } r = 1 \dots M_x - 1 \end{cases}$$

So $\frac{N+(M_x-1)}{M_x}$ is the smallest multiple of M_x greater than or equal to N, so $\frac{N+(M_x-1)}{M_x}$ **blocks are needed or more than needed to run a total of N threads.**

Problem: Max. grid dim. in 1-direction is 65535, $\equiv N_i^{\max}$.
So $\frac{N+(M_x-1)}{M_x} = N_i^{\max} \Longrightarrow N = N_i^{\max} M_x - (M_x - 1) \leq N_i^{\max} M_x$. i.e. number of threads N is limited by $N_i^{\max} M_x$.
Solution.

- number of threads per block in x-direction $\equiv M_x \Longrightarrow$ `blockDim.x`

- number of blocks in grid $\equiv N_x \Longrightarrow$ `gridDim.x`
- N_xM_x total number of threads in x-direction. Increment by N_xM_x. So next scheduled execution by GPU at the $k = N_xM_x$ thread.

Sanders and Kandrot (2010) [19] made an important note, on pp. 176-177 Ch. 9 Atomics of Section 9.4 Computing Histograms, an important *rule of thumb* on the number of blocks.

First, consider N^{threads} total threads. The extremes are either N^{threads} threads on a single block, or N^{threads} blocks, each with a single thread.

Sanders and Kandrot gave this tip:

number of blocks, i.e. `gridDim.x` $\Longleftarrow N_x \sim 2\times$ number of GPU multiprocessors, i.e. twice the number of GPU multiprocessors. In the case of my GeForce GTX 980 Ti, it has 22 Multiprocessors.

32.1. **global thread Indexing: 1-dim., 2-dim., 3-dim.** Consider the problem of *global thread indexing.* This was asked on the NVIDIA Developer's board (cf. Calculate GLOBAL thread Id). Also, there exists a "cheatsheet" (cf. CUDA Thread Indexing Cheatsheet). Let's consider a (mathematical) generalization.

Consider again (cf. 32) the following notation:

- `threadIdx.x` $\Longleftarrow i_x$, $0 \leq i_x \leq M_x - 1$, $\quad i_x \in \{0 \dots M_x - 1\} \equiv I_x$, of "cardinal length/size" of $|I_x| = M_x$
- `blockIdx.x` $\Longleftarrow j_x$, $0 \leq j_x \leq N_x - 1$, $\quad j_x \in \{0 \dots N_x - 1\} \equiv J_x$, of "cardinal length/size" of $|J_x| = N_x$
- `blockDim.x` $\Longleftarrow M_x$
- `gridDim.x` $\Longleftarrow N_x$

Now consider formulating the various cases, of a grid of dimensions from 1 to 3, and blocks of dimensions from 1 to 3 (for a total of 9 different cases) mathematically, as the CUDA Thread Indexing Cheatsheet did, similarly:

- *1-dim. grid* of *1-dim. blocks.* Consider $J_x \times I_x$. For $j_x \in J_x$, $i_x \in I_x$, then $k_x = j_xM_x + i_x$, $k_x \in \{0 \dots N_xM_x - 1\} \equiv K_x$. The condition that k_x be a valid global thread index is that K_x has equal cardinality or size as $J_x \times I_x$, i.e.

$$|J_x \times I_x| = |K_x|$$

(this must be true). This can be checked by checking the most extreme, maximal, case of $j_x = N_x - 1$, $i_x = M_x - 1$:

$$k_x = j_xM_x + i_x = (N_x - 1)M_x + M_x - 1 = N_xM_x - 1$$

and so k_x ranges from 0 to $N_xM_x - 1$, and so $|K_x| = N_xM_x$.

Summarizing all of this in the following manner:

$$J_x \times I_x \longrightarrow K_x \equiv K^{N_xM_x} = \{0 \dots N_xM_x - 1\}$$

$$(j_x, i_x) \longmapsto k_x = j_xM_x + i_x$$

For the other cases, this generalization we've just done is implied.

- *1-dim. grid* of *2-dim. blocks*

$$J_x \times (I_x \times I_y) \longrightarrow K^{N_xM_xM_y} \equiv \{0 \dots N_xM_xM_y - 1\}$$

$$(j_x, (i_x, i_y)) \longmapsto k = j_xM_xM_y + (i_x + i_yM_x) = j_x|I_x \times I_y| + (i_x + i_yM_x) \in \{0 \dots N_xM_xM_y - 1\}$$

The "most extreme, maximal" case that can be checked to check that the "cardinal size" of $K^{N_xM_xM_y}$ is equal to $J_x \times (I_x \times I_y)$ is the following, and for the other cases, will be implied (unless explicitly written or checked out):

$$k = j_xM_xM_y + (i_x + i_yM_x) = (N_x - 1)M_xM_y + ((M_x - 1) + (M_y - 1)M_x) = (N_xM_xM_y - 1)$$

The thing to notice is this emerging, general pattern, what could be called a "global view" of understanding the threads and blocks model of the GPU (cf. njuffa's answer:

$$\text{total number of threads} = \text{block index (Id)} \cdot \text{total number of threads per blcok} + \text{thread index on the block}$$

But as we'll see, that's not the only way of "flattening" the index, or transforming into a 1-dimensional index.

- *1-dim. grid* of *3-dim. blocks*

$$J_x \times (I_x \times I_y \times I_z) \longrightarrow K^{N_xM_xM_yM_z}$$

$$(j_x, (i_x, i_y, i_z)) \longmapsto k = j_x(M_xM_yM_z) + (i_x + i_yM_x + i_ZM_xM_y) \in \{0 \dots N_xM_xM_yM_z - 1\}$$

- *2-dim. grid* of *1-dim. blocks*

$$(J_x \times J_y) \times I_x \longrightarrow L^{N_xN_y} \times I_x \longrightarrow K^{N_xN_yM_x}$$

$$((j_x, j_y), i_x) \longmapsto ((j_x + N_xj_y), i_x) \longmapsto k = (j_x + N_xj_y) \cdot M_x + i_x \in \{0 \dots N_xN_yM_x - 1\}$$

- *2-dim. grid* of *2-dim. blocks*

$$(J_x \times J_y) \times (I_x, I_y) \longrightarrow L^{N_xN_y} \times (I_x, I_y) \longrightarrow K^{N_xN_yM_x}$$

$$((j_x, j_y), (i_x, i_y)) \longmapsto ((j_x + N_xj_y), (i_x, i_y)) \longmapsto k = (j_x + N_xj_y) \cdot M_xM_y + i_x + M_xi_y$$

But this *isn't the only way of obtaining* a "flattened index." Exploit the commutativity and associativity of the Cartesian product:

$$J_x \times J_y \times I_x \times I_y = (J_x \times I_x) \times (J_y \times I_y) \longrightarrow K^{N_xM_x} \times K^{N_yM_y} \longrightarrow K^{N_xN_yM_xM_y}$$

$$((j_x, j_y, i_x, i_y) = ((j_x, i_x), (j_y, i_y)) \longmapsto (i_x + M_xj_x, i_y + M_yj_j) \equiv (k_x, k_y) \longmapsto \begin{array}{c} k = k_x + k_yN_xM_x = \\ = (i_x + M_xj_x) + (i_y + M_yj_y)M_xN_x \end{array}$$

Indeed, checking the "maximal, extreme" case,

$$k = k_x + k_yN_xM_x = M_xN_x - 1 + (M_yN_y - 1)(N_xM_x) = M_yM_yN_xM_x - 1$$

and so k ranges from 0 to $M_yM_yN_xM_x - 1$.

- *3-dim. grid* of *3-dim. blocks*

$$\begin{array}{c} (J_x \times J_y \times J_z) \times (I_x \times I_y \times I_z) = \\ = (J_x \times I_x) \times (J_y \times I_y) \times (J_z \times I_z) \end{array} \longrightarrow K^{N_xM_x} \times K^{N_yM_y} \times K^{N_zM_z} \longrightarrow K^{N_xN_yN_zM_xM_yM_z}$$

$$\begin{array}{c} ((j_x, j_y, j_z), (i_x, i_y, i_z)) = \\ = ((j_x, i_x), (j_y, i_y), (j_z, i_z)) \end{array} \longmapsto \begin{array}{c} (i_x + M_xj_x, i_y + M_yj_j, i_z + M_zj_z) \equiv \\ \equiv (k_x, k_y, k_z) \end{array} \longmapsto k = k_x + k_yN_xM_x + k_zN_xM_xN_yM_y$$

Indeed, checking the "extreme, maximal" case for k:

$$k = k_x + k_yN_xM_x + k_zN_xM_xN_yN_y =$$
$$= (N_xM_x - 1) + (N_yM_y - 1)N_xM_x + (N_zM_z - 1)N_xM_xN_yM_y = N_xN_yN_zM_xM_yM_z - 1$$

33. Row-Major ordering vs. Column Major ordering, as flatten

So-called row-major ordering and column major ordering should be formalized, to deal with continguous memory access in reading or writing to a matrix, or lack thereof.

Given

$$\begin{aligned} A &\in \mathrm{Mat}_{\mathbb{R}}(m,n) \\ A &: \{1,2,\dots m\} \times \{1,2,\dots n\} \to \mathbb{R} \\ A &: (i,j) \mapsto A(i,j) \in \mathbb{R} \\ \text{or} \quad A &: \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} \to \mathbb{R} \\ A &: (i,j) \mapsto A(i,j) \in \mathbb{R} \end{aligned}$$

Consider isomorphism "flatten":

$$\begin{aligned} \mathrm{Mat}_{\mathbb{R}}(m,n) &\xrightarrow{\text{flatten}} \mathbb{R}^{mn} \\ \{1,2,\dots m\} \times \{1,2,\dots n\} &\to \{1,2,\dots mn\} \\ \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} &\to \{0,1,\dots mn-1\} \end{aligned} \tag{64}$$

There are 2 kinds of flatten:

Row-major ordering is the one we're (psychologically) used to, if we read continguously from left to right, horizontally, along a row.

Definition 10 (row-major ordering)**.**

$$\begin{array}{cc} \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} \to \{0,1,\dots mn-1\} & \{0,1,\dots mn-1\} \to \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} \\ (i,j) \mapsto in+j & k \mapsto (k/n, k \mod n) \\ \{1,2,\dots m\} \times \{1,2,\dots n\} \to \{1,2,\dots mn\} & \{1,2,\dots mn\} \to \{1,2,\dots m\} \times \{1,2,\dots n\} \\ (i,j) \mapsto (i-1)n+j & k \mapsto (\lceil k/n \rceil, k \mod n) \end{array} \tag{65}$$

Definition 11 (column-major ordering)**.**

$$\begin{array}{cc} \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} \to \{0,1,\dots mn-1\} & \{0,1,\dots mn-1\} \to \{0,1,\dots m-1\} \times \{0,1,\dots n-1\} \\ (i,j) \mapsto i+jm & k \mapsto (k \mod m, k/m) \\ \{1,2,\dots m\} \times \{1,2,\dots n\} \to \{1,2,\dots mn\} & \{1,2,\dots mn\} \to \{1,2,\dots m\} \times \{1,2,\dots n\} \\ (i,j) \mapsto i+(j-1)m & k \mapsto (k \mod m, \lceil k/m \rceil) \end{array} \tag{66}$$

33.1. **(CUDA) Constant Memory.** cf. Chapter 6 Constant Memory of Sanders and Kandrot (2010) [19]

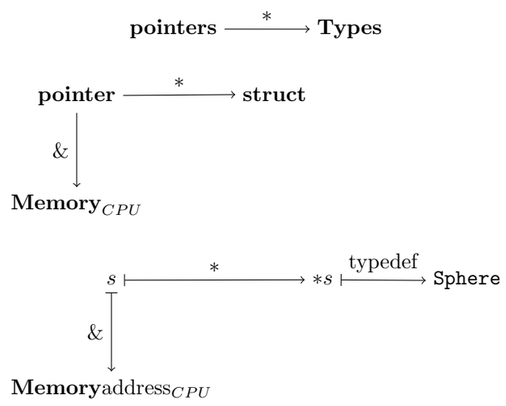
Refer to the ray tracing examples in Sanders and Kandrot (2010) [19], and specifically, here: raytrace.cu, rayconst.cu.

Without constant memory, then this had to be done:

- *definition* (in the code) - Consider **struct** as a subcategory of **Types** since **struct** itself is a category, equipped with objects and functions (i.e. methods, modules, etc.).
  So for **struct**, Obj**struct** ∋ `Sphere`. $\implies$

```
struct sphere { ... }
```

- Usage, "instantiation", i.e. creating, or "making" it (the `struct`):



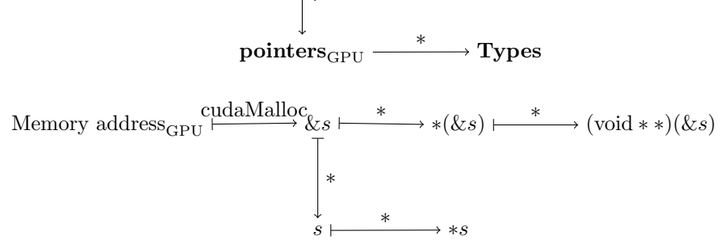
$\implies$

```
Sphere *s
```

Recalling Eq. 63, for `SPHERES` == 40 (i.e. for example, 40 spheres)

```
cudaMalloc((void **) &s, sizeof(Sphere)*SPHERES)
```

$\impliedby$



and syntax-wise,

$$\textbf{pointers} \times \mathbb{N}^{+} \xrightarrow{\text{cudaMalloc}} \texttt{cudaError\_r}$$

$$((\text{void} **)(s), \text{sizeof(Sphere)} * \text{SPHERES}) \xmapsto{\text{cudaMalloc}} \text{cudaSuccess (for example)}$$

Now consider

```
cudaMemcpy(s, temp_s, sizeof(Sphere) * SPHERES, cudaMemcpyHostToDevice)
```

$$\begin{array}{c} \text{cudaMemcpy}(s, temps, \text{sizeof(Sphere)} * \text{SPHERES}, \text{cudaMemcpyHostToDevice}) \\ \textbf{Memory}_{CPU} \longrightarrow \textbf{Memory}_{GPU} \end{array}$$

$$\begin{array}{ccc} \textbf{Memory}\text{address}_{CPU} & \longmapsto & \textbf{Memory}\text{address}_{GPU} \\ * \downarrow & & * \downarrow \\ \texttt{temp\_s} & \longmapsto & s \end{array}$$

The lesson then is this, in light of how long ray tracing takes with constant memory and without constant memory - `cudaMemcpy` between host to device, CPU to GPU, is a costly operation. Here, in this case, we're copying from the host memory to memory on the GPU. It copies to a global memory on the GPU.

Now, using **constant memory**,
we no longer need to do `cudaMalloc`, allocate memory on the GPU, for s, pointer to a `Sphere`.

Instead, we have

```
__constant__ Sphere s[SPHERES];
```

In this particular case, we want it to have global scope.
Note, it is still on host memory.

Notice that

$$\begin{array}{c} \text{cudaMemcpyHostToDevice}(-, -, \text{sizeof(Sphere)} * \text{SPHERES}) \\ \textbf{Memory}_{CPU} \longrightarrow \textbf{Memory}_{GPU} \end{array}$$

$$\begin{array}{ccc} \textbf{Memory}\text{address}_{CPU} & \longmapsto & \textbf{Memory}\text{adddress}_{GPU} \\ * \downarrow & & * \downarrow \\ \texttt{temp\_s} & \longmapsto & s \\ \text{typedef} \downarrow & & \text{typedef} \downarrow \\ \text{array of Sphere's} & \longmapsto & \text{array of Sphere's} \end{array}$$

So notice that we have a bijection, and on one level, we can think of the bijection from `temp_s`, an array of Sphere's to s, an array of Sphere's. So notice that the types and memory size of `temp_s` and s must match.

And for this case, that's all there is to *constant memory*. What's going on involves the so-called *warp*, a collection of threads, "woven together" and get executed in lockstep. NVIDIA hardware broadcasts a single memory read to each half-warp. "If every thread in a half-warp requests data from the same address in constant memory, your GPU will generate only a single read request and subsequently broadcast the data to every thread." (cf. Sanders and Kandrot (2010) [19]). Furthermore, "the hardware can aggressively cache the constant data on the GPU."

33.2. **(CUDA) Texture Memory.**

33.3. **Do (smooth) manifolds admit a triangulation?** Topics in Geometric Topology (18.937)

Piecewise Linear Topology (Lecture 2)

Part 8. Monte Carlo methods

34. Monte Carlo methods

cf. Sec. 11.1.3 *Second Illustration, Particles in a Box*, Hjorth-Jensen (2015) [17].

Recall that a random variable $X : \Omega \to E$, $\Omega \equiv$ set of possible outcomes, $E \equiv$ measurable space, with $E = \mathbb{R}$ (usually).

Probability that X takes on value in measurable set $S \subseteq E$,

$$\Pr(X \in S) = P(\{\omega \in \Omega | X(\omega) \in S\})$$

If we assume "all particles in left half have equal probabilities of going to the right half" (pp. 346, Sec. 11.1.3 of Hjorth-Jensen (2015) [17]),

Let $\Omega_{LR,RL} = \{L \to R, R \to L\}$, where
$L \to R \equiv$ particle is on left half and goes to the right,
$R \to L \equiv$ particle is on right half and goes to the left,

Let $X_{LR,RL} : \Omega_{LR,RL} \to \mathbb{R}$ with
$X_{LR,RL} : L \to R \mapsto 0$
$X_{LR,RL} : R \to L \mapsto 1$

Let $\Omega_{LR} = \{L, R\}$, where
$L \equiv$ particle on left half,
$R \equiv$ particle on right half.

Let $X_{LR} : \Omega_{LR} \to \{0, 1\}$ where
$X_{LR} : L \mapsto 0$
$X_{LR} : R \mapsto 1$.

Let $n_l \equiv$ number of particles at any time on the left side, and
let $n_r = N - n_l$ for those on the right side.

The key insight is that at any time,

$$\Pr(X_{LR} = 0) = P(\{L\}) = \frac{n_l}{N}$$
$$\Pr(X_{LR} = 1) = P(\{R\}) = \frac{n_r}{N}$$

Assume particle must move at each time step.
Let state of the system be characterized by $n_l = 0, 1, \ldots N$.
Consider $\forall t = 0, 1, \ldots,$

$$n_l = n_l(t) = \begin{cases} n_l(t-1) - 1 & \text{if } L \to R \\ n_l(t-1) + 1 & \text{if } R \to L \end{cases}$$

i.e.

$$n_l(t-1) \xmapsto{X_{LR,RL}} \begin{cases} n_l(t-1) - 1 = n_l(t) & \text{if } X_{LR,RL} = 0 \\ n_l(t-1) + 1 = n_l(t) & \text{if } X_{LR,RL} = 1 \end{cases}$$

Let a system's state s be characterized by a finite number of state variables: $s = (s_1, s_2, \ldots s_n)$. So for $s \in S$, S, the space of all possible states, can be infinite (infinite number of states S that s can be in), but s is characterized by a finite number of variables (parameters), such as only energy and pressure, only.

Consider time evolution operators in general:
Consider system with state space X, time evolution $F_{t,s}$ is a mapping (for 2 different cases)

$$F_{t,s} : X \to X \qquad \forall t, s \in \mathbb{R}$$
$$F_{t,s} : X \to X \qquad \forall t, s \in \mathbb{Z}$$

s.t.

$$F_{u,t}(F_{t,s}(x)) = F_{u,s}(x)$$

Time evolution F is **homogeneous**, if

$$F_{u,t} = F_{u-t,0} \qquad \forall\, u,t \in \mathbb{R}(\text{or } \mathbb{Z})$$

In this case of a homogeneous system, mappings $G_t = F_{t,0}$ form a 1-parameter group of transformations of X, i.e.

$$G_{t+s} = G_t G_s$$

Specialize to homogeneous time evolution G_t, $t \in \mathbb{Z}$, so that $G_t = (G_1)^t$, $\forall\, t \in \mathbb{Z}$.
Thus, consider $G_1 : S \to S$.
Let $\Omega_{S\to S} = \{s \mapsto s'\}$, $\forall\, s, s' \in S$, $s \mapsto s'$ that are allowed or considered.
Let $X : \Omega_{S\to S} \to E$.
Assign $P(s \mapsto s')$ $\quad \forall\, s \mapsto s' \in \Omega_{S\to S}$.
Thus, apply $G_1 : S \to S$ according to $X = X : \Omega_{S\to S} \to E$.
Thus, given $\boxed{S, \Omega_{S\to S}}$ (all possible state transitions, not all permutations), random variable $\boxed{X = X : \Omega_{S\to S} \to E}$, where $E = \mathbb{R}$ (usually), and P s.t.

$$\boxed{\Pr(X \in S) = P(\{\omega \in \Omega | X(\omega) \in S\}) \text{ with } S \subseteq E} \tag{67}$$

then

$$\boxed{G_1 : S \to S}, \text{ governed by } X \tag{68}$$

34.1. **Radioactive decay via Monte Carlo.** Given $\omega \equiv$ transition probability.

$$\frac{dN(t)}{dt} = -\omega N(t) N(t) = N(0) e^{-\omega t}$$

For single type of nuclei (called "X").
Let $S_i = \{0,1\} = \mathbb{Z}_2$ (decay, no decay).

$$\Omega_{S_i \to S_i} = \{1 \to 0, 1 \to 1\}$$
$$X_i = X_i : \Omega_{S_i \to S_i} \to E = \mathbb{R}$$
$$X_i : 1 \to 0 \mapsto 0$$
$$X_i : 1 \to 1 \mapsto 1$$
$$\Pr(X_i = 0) = \omega$$
$$\Pr(X_i = 1) = 1 - \omega$$
$$G_1 : S_i \to S_i$$
$$G_1 : 1 \mapsto \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X = 1 \end{cases}$$

$\forall$ time t,

$$S_t = \{N(t)\}$$
$$S_{t+1} = \{0, 1, \dots N(t)\}$$
$$\Omega_{S_t \to S_{t+1}} = \{N(t) \to 0, N(t) \to 1, \dots N(t) \to N(t)\}$$
$$X = X : \Omega_{S_t \to S_{t+1}} \to E = \mathbb{R}$$
$$X : N(t) \to 0 \mapsto 0$$
$$X : N(t) \to 1 \mapsto 1$$
$$\vdots$$
$$X : N(t) \to n(t) \mapsto N(t)$$
$$\Pr(X = 0) = \omega^{N(t)}$$
$$\Pr(X = 1) = \omega^{N(t)-1}(1-\omega)$$
$$\vdots$$
$$\Pr(X = N(t)) = (1-\omega)^{N(t)}$$
$$G_1 : S_t \to S_{t+1}$$
$$G_1 : N(t) \to \begin{cases} 0 & \text{if } X = 0 \\ 1 & \text{if } X = 1 \\ \vdots & \\ N(t) & \text{if } X = N(t) \end{cases}$$

Clearly, $X = \sum_{i=0}^{N(t)} X_i$ with $P(X) = \prod_{i=0}^{N(t)} P(X_i)$ and so simply sum the result of X_i's to obtain $N(t+1)$.

34.1.1. *2 nuclei radioactive decay with Monte Carlo method.* If nucleus "X" decays to daughter nucleus "Y", which can also decay, then the system of first-order differential equations describing the system are

$$\frac{dN_X(t)}{dt} = -\omega_X N_X(t)$$
$$\frac{dN_Y(t)}{dt} = -\omega_Y N_Y(t) + \omega_X N_X(t)$$

Consider

$$\Omega_{S\to S} = \{X \to X, X \to Y, Y \to 0, Y \to Y\}$$
$$\Pr(X = X \to X) = 1 - \omega_X$$

Let $s = (x, y) \in S = \{(1,0), (0,1), (0,0)\}$.

$$\Omega_{S\to S} = \{X \to X, X \to Y, Y \to 0, Y \to Y\} \equiv$$
$$\{(1,0) \to (1,0), (1,0) \to (0,1), (0,1) \to (0,0), (0,1) \to (0,1)\}$$

So for $X : \Omega_S \to S \to E = \mathbb{R}$,

$$p(X \to X) = 1 - \omega_X$$
$$p(X \to Y) = \omega_X$$
$$p(Y \to 0) = \omega_Y$$
$$p(Y \to Y) = 1 - \omega_Y$$
$$G_1 : (1,0) \mapsto \begin{cases} (1,0) & \text{if } X = X \to X \\ (0,1) & \text{if } X = X \to Y \end{cases}$$
$$G_1 : (0,1) \mapsto \begin{cases} (1,0) & \text{if } Y = Y \to 0 \\ (0,1) & \text{if } Y = Y \to Y \end{cases}$$

35. Ising Model

EY : 20171223 Things to check from Hjorth-Jensen (2015) [17]:

2-dim. Ising model, with $\mathcal{B} \equiv h_j = 0$, undergoes phase transition of 2nd. order: meaning below given critical temperature T_C, there's spontaneous magnetization with $\langle \mathcal{M} \rangle \equiv \langle \mathbf{M} \rangle \neq 0$. $\langle \mathbf{B} \rangle \to 0$ at T_C with *infinite* slope, a behavior called *critical phenomena*. Critical phenomenon normally marked by 1 or more thermodynamical variables which is 0 above a critical point. In this case, $\langle \mathbf{B} \rangle \neq 0$, such a parameter normally called *order parameter*.

Critical phenomena; we still don't have a satisfactory understanding of system's properties close to the critical point, even for simplest 3-dim. systems. Even mean-field models can predict wrong physics; mean-field theory results in a 2nd.-order phase transition for 1-dim. Ising model, wherea 1-dim. Ising model doesn't predict any spontaneous magnetization at any finite temperature T.

e.g. Consider 1-dim. N-spin system. Assume periodic boundary conditions. Consider state of all spins up, with total energy $-NJ$ and magnetization N. Flip half of spins (e.g. all spins of index $i > N/2$) so 1st half of spins point upwards and last half points downwards. Energy is $-NJ + 4J$, net magnetization 0. This is an example of a possible disordered state with net magnetization 0. Change in energy is too small to stabilize disordered state (to $-NJ$).

35.1. Metropolis algorithm setup.

Let

selection probabilities $g(\mu, \nu) :=$ prob. that state ν selected by algorithm out of all states, given that 1 is in state μ.

acceptance probabilities $A(\mu, \nu)$; if new state ν accepted, move to state ν; if ν is not accepted, then we stay in μ

Given $L = |\Lambda|$ total sites on lattice, given single spin-flip, then there are L total new states ν from present state μ. Metropolis algorithm assumes selection probability equal to the L states $g(\mu, \nu) = \frac{1}{L}$.

Suppose system in thermal equilibrium, so system's energy only fluctuates within small range.

EY:20171223

$$\begin{array}{lll} \uparrow\downarrow\uparrow \mapsto \uparrow\uparrow\uparrow & 2J \mapsto -2J & \Delta E = -4J \\ \uparrow\uparrow\downarrow \mapsto \uparrow\uparrow\uparrow & 0 \mapsto -2J & \Delta E = -2J \\ \uparrow\uparrow\uparrow \mapsto \uparrow\downarrow\downarrow & 0 \mapsto 0 & \Delta E = 0 \\ \uparrow\uparrow\uparrow \mapsto \uparrow\downarrow\uparrow & 0 \mapsto 2J & \Delta E = 2J \\ \uparrow\uparrow\uparrow \mapsto \uparrow\downarrow\uparrow & 0 \mapsto 2J & \Delta E = 2J \end{array}$$

$2^3 = 8$. $8(8-1) = 8 \cdot 7 = 56$ transitions. $\frac{56}{2} = 28$ (up down symmetry)

Nevertheless, by detailed balance

$$\frac{P(\mu,\nu)}{P(\nu,\mu)} = \frac{g(\mu,\nu)A(\mu,\nu)}{g(\nu,\mu)A(\nu,\mu)} = \frac{A(\mu,\nu)}{A(\nu,\mu)} = \frac{P_\beta(\nu)}{P_\beta(\mu)} = \frac{\frac{1}{2}\exp(-\beta(H_\nu))}{\frac{1}{2}\exp(-\beta(H_\mu))} = \exp(-\beta(H_\nu - H_\mu))$$

Thus, select acceptance probability for our algorithm to satisfy:

$$\frac{A(\mu,\nu)}{A(\nu,\mu)} = \exp(-\beta(H_\nu - H_\mu)) \tag{69}$$

If $H_\mu < H_\nu$, then $A(\nu,\mu) > A(\mu,\nu)$ (go to lower energies).

Thus, the Metropolis acceptance algorithm is

$$A(\mu,\nu) = \begin{cases} \exp(-\beta(H_\nu - H_\mu)) & \text{if } H_\nu - H_\mu > 0 \\ 1 & \text{otherwise} \end{cases} \tag{70}$$

Compare the algorithm's steps given in Wikipedia, "Ising model" vs. pp. 435, 13.5 The Metropolis Algorithm and the 2-dim. Ising Model of Hjorth-Jensen (2015) [17], respectively:

(1) Pick spin site using selection probability $g(\mu,\nu)$ and calculate contribution to energy involving this spin, $E_b \equiv E_\mu$.
(2) Flip spin and calculate new contribution, E_t.
(3) If $E_t < E_\mu$, keep flipped value.
(4) If $E_t > E_\mu$, only keep with probability $\exp(-\beta(H_\nu - H_\mu))$
(5) Repeat

(1) Establish initial state with energy E_b by positioning yourself at random configuration in lattice
(2) Change initial configuration by flipping 1 spin only. Compute energy of this trial state E_t.
(3) Calculate $\Delta E_t - E_b$.
(4) If $\Delta E \leq 0$, accept new configuration.
(5) If $\Delta E > 0$, calculate $w = \exp(-(\beta \Delta E))$.
(6) Compare w with random number r. If $r \leq w$, accept new configuration: else keep old configuration.
(7) Next, update various expectation values.
(8) Repeat.

Hjorth-Jensen (2015) [17] notes that each time you sweep through lattice, i.e., when you've summed over all spins, this constitutes a "Monte Carlo cycle." One could think of one of such a cycle as a measurement. At the end, you should divide the expectation value by total number of cycles. You can choose to divide by number of spins or not. If you divide with number of spins as well, result, e.g. for the energy is now energy per spin.

Calculate all possible $\Delta E \equiv H_\nu - H_\mu$ beforehand for a given Λ_d for computational efficiency.

Since all we care about is $\Delta E \equiv H_\nu - H_\mu$, use symmetry to find unique values of $\Delta E \equiv H_\nu - H_\mu$.

There are $2^{|\mathbf{y}(\mathbf{k})|}$ total possible spin configurations, given a chosen lattice site $\mathbf{k}$, and its (nearest-neighbor) edges $\mathbf{y}(\mathbf{k})$.

For $d = 1$, start with $2^{|\mathbf{y}(\mathbf{k})|} = 2^2$. By symmetry, subtract as follows:

$$2^2 - \left(\binom{2}{1} - 1\right) = 2^2 - 1 = 3$$

Then unique $\Delta E \equiv H_\nu - H_\mu$ changes for single spin flips are

$$\begin{array}{lll} \uparrow\uparrow\uparrow \mapsto \uparrow\downarrow\uparrow & -2J \mapsto 2J & \Delta E = 4J \\ \uparrow\uparrow\downarrow \mapsto \downarrow\uparrow\downarrow & 0 \mapsto 2J & \Delta E = 2J \\ \uparrow\uparrow\downarrow \mapsto \uparrow\downarrow\downarrow & 0 \mapsto 0 & \Delta E = 0 \\ \uparrow\uparrow\downarrow \mapsto \uparrow\uparrow\uparrow & 0 \mapsto -2J & \Delta E = -2J \\ \downarrow\uparrow\downarrow \mapsto \downarrow\downarrow\downarrow & 2J \mapsto -2J & \Delta E = -4J \end{array}$$

Likewise, for $d = 2$,

$$2^{|\mathbf{y}(\mathbf{k})|} = 2^4 = 16$$

$$2^4 - \left(\binom{4}{1} - 1\right) - \left(\binom{4}{2} - 1\right) - \left(\binom{4}{3} - 1\right) = 2^4 - (4-1) - (6-1) - (4-1) = 16 - 3 - 5 - 3 = 5$$

Expect 5 unique H_μ configurations.

$$\begin{array}{lll} \begin{array}{c}\uparrow \\ \uparrow\uparrow\uparrow \\ \uparrow\end{array} \mapsto \begin{array}{c}\uparrow \\ \uparrow\downarrow\uparrow \\ \uparrow\end{array} & -4J \mapsto 4J & \Delta E = 8J \\ \begin{array}{c}\uparrow \\ \uparrow \\ \downarrow\uparrow\uparrow\end{array} \mapsto \begin{array}{c}\uparrow \\ \uparrow \\ \downarrow\downarrow\uparrow\end{array} & -2J \mapsto 2J & \Delta E = 4J \\ \begin{array}{c}\uparrow \\ \uparrow \\ \downarrow\uparrow\uparrow\end{array} \mapsto \begin{array}{c}\uparrow \\ \uparrow \\ \downarrow\downarrow\uparrow\end{array} & 0 \mapsto 0 & \Delta E = 0J \\ \begin{array}{c}\downarrow \\ \downarrow \\ \uparrow\uparrow\downarrow\end{array} \mapsto \begin{array}{c}\downarrow \\ \downarrow \\ \uparrow\downarrow\downarrow\end{array} & 2J \mapsto -2J & \Delta E = -4J \\ \begin{array}{c}\downarrow \\ \downarrow \\ \downarrow\uparrow\downarrow\end{array} \mapsto \begin{array}{c}\downarrow \\ \downarrow \\ \downarrow\downarrow\downarrow\end{array} & 4J \mapsto -4J & \Delta E = -8J \end{array}$$

Near critical point methods include Wolff algorithm; I found a parallelized version; cf. Parallelization of the Wolff Single-Cluster Algorithm; a review of these "new Monte Carlo algorithms" is given by Barkema and Newman; cf. New Monte Carlo algorithms for classical spin systems

35.1.1. *Periodic boundary conditions for "grids".* Consider d-dim. "grid": $\begin{matrix} i_1 = 0, 1, \dots L_1 - 1 \\ i_2 = 0, 1, \dots L_2 - 1 \\ \vdots \\ i_d = 0, 1, \dots L_d - 1 \end{matrix}$ e.g. $d = 3$, $\begin{matrix} i = 0, 1, \dots L_x - 1 \\ j = 0, 1, \dots L_y - 1 \\ k = 0, 1, \dots L_z - 1 \end{matrix}$

Consider for periodic boundary conditions

$$\begin{matrix} i_1 \in \mathbb{Z}_{L_1} \\ i_2 \in \mathbb{Z}_{L_2} \\ \vdots \\ i_d \in \mathbb{Z}_{L_d} \end{matrix}$$

e.g. $d = 3$, so $\begin{matrix} i \in \mathbb{Z}_{L_x} \\ j \in \mathbb{Z}_{L_y} \\ z \in \mathbb{Z}_{L_z} \end{matrix}$

Suppose the "stencil" is given by $\nu_1, \nu_2 \dots \nu_d$ where

$$\begin{matrix} \nu_1 \in \{-1, 1\} \\ \nu_2 \in \{-1, 1\} \\ \vdots \\ \nu_d \in \{-1, 1\} \end{matrix}$$

e.g. $\begin{matrix} \nu_x \in \{-1, 1\} \\ \nu_y \in \{-1, 1\} \\ \nu_z \in \{-1, 1\} \end{matrix}$

Consider the implementation given by Hjorth-Jensen (2015) [17] on pp. 436 for `ising_2dim.cpp` in Ch. 13 Monte Carlo Methods in Statistical Physics (which is unnecessary complicated) for the periodic boundary condition:

```
inline int periodic(int i, int limit, int add) {
return (i+ limit+add) % (limit);
}
```

This is because

$$(i_i + \nu_i + L_i) \equiv (i_i + \nu_i) \bmod L_i$$

since

$$\frac{(i_i + \nu_i + L_i) - (i_i + \nu_i)}{L_i} = 1 \in \mathbb{Z}$$

So $i_i + \nu_i + L_i \sim i_i + \nu_i$, modulo $L - i$.

Thus we seek to calculate for periodic boundary conditions to be implemented

$$[i_i + \nu_i] \in \mathbb{Z}_{L_i} \tag{71}$$

35.2. **Single-spin selection as the first step of the Metropolis algorithm, and the case for its thorough, whole modification for the parallel computation case.** It appears that the heart of the Metropolis algorithm is that we flip a *single* spin, i.e. *single-spin flip*.

Let random variable $X \in \{0, 1\}$, s.t. $X = 1$ if some single lattice site k is chosen, $X = 0$ if particular site k is not chosen.

Suppose we choose $p(X = 1) = \frac{1}{L}$ where $L =$ total number of sites (e.g. $L = L_x L_y$ for a $L_x \times L_y$ 2-dim. grid). Thus, we have a uniform distribution for the choice of spin to flip.

Now

$$E[X] = \frac{1}{L} = \text{ expectation value of } X$$

which can intuitively be interpreted as the average number of sites chosen at some particular site k.

Since site selection is completely random, i.e. for X_k, $k \in \Lambda$, a random variable for site k in lattice Λ, the X_k's are independent and identically distributed, $\sum_{k=0}^{L-1} X_k$ represents the total number of sites chosen on lattice Λ.

$$E\left[\sum_{k=0}^{L-1} X_k\right] = \sum_{k=0}^{L-1} E[X_k] = L\left(\frac{1}{L}\right) = 1$$

which can intuitively be interpreted that on average, 1 site will be selected on lattice Λ.

35.2.1. *Single-spin selection for serial Metropolis method, explicitly discussed.* It should be clarified that when it says to "pick a single spin from the lattice grid", the probability of choosing *a* (or *some*) spin in this step is *1*, i.e.

$$P(\text{some}\mathbf{k} \in \Lambda) = 1 = \text{ probability of choosing some spin site } \mathbf{k} \text{ in lattice grid } \Lambda$$

Indeed, since the probability of picking a *particular* site $\mathbf{k}$ from Λ is $\frac{1}{|\Lambda|}$, and since the picking of a particular site $\mathbf{k}$ is *independent* of picking a different particular site $\mathbf{k}'$, then

$$\sum_{\mathbf{k} \in \Lambda} P(\mathbf{k}) = \sum_{\mathbf{k} \in \Lambda} \frac{1}{|\Lambda|} = \Lambda \frac{1}{\Lambda} = 1$$

For the example of a 2-dim. $L_x \times L_y$ grid, the above expressions specialize to

$$P(\text{some}(i, j) \in \{0, 1, \dots L_x - 1\} \times \{0, 1, \dots L_y - 1\}) = 1$$

$$\sum_{(i,j) \in \{0,1,\dots L_x - 1\} \times \{0,1,\dots L_y - 1\}} P((i, j)) = \sum_{j=0}^{L_y - 1} \sum_{i=0}^{L_x - 1} P((i, j)) = \sum_{j=0}^{L_y - 1} \sum_{i=0}^{L_x - 1} \frac{1}{L_x L_y} = L_x L_y \frac{1}{L_x L_y} = 1$$

I extensively expounded on this point because it becomes important in the actual *implementation* of this step in serial code.

Let's say that we are already given a random number generator. Because $|\Lambda|$, the size of the lattice grid ($L_x L_y$ for the 2-dim. case) can be huge ($L_x L_y \sim 10^6$ for a 1024×1024 lattice grid), it doesn't make sense to call a function that'll randomly pick 1 integer from $\{0, 1, \dots |\Lambda| - 1\}$, but rather to employ a random number generator that'll return a float between $0, 1$, i.e. in $[0., 1.)$, multiply it by L_x, and take the integer "floor", and then take another float in $[0., 1.)$, multiply it by L_y, and take the integer "floor."

But notice that in this procedure, we end up always picking *a spin*. What we cannot do, for example in the 2-dim. lattice grid case, is to take 2 for loops, 1 for $j \in \{0, 1, \dots L_y - 1\}$, another for $i \in \{0, 1, \dots L_x - 1\}$, and then assign a probability $\frac{1}{L_x L_y}$ of whether spin at site (i, j) will be chosen. This is because there is the possibility that no spins will get chosen after completing these 2 for loops - we have then not assigned the correct probability for this event, nor implemented correctly that we pick *a spin* from Λ.

35.2.2. *Single-spin selection is not tenable for (CUDA) parallel code because you idle all threads but 1.* However, for CUDA, which is a SIMT (Single-Instruction Multiple Thread) architecture, selecting a single spin at 1 time is not in anyway a reasonable action. Clearly, all the other threads, but 1, will be idled. After a choice of a single site, only 1 thread would be doing the next computations in the Metropolis algorithm. Nothing parallel is occurring.

This deficiency with this first step of picking a single spin has not been addressed in previous papers on a parallel implementation of the Ising model. Rather, most of the papers, at this point, discuss linear congruential generators for random numbers,

which I here make outdated with the implementation of CURAND, CUDA's random number generator, available to each single thread.

35.2.3. *Why then was there a single spin selection as the first step of the Metropolis algorithm?* Then why was there a single-spin selection as a first step of the Metropolis algorithm? What was the rationale for it? Consider that the number of all possible spin configurations, $\sigma \equiv \{\sigma_{\mathbf{k}}\}_{\mathbf{k}\in\Lambda}$, given a lattice grid Λ, is $2^{|\Lambda|}$, where we have a factor of 2 (spin up or spin down) for each lattice site, i.e. each spin in Λ. Also, recall that the partition function $Z = Z(\beta) = \sum_\sigma \exp(-\beta H(\sigma))$, where $\beta = \frac{1}{T}$, i.e. it's the inverse of the system's temperature, and summation is over all possible spin configurations σ, a summation of $2^{|Lambda|}$ terms. Calculating directly Z was never practical and so another method was sought out.

Rather, observe that the probability of spin configuration σ is $P_\beta(\sigma) = \frac{\exp(-\beta H(\sigma))}{Z_\beta}$. Clearly, P, as a function of σ is highly peaked at a few, certain spin configurations, while being near 0 for (many) others. So we shouldn't search the entire spin configuration space of σ's but rather deduce the most probable spin configuration.

35.2.4. *"Completely random" spin-selection vs. "biased"; Parallel utilization vs. serial code.* So we have the strategy that we want to minimize the Hamiltonian, $H(\sigma)$, and so we'll calculate changes in energy ΔH with a single spin flip at a particular site, ahead of us as the following steps in a Metropolis algorithm, which will help us find this most probable spin configuration, given β.

But beforehand, for the first step, if we are, as an extreme example, choosing the exact same spin to flip, we are not searching at all the entire spin configuration space of σ's, our search is extremely "biased." On the other hand, if we pick a single site with uniform probability distribution assigned to each site in Λ, then our search is "completely random." Suppose this spectrum of schemes for picking a single site lie on an axis of "randomness", from "completely biased" to "completely random."

Also, consider another axis orthogonal to the axis, from being "biased" to "completely random." An essentially serial (CPU) implementation of the single spin selection process would involve a single thread that would perform mathematical operations for a single spin site. Beforehand, it would pick a single spin, and no other, and then sequentially follow the Metropolis algorithm. For instance, some implement a "sweep" of the lattice grid with 2 or more for loops to loop through each grid coordinate. But keep in mind that this for loop is completely sequential, in order of the given start value, incremented up to the end value, e.g. from 0 to $L_x - 1$ for coordinate i.

Thus, we are back to the problem that we are idling all thread, but 1, for a parallel implementation. We cannot naively parallelize those 2 for loops immediately above, in a "sweep" because we need to guarantee a single spin is selected for each thread. But with the selection of a single spin, suppose 2 threads choose the same spin site, and so 1 operation would have to proceed after the completion of the other, defeating parallelization.

The single spin selection step, the first step of the Metropolis algorithm, cannot be saved for the parallel implementation.

For the sake of argument, the "most parallel" scheme would be to choose all spin sites to flip or not flip. On an axis for the spectrum from "serial" to "parallel", this would be "parallel", opposite to the serial "sweep" method for single spin selection. Of course, choosing all the spins once would run into problems with the calculation of $\Delta H(\sigma)$ since it depends on the values of the spins of its nearest neighbor, at its present configuration. Then there are so-called "checkerboard decomposition" to update even sites first, and then odd sites next, to avoid calculating ΔH from adjacent spins that could have changed values already in a parallel update.

However, the concern with these schemes, where all sites are updated in parallel or in a checkerboard decomposition, is that, on the axis from "biased" to "completely random", the spin selection is not "completely random" and so we are "biased" in our search over all possible spin configurations σ.

I would argue that the subsequent Metropolis step where we use the transition probability $A(\mu,\nu) = \exp(-\beta(H_\nu - H_\mu))$ is all that's necessary to ensure that we are obtaining the spin configuration σ that is the most probable spin configuration.

Nevertheless, this discussion was made to make explicit that any parallel implementation of the Metropolis algorithm necessitates scrapping the first step of single-spin selection. Selecting which spins to choose, whether in a checkerboard or double checkerboard decomposition, at once in parallel is open for discussion.

Newman and Barkema (1999) [20]

Weigel (2011) used the double checkerboard Metropolis method Simulating spin models on GPU by Martin Weigel https://arxiv.org/pdf/1006.3865.pdf

Minicourse on Machine Learning for Many-Body Physics; September 25-29, 2017 in São Paulo, Brazil. ICTP-SAIFR/IFT-UNESP http://www.ictp-saifr.org/?page_id=15446

Part 9. Computational Fluid Dynamics (CFD); Computational Methods

36. On Computational Methods for Aerospace Engineering, via Darmofal, Spring 2005

Notes to follow along Darmofal (2005) [21]

36.1. **On Lecture 1, Numerical Integration of Ordinary Differential Equations.** For the 1-dim. case,

$$m_p \frac{du}{dt} = m_p g - D(u)$$

Recall the velocity vector field $u = u(t,x) \in \mathfrak{X}(\mathbb{R}\times\mathbb{R})$. This is *not* what we want in this case; we want for particles the tangent bundle.

$$D = D(u) = \frac{1}{2}\rho_g \pi a^2 u^2 C_D(\mathrm{Re})$$
$$\mathrm{Re} = \frac{2\rho_g u a}{\mu_g}$$
$$C_D = \frac{24}{\mathrm{Re}} + \frac{6}{1+\sqrt{\mathrm{Re}}} + 0.4$$

Darmofal (2005) [21] then made a brief aside/note on linearization.
Consider perturbation method (linearization)

$$u(t) = u_0 + \widetilde{u}(t)$$

e.g. constant (in time).
If $\frac{du}{dt} = f(u,t)$,

$$\frac{d\widetilde{u}}{dt} = f(u_0+\widetilde{u},t) = f(u_0,t) + \left.\frac{\partial f}{\partial u}\right|_{u_0,0}\widetilde{u} + \left.\frac{\partial f}{\partial t}\right|_{u_0,0} t + \mathcal{O}(t^2,\widetilde{u}t,\widetilde{u}^2)$$

$a = 0.01\,m$, $\rho_p = 917\,\mathrm{kg}/m^3$ $\quad\quad$ $\rho_g = 0.9\,\mathrm{kg}/m^3$
$m_p = \rho_p \frac{4}{3}\pi a^3 = 0.0038\,\mathrm{kg}$
$\mu_g = 1.69\times 10^{-5}\,\mathrm{kg}/(m\,\mathrm{sec})$
$g = 9.8\,m/s^2$
In the 3-dim. case,

$$m_p \frac{d\mathbf{u}}{dt} = m_p g - D(u)\frac{\mathbf{u}}{|u|}$$

Consider curve $\begin{matrix} x:\mathbb{R}\to N = \mathbb{R} \\ x(t)\in\mathbb{R}\end{matrix}$, $u(t) \equiv \frac{dx}{dt} \in \Gamma(TN) = \Gamma(T\mathbb{R})$

$$\frac{du}{dt} = g - \frac{D(u)}{m_p}$$

$$\frac{du}{dt} = (u(t+\Delta t) - u(t))\frac{1}{\Delta t} + \mathcal{O}(\Delta t)$$

36.2. **Multi-step methods generalized.** This subsection corresponds to Lecture 3: Convergence of Multi-Step Methods, but is a further generalization to the presented multi-step methods.

The problem to solve, the ODE to compute out, is

$$\frac{du}{dt}(t) = f(u(t), t) \tag{72}$$

Make the following ansatz:

$$\frac{du}{dt}(t) = \sum_{\nu=0}^{N} \frac{1}{h} C_\nu u(t - \nu h) = \sum_{\xi=1}^{P} \beta_\xi f(u(t - \xi h), t - \xi h) \tag{73}$$

Do the Taylor expansion:

$$\sum_{\nu=0}^{N} \frac{1}{h} C_\nu \left[u(t) + \left(\frac{du}{dt} \right)(t) \cdot (-\nu h) + \sum_{j=2}^{n} \frac{u^{(j)}(t)}{j!} (-\nu h)^j + \mathcal{O}(h^n) \right] = \sum_{\xi=1}^{P} \beta_\xi \frac{du}{dt}(t - \xi h) =$$
$$= \sum_{\xi=1}^{P} \beta_\xi \left[\frac{du}{dt} + \sum_{j=2}^{n} \frac{u^{(j)}(t)}{j!} (-\xi h)^j + \mathcal{O}(h^n) \right]$$

36.3. **Convection (Discretized).** While I am following Lecture 7 of Darmofal (2005) [21], I will generalize to a "foliated, spatial" (smooth) manifold N, parametrized by time $t \in \mathbb{R}$, $\mathbb{R} \times N$, with $\dim N = n = 1, 2$ or 3 and to *CUDA* C/C++ parallel programming.

Consider n-form $m \in \Omega^N(\mathbb{R} \times N)$, $\dim N = n$. Then

$$\frac{d}{dt} m = \frac{d}{dt} \int_V \rho \mathrm{vol}^n = \int_V \mathcal{L}_{\frac{\partial}{\partial t} + \mathbf{u}} \rho \mathrm{vol}^n = \int_V \frac{\partial \rho}{\partial t} \mathrm{vol}^n + \mathbf{d} i_{\mathbf{u}} \rho \mathrm{vol}^n = \int_V \left(\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho u) \right) \mathrm{vol}^n =$$
$$= \int_V \frac{\partial \rho}{\partial t} \mathrm{vol}^n + \int_{\partial V} \rho i_{\mathbf{u}} \mathrm{vol}^n = \dot{m} \tag{74}$$

where recall

$$\mathrm{div} : \mathfrak{X}(\mathbb{R} \times N) \to C^\infty(\mathbb{R} \times N)$$
$$\mathrm{div}(\rho \mathbf{u}) = \frac{1}{\sqrt{g}} \frac{\partial (\sqrt{g} u^i \rho)}{\partial x^i}$$

36.3.1. *1-dimensional case for Convection from mass (scalar) conservation.* Consider Cell i, between $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$, i.e. $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \subset \mathbb{R}$. In this case, Eq. 74, for mass conservation with sources, becomes

$$\int_V \frac{\partial \rho}{\partial t} \mathrm{vol}^n + \int_{\partial V} \rho i_{\mathbf{u}} \mathrm{vol}^n = \int_V \frac{\partial \rho}{\partial t} dx + \int_{\partial V} \rho u^i = \int_{x_L}^{x_R} \frac{\partial \rho}{\partial t} dx + (\rho(x_R) u(x_R) - \rho(x_L) u(x_L)) = \frac{d}{dt} \int_{x_L}^{x_R} \rho(x) dx$$

In the case of $\frac{d}{dt} m = 0$, on a single cell i,

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial \rho}{\partial t} dx + \rho(x) u(x)|_{x_i + \frac{1}{2}} - \rho(x) u(x)|_{x_i - \frac{1}{2}} = 0$$

This is one of the first main approximations Darmofal (2005) [21] makes, in Eq. 7.10, Section 7.3 Finite Volume Method for Convection, for the *finite volume method*:

$$\overline{m}_i := \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \rho(x) dx \tag{75}$$

where $\Delta x_i \equiv x_{i+\frac{1}{2}} - x_{i+\frac{1}{2}}$.

And so

$$\Delta x_i \frac{\partial}{\partial t} \overline{m}_i + \rho(x) u(x)|_{x_{i+\frac{1}{2}}} - \rho(x) u(x)|_{x_{i-\frac{1}{2}}} = 0 \tag{76}$$

We want to discretize this equation also in time.

Consider as first approximation,

$$\overline{m}(x, t) = \overline{m}_i(t) \qquad \forall\, x_{i-\frac{1}{2}} < x < x_{i+\frac{1}{2}} \tag{77}$$

Consider then initial time t, time step Δt.

36.3.2. *1-dimensional "Upwind" Interpolation for Finite Volume.* This is the "major" approximation for the so-called "Upwind" interpolation approximation:

$$\rho(x_{i+\frac{1}{2}}, t + \Delta t) = \begin{cases} \overline{m}_i(t) & \text{if } u(x_{i+\frac{1}{2}}, t) > 0 \\ \overline{m}_{i+1}(t) & \text{if } u(x_{i+\frac{1}{2}}, t) < 0 \end{cases} \tag{78}$$

Then use the so-called "forward" time approximation for $\frac{d}{dt} \overline{m}_i(t)$:

$$\Delta x_i \frac{\overline{m}_i(t + \Delta t) - \overline{m}_i(t)}{\Delta t} + (\rho u)(t, x_{i+\frac{1}{2}}) - (\rho u)(t, x_{i-\frac{1}{2}}) = 0$$

Darmofal (2005) [21] didn't make this explicit in Lecture 7, but in the approximation for $\rho(x_{i+\frac{1}{2}}, t + \Delta t)$, Eq. 78, it's supposed that it's valid at time t: $\rho(x_{i+\frac{1}{2}}, t) \approx \rho(x_{i+\frac{1}{2}}, t + \Delta t)$, since it's the value of ρ for time moving forward from t (this is implied in Darmofal's code convect1d

$$\rho(x_{i+\frac{1}{2}}, t) u(x_{i+\frac{1}{2}}, t) = \begin{cases} \overline{m}_i(t) u(x_{i+\frac{1}{2}}, t) & \text{if } u(x_{i+\frac{1}{2}}, t) > 0 \\ \overline{m}_{i+1}(t) u(x_{i+\frac{1}{2}}, t) & \text{if } u(x_{i+\frac{1}{2}}, t) < 0 \end{cases}$$

Then

$$\frac{\Delta x_i}{\Delta t} (\overline{m}_i(t + \Delta t) - \overline{m}_i(t)) +$$
$$+ \begin{cases} \overline{m}_i(t) u(x_{i+\frac{1}{2}}, t) & \text{if } u(x_{i+\frac{1}{2}}, t) > 0 \\ \overline{m}_{i+1}(t) u(x_{i+\frac{1}{2}}, t) & \text{if } u(x_{i+\frac{1}{2}}, t) < 0 \end{cases} -$$
$$- \begin{cases} \overline{m}_{i-1}(t) u(x_{i-\frac{1}{2}}, t) & \text{if } u(x_{i-\frac{1}{2}}, t) > 0 \\ \overline{m}_i(t) u(x_{i-\frac{1}{2}}, t) & \text{if } u(x_{i-\frac{1}{2}}, t) < 0 \end{cases} =$$
$$= 0 \tag{79}$$

A note on 1-dimensional gridding: Consider total length $L_0 \in \mathbb{R}^+$. For N^{cells} total cells in x-direction. $i = 0 \dots N^{\text{cells}} - 1$.

$$x_{i-\frac{1}{2}} = i \Delta x \qquad i = 0, 1 \dots N^{\text{cells}} - 1$$
$$x_{i+\frac{1}{2}} = (i + 1) \Delta x \qquad i = 0, 1 \dots N^{\text{cells}} - 1$$
$$x_i = x_{i-\frac{1}{2}} + \frac{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}}{2} = \frac{x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}}}{2} = (2i + 1) \frac{\Delta x}{2} \qquad i = 0, 1 \dots N^{\text{cells}} - 1$$

At this point, instead of what is essentially the so-called "Upwind Interpolation", which Darmofal is doing in Lecture 7 of Darmofal (2005) [21], and on pp. 76, Chapter 4 Finite Volume Methods, Subsection 4.4.1 Upwind Interpolation (UDS) of Ferziger and Peric (2002) [22], which is essentially a zero-order approximation, let's try to do better.

Consider the interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \subset \mathbb{R}$.

For the 1-dimensional case of (pure) convection,

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial \rho(t, x)}{\partial t} dx + \rho(t, x_{i+\frac{1}{2}}) u(t, x_{i+\frac{1}{2}}) - \rho(t, x_{i-\frac{1}{2}}) u(t, x_{i-\frac{1}{2}}) = \frac{d}{dt} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \rho(x) dx$$

Given $\rho(t,x_{i-\frac{1}{2}}),\rho(t,x_{i+\frac{1}{2}})\in\mathbb{R}$, do (polynomial) interpolation:

$$\mathbb{R}\times\mathbb{R}\xrightarrow{\text{interpolation}}\mathbb{R}[x]\equiv\mathcal{P}_{n=1}(\mathbb{R})$$
$$\rho(t,x_{i-\frac{1}{2}}),\rho(t,x_{i+\frac{1}{2}})\mapsto\frac{(x-x_{i-\frac{1}{2}})\rho(t,x_{i+\frac{1}{2}})-(x-x_{i+\frac{1}{2}})\rho(t,x_{i-\frac{1}{2}})}{h}=\rho_{n=1}(t,x)$$

where $h\equiv x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}$ and $\mathcal{P}_{n=1}(\mathbb{R})$ is the set of all polynomials of order $n=1$ over field $\mathbb{R}$ (real numbers).

In general,

$$\mathbb{R}\times\mathbb{R}\xrightarrow{\text{interpolation}}\mathbb{R}[x]\equiv\mathcal{P}_{n=1}(\mathbb{R})$$
$$\rho(t,x_L),\rho(t,x_R)\mapsto\frac{(x-x_L)\rho(t,x_R)-(x-x_R)\rho(t,x_L)}{(x_R-x_L)}=\rho_{n=1}(t,x)$$

We interchange the operations of integration and partial derivative - I (correct me if I'm wrong) give two possible reasons why we can do this: the spatial manifold N is fixed in time t, and if the grid cell itself is fixed in time, then the partial derivative in time can be moved out of the integration limits.

So, interchanging $\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}dx$ and $\frac{\partial}{\partial t}$:

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\frac{\partial\rho(t,x)}{\partial t}dx=\frac{\partial}{\partial t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\rho(t,x)dx$$

So then

$$\implies\frac{\partial}{\partial t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\rho_{n=1}(t,x)=\frac{\partial}{\partial t}(\rho(t,x_{i+\frac{1}{2}})+\rho(t,x_{i-\frac{1}{2}}))\frac{\Delta x}{2}$$

where $\Delta x=x_{i+\frac{1}{2}}-x_{i+\frac{1}{2}}$.

In general,

$$\frac{\partial}{\partial t}\int_{x_L}^{x_R}\rho_{n=1}(t,x)=\frac{\partial}{\partial t}(\rho(t,x_R)+\rho(t,x_L))\frac{(x_R-x_L)}{2}$$

Then, discretizing,

(80)
$$\implies\left[(\rho(t+\Delta t,x_{i+\frac{1}{2}})+\rho(t+\Delta t,x_{i-\frac{1}{2}}))-(\rho(t,x_{i+\frac{1}{2}})+\rho(t,x_{i-\frac{1}{2}}))\right]\frac{\Delta x}{2}\left(\frac{1}{\Delta t}\right)+\rho(t,x_{i+\frac{1}{2}})u(t,x_{i+\frac{1}{2}})-\rho(t,x_{i-\frac{1}{2}})u(t,x_{i-\frac{1}{2}})=$$
$$=\dot{m}_{[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]}(t)$$

To obtain $\rho(t,x_{i-\frac{1}{2}})$, consider

$$\frac{\partial\rho}{\partial t}+\text{div}(\rho u)=\frac{d\rho}{dt}=0$$

which is valid at every point on N.

Consider for $\dim N=1$,

$$\frac{\partial\rho}{\partial t}(t,x)+\frac{\partial(\rho u)}{\partial x}(t,x)$$

Now, we want $x=x_{i-\frac{1}{2}}$.

Consider

$$\frac{\partial\rho(t,x_{i-\frac{1}{2}})}{\partial t}\approx\frac{\rho(t+\Delta t,x_{i-\frac{1}{2}})-\rho(t,x_{i-\frac{1}{2}})}{\Delta t}$$

Next, consider the (polynomial) interpolation for the $\frac{\partial(\rho u)}{\partial x}(t,x)$ term:

$$\mathbb{R}\times\mathbb{R}\times\mathbb{R}\xrightarrow{\text{interpolate}}\mathbb{R}[x]\equiv\mathcal{P}_{n=2}(\mathbb{R})$$
$$\rho(t,x_{i-\frac{3}{2}})u(t,x_{i-\frac{3}{2}}),\rho(t,x_{i-\frac{1}{2}})u(t,x_{i-\frac{1}{2}}),\rho(t,x_{i+\frac{1}{2}})u(t,x_{i+\frac{1}{2}})\xmapsto{\text{interpolate}}(\rho u)_{n=2}(t,x)$$

Thus, we can calculate, by plugging into,

$$\frac{\partial(\rho u)_{n=2}(t,x_{i-\frac{1}{2}})}{\partial x}$$

In general, for

$$\mathbb{R}\times\mathbb{R}\times\mathbb{R}\xrightarrow{\text{interpolate}}\mathbb{R}[x]\equiv\mathcal{P}_{n=2}(\mathbb{R})$$
$$\rho(t,x_{LL})u(t,x_{LL}),\rho(t,x_L)u(t,x_L),\rho(t,x_R)u(t,x_R)\xrightarrow{\text{interpolate}}(\rho u)_{n=2}(t,x)$$

we have

$$\frac{\partial(\rho u)_{n=2}(t,x_L)}{\partial x}=\frac{1}{(x_L-x_{LL})(x_L-x_R)(x_{LL}-x_R)}\cdot$$
$$\cdot\Big((x_L-x_{LL})^2(\rho u)(x_R)+(x_L-x_{LL})(x_{LL}-x_R)(\rho u)(x_L)-(x_L-x_R)^2(\rho u)(x_{LL})+(x_L-x_R)(x_{LL}-x_R)(\rho u)(x_L)\Big)$$

Thus,

(81)
$$\implies\begin{array}{c}\rho(t+\Delta t,x_{i-\frac{1}{2}})-\rho(t,x_{i-\frac{1}{2}})+\frac{\partial(\rho u)_{n=2}}{\partial x}(t,x_{i-\frac{1}{2}})\Delta t=0\text{ or}\\ \rho(t+\Delta t,x_{i-\frac{1}{2}})=\rho(t,x_{i-\frac{1}{2}})-\frac{\partial(\rho u)_{n=2}}{\partial x}(t,x_{i-\frac{1}{2}})\Delta t\end{array}$$

Now a note on the 1-dimensional grid, "gridding": for cell $i=0,\dots N^{\text{cell}}-1$, N^{cell} cells total in the x-direction, then

$$x_{i-\frac{1}{2}}=ih$$
$$x_{i-\frac{1}{2}}=(i+1)h$$

and so $x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}=h$, meaning the cell width or cell size is h.

Thus, in summary,

$$\rho(t+\Delta t,x_{i-\frac{1}{2}})=\rho(t,x_{i-\frac{1}{2}})-(\rho(t,x_{i+\frac{1}{2}})u(t,x_{i+\frac{1}{2}})-\rho(t,x_{i-\frac{3}{2}})u(t,x_{i-\frac{3}{2}}))\left(\frac{1}{2h}\right)\Delta t$$

(82)
$$\left[(\rho(t+\Delta t,x_{i+\frac{1}{2}})+\rho(t+\Delta t,x_{i-\frac{1}{2}}))-(\rho(t,x_{i+\frac{1}{2}})+\rho(t,x_{i-\frac{1}{2}}))\right]\frac{h}{2}\left(\frac{1}{\Delta t}\right)+\rho(t,x_{i+\frac{1}{2}})u(t,x_{i+\frac{1}{2}})-\rho(t,x_{i-\frac{1}{2}})u(t,x_{i-\frac{1}{2}})=$$
$$=\dot{m}_{[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]}(t)$$

If one was to include Newtonian gravity, consider this general expression for the time derivative of the momentum flux Π:

(83)
$$\Pi=\int_{B(t)}\rho u^i\text{vol}^n\otimes e_i$$
$$\dot{\Pi}=\int_{B(t)}\frac{\partial(\rho u^i)}{\partial t}\text{vol}^n\otimes e_i+\int_{B(t)}d(\rho u^i i_u\text{vol}^n)\otimes e_i=\int_{B(t)}\frac{\partial(\rho u^i)}{\partial t}\text{vol}^n\otimes e_i+\int_{\partial B(t)}\rho u^i i_u\text{vol}^n\otimes e_i$$

In 1-dim.,

$$\dot{\Pi}=\int_{B(t)}\frac{\partial(\rho u)}{\partial t}dx+\int_{\partial B}\rho u^2=\int_B\frac{GMdm}{r^2}=GM\int_B\frac{\rho\text{vol}^n}{r^2}=GM\int_B\frac{\rho dx}{(R-x)^2}$$

Considering a first-order polynomial interpolation for ρ,$\rho_{n=1}$,

$$\frac{\partial}{\partial t}((\rho u)(t,x_{i+\frac{1}{2}})+(\rho u)(t,x_{i-\frac{1}{2}}))\frac{h}{2}+\rho u^2(t,x_{i+\frac{1}{2}})-\rho u^2(t,x_{i-\frac{1}{2}})=GM\int\frac{\rho_{n=2}dx}{(R-x)^2}$$

Note that we need another equation, at $x=x_{i-\frac{1}{2}}$, similar to above:

$$\frac{\partial(\rho u)}{\partial t}+\frac{\partial(\rho u^2)}{\partial x}=\frac{GM\rho}{(R-x)^2}$$
$$\implies\rho u(t+\Delta t,x_{i-\frac{1}{2}})-\rho u(t,x_{i-\frac{1}{2}})+(\rho u^2(t,x_{i+\frac{1}{2}})-\rho u^2(t,x_{i-\frac{3}{2}}))\left(\frac{1}{2h}\right)\Delta t=\Delta t\int GM\frac{\rho dx}{(R-x)^2}$$

As a recap, the 1-dimensional setup is as follows:

$$\mathbb{R}\times N=\mathbb{R}\times\mathbb{R}\xrightarrow{\text{discretization}}\mathbb{Z}\times\mathbb{Z}$$
$$(t,x)\xmapsto{\text{discretization}}(t_0+(\Delta t)j, x_{i-\frac{1}{2}}=ih),\qquad i,j\in\mathbb{Z}$$

Initial conditions for $\rho\in C^\infty(\mathbb{R}\times\mathbb{R})$: $\rho(t_0,x)\in C^\infty(\mathbb{R}\times\mathbb{R})$.

Choices for $u\in\mathfrak{X}(\mathbb{R}\times\mathbb{R})$:

- $u(t,x)=u(x)$ (i.e. time-independent velocity vector field)
- $u(t,x)$ determined by Newtonian gravity (that's an external force on the fluid)

36.3.3. *Note on 1-dimensional gridding.* For, $[0,1]\subset\mathbb{R}$
N cells,
Then $1/N=\Delta x$. Then consider

$$x_j=j\Delta x\qquad j=0,1,\dots N$$

36.4. **2-dim. and 3-dim. "Upwind" interpolation for Finite Volume.** I build on Lecture 7 of Darmofal (2005) [21].
Consider a rectangular grid.

Consider cell C^2_{ij}, $i=0\dots N_x-1$, $j=0\dots N_y-1$. Then there's $N_x\cdot N_y$ total cells, N_x cells in x-direction .
N_y cells in y-direction

There are 2 possibilities: rectangles of all the same size, with width l^x and length l^t each, or each rectangle for each cell C^2_{ij} is different, of dimensions $l^x_i\times l^y_j$.

Consider cells centered at $x_{2i+1}=l^x\frac{(2i+1)}{2}=\sum_{k=0}^{i-1}l^x_k+\frac{l^2_i}{2}$.
On the "left" sides, $x_{2i}=l^xi=\sum_{k=0}^{i-1}l^x_k$
"right" sides, $x_{2(i+1)}=l^x(i+1)=\sum_{k=0}^{i}l^x_k$.
So cells are centered at

$$(x_{2i+1},y_{2j+1})=(l^x\frac{(2i+1)}{2},l^y\frac{(2j+1)}{2})=\left(\sum_{k=0}^{i-1}l^x_k+\frac{l^x_i}{2},\sum_{k=0}^{j-1}l^y_k+\frac{l^y_j}{2}\right)$$

So this cell C^2_{ij}, a 2-(cubic) simplex has 4 1-(cubic) simplices (edges): so 1-(cubic) simplices $\{C^1_{i\pm1,j},C^1_{i,j\pm1}\}$
The center of these simplices are the following:

$$x_{C^1_{i+1,j}}=(x_{2i+1+1},y_{2j+1})=(l^x(i+1),l^y\frac{(2j+1)}{2})=\left(\sum_{k=0}^{i}l^x_k,\sum_{k=0}^{j-1}l^y_k+\frac{l^y_j}{2}\right)\text{ so then}$$
$$x_{C^1_{i\pm1,j}}=(x_{2i+1\pm1},y_{2j+1})=(l^x\left(\frac{2i+1\pm1}{2}\right),l^y\frac{(2j+1)}{2})=\left(\sum_{k=0}^{\frac{2i-1\pm1}{2}}l^x_k,\sum_{k=0}^{j-1}l^y_k+\frac{l^y_j}{2}\right)$$
$$x_{C^1_{i,j\pm1}}=(x_{2i+1},y_{2j+1\pm1})=(l^x\left(\frac{2i+1}{2}\right),l^y\frac{(2j+1\pm1)}{2})=\left(\sum_{k=0}^{i-1}l^x_k+\frac{l^x_i}{2},\sum_{k=0}^{\frac{2j-1\pm1}{2}}l^y_k\right)$$

We want the flux. So for

$$\overline{\rho}_{ij}:=\frac{1}{l^x_il^y_j}\int_{C^2_{ij}}\rho\text{vol}^2$$

then the flux through 1-(cubic) simplices (faces), $\int\rho i_{\mathbf{u}}\text{vol}^2$,

$$\int_{C^1_{i+1,j}}\rho i_{\mathbf{u}}\text{vol}^2=\begin{cases}l^y_j\overline{\rho}_{ij}u^x(x_{C^1_{i+1,j}}) & \text{if } u^x(x_{C^1_{i+1,j}})>0\\ l^y_j\overline{\rho}_{i+1,j}u^x(x_{C^1_{i+1,j}}) & \text{if } u^x(x_{C^1_{i+1,j}})<0\end{cases}$$
$$\int_{C^1_{i-1,j}}\rho i_{\mathbf{u}}\text{vol}^2=\begin{cases}-l^y_j\overline{\rho}_{i-1,j}u^x(x_{C^1_{i-1,j}}) & \text{if } u^x(x_{C^1_{i-1,j}})>0\\ -l^y_j\overline{\rho}_{i,j}u^x(x_{C^1_{i-1,j}}) & \text{if } u^x(x_{C^1_{i-1,j}})<0\end{cases}$$

Likewise,

$$\int_{C^1_{i,j+1}}\rho i_{\mathbf{u}}\text{vol}^2=\begin{cases}l^x_i\overline{\rho}_{ij}u^y(x_{C^1_{i,j+1}}) & \text{if } u^y(x_{C^1_{i,j+1}})>0\\ l^x_i\overline{\rho}_{i,j+1}u^y(x_{C^1_{i,j+1}}) & \text{if } u^y(x_{C^1_{i,j+1}})<0\end{cases}$$

and so on.

36.4.1. *3-dim. "Upwind" interpolation for finite volume.* For a rectangular prism (cubic),
for cell C^3_{ijk}, $i=0\dots N_x-1$, $j=0\dots N_y-1$, $k=0\dots N_z-1$, $N_x\cdot N_y\cdot N_z$ total cells.
Cells centered at

$$(x_{2i+1},y_{2j+1},z_{2j+1})=(l^x\frac{(2i+1)}{2},l^y\frac{(2j+1)}{2},l^z\frac{(2k+1)}{2})=\left(\sum_{l=0}^{i-1}l^x_l+\frac{l^x_i}{2},\sum_{l=0}^{j-1}l^y_l+\frac{l^y_j}{2},\sum_{l=0}^{k-1}l^z_l+\frac{l^y_k}{2}\right)$$

For the 3-(cubic) simplex, C^3_{ijk}, it has 6 2-(cubic) simplices (faces). So for C^3_{ijk}, consider $\{C^2_{i\pm1,jk},C^2_{ij\pm1,k},C^2_{ijk\pm1}\}$.
The center of these faces, such as for $C^2_{i\pm1,jk}$, $x_{C^2_{i\pm1,jk}}$, for instance,

$$x_{C^2_{i\pm1,jk}}=(x_{2i+1\pm1},y_{2j+1},z_{2k+1})=(l^x\left(\frac{2i+1\pm1}{2}\right),l^y\frac{(2j+1)}{2},l^z\frac{(2j+1)}{2})=\left(\sum_{l=0}^{\frac{2i-1\pm1}{2}}l^x_l,\sum_{l=0}^{j-1}l^y_l+\frac{l^y_j}{2},\sum_{l=0}^{l-1}l^z_l+\frac{l^z_k}{2}\right)$$

We want the flux. So for

$$\overline{\rho}_{ijk}:=\frac{1}{l^x_il^y_jl^z_k}\int_{C^3_{ijk}}\rho\text{vol}^3$$

then the flux through 2-(cubic) simplices (faces), $\int\rho i_{\mathbf{u}}\text{vol}^3$,

$$\int_{C^2_{i+1,jk}}\rho i_{\mathbf{u}}\text{vol}^3=\begin{cases}l^y_jl^z_k\overline{\rho}_{ijk}u^x(x_{C^2_{i+1,jk}}) & \text{if } u^x(x_{C^2_{i+1,jk}})>0\\ l^y_jl^z_k\overline{\rho}_{i+1,jk}u^x(x_{C^2_{i+1,jk}}) & \text{if } u^x(x_{C^2_{i+1,jk}})<0\end{cases}$$

and so on.

To reiterate the so-called "upwind" interpolation method, in generality, recall that we are taking this equation:

$$\int_{C^n_{ij}}\frac{\partial\rho}{\partial t}\text{vol}^n+\int_{\partial C^n_{ij}}\rho u_{\mathbf{u}}\text{vol}^n=\dot{M}_{ij}$$

and discretizing it to obtain

$$\frac{\partial}{\partial t}\overline{\rho}_{ij}|\text{vol}^n|+\int_{\partial C^n_{ij}}\rho i_{\mathbf{u}}\text{vol}^n=\dot{M}_{ij}$$
$$\implies\frac{\partial}{\partial t}\overline{\rho}_{ij}=\frac{-1}{|\text{vol}^n|}\int_{\partial C^n_{ij}}\rho i_{\mathbf{u}}\text{vol}^n+\frac{1}{|\text{vol}^n|}\dot{M}_{ij}$$

37. Finite Difference

References/Links that I used:

- Chapter 6 The finite difference method, by Pascal Frey
- Numerical Methods for Partial Differential Equations by Volker John
- *Wikipedia* "Finite Difference". Wikipedia has a section on Difference operators which appears powerful and general, but I haven't understood how to apply it. In fact, see my jupyter notebook on the `CompPhys` github, `finitediff.ipynb` on how to calculate the coefficients in arbitrary (differential) order, and (error) order (of precision, error, i.e. $O(h^p)$) for finite differences, approximations of derivatives.

From `finitediff.ipynb`, I derived this formula

$$f'(x) = \frac{1}{h}\sum_{\nu=1}^{3} C_\nu \cdot (f(x+\nu h) - f(x-\nu h)) + \mathcal{O}(h^7) \text{ for}$$

$$C_1 = \frac{3}{4} \tag{84}$$
$$C_2 = \frac{-3}{20}$$
$$C_3 = \frac{1}{60}$$

37.1. Finite Difference with Shared Memory (CUDA C/C++).

References/Links that I used:

- Finite Difference Methods in CUDA C++, Part 2, by Dr. Mark Harris
- GPU Computing with CUDA Lecture 3 - Efficient Shared Memory Use, Christopher Cooper of Boston University, August, 2011. UTFSM, Valparaíso, Chile.

cf. Finite Difference Methods in CUDA C++, Part 2, by Dr. Mark Harris

In x-derivative, $\frac{\partial f}{\partial x}$, $\forall$ thread block, $(j_x, j_y) \in \{\{0 \dots N_x - 1\} \times \{0 \dots N_y - 1\}$. $m_x \times s_{\text{Pencils}}$ elements $\in$ tile e.g. $64 \times$ sPencils.

In y-derivative, $\frac{\partial f}{\partial y}$, (x, y)-tile of sPencils $\times\, 64 = s_{\text{Pencils}} \times m_y$.

Likewise, $\frac{\partial f}{\partial z} \to (x, z)-$tile of sPencils $\times\, 64 = s_{\text{Pencils}} \times m_z$.

Consider for the y derivative, the code for `__global__ void derivative_y(*f, *d_f)` (finitediff.cu):

`int i` $\Longleftarrow i = j_x M_x + i_x \in \{0 \dots N_x M_x - 1\}$ (since $j_x M_x + i_x = (M_x - 1) + (N_x - 1)M_x$, i.e. the "maximal" case)

`int j` $\Longleftarrow j = i_y \in \{0 \dots M_y - 1\}$

`int k` $\Longleftarrow k = j_y \in \{0 \dots N_y - 1\}$

`int si` $\left\{\Longleftarrow si = i_x \in \{0 \dots M_x - 1\}\right\}$

`int sj` $\Longleftarrow sj = j + 4 \in \{4 \dots M_y + 3\}$. Notice that $r = 4$. Then generalize to $s_j = j + r \in \{r, \dots, M_y + r - 1\}$

`int globalIdx` $\Longleftarrow l = k m_x m_y + j m_x + i \in \{0, \dots, (N_y - 1)m_x m_y + (M_y - 1)m_x + N_x M_x - 1\}$ since

$$(N_y - 1)m_x m_y + (M_y - 1)m_X + N_x M_x - 1$$

`s_f[sj][si]` $\Longleftarrow s_f[s_j][s_i] \equiv (s_f)_{s_j, s_i} = f(l)$, $l \in \{0 \dots (N_y - 1)m_x m_y + (M_y - 1)m_x + N_x M_x - 1\}$ with

$$s_f \in \text{Mat}_{\mathbb{R}}(M_y + r, M_x)$$

If $j < 4$, $j < r$, $j = s_j - r \in \{0 \dots r - 1\}$ and so

$$(s_f)_{(s_j - r), s_i} = (s_f)_{s_j + m_y - 1 - r, s_i}$$
$$\Longleftrightarrow$$
$$\left\{\{0, \dots r - 1\} \times \{0 \dots M_x - 1\} \leftarrow \{m_y - 1, \dots M_y + m_y - 2\} \times \{0 \dots M_x - 1\}\right\}$$

Then, the actual approximation method:

$$\frac{\partial f}{\partial y}(l) = \frac{\partial f}{\partial y}(i, j, k) = \sum_{\nu=1}^{r} c_\nu ((s_f)_{s_j + \nu, s_i} - (s_f)_{s_j - \nu, s_i})$$

The shared memory tile here is

`__shared__ float s_f[m_y+8][sPencils]` $\Longleftarrow s_f \in \text{Mat}_{\mathbb{R}}(m_y + 2r, s_{\text{Pencil}})$

By using the shared memory tile, each element from global memory is read only once. (cf. Finite Difference Methods in CUDA C++, Part 2, by Dr. Mark Harris)

Consider expanding the number of pencils in the shared memory tile, e.g. 32 pencils.

Harris says that "with a 1-to-1 mapping of threads to elements where the derivative is calculated, a thread block of 2048 threads would be required." Consider then letting each thread calculate the derivative for multiple points.

So Harris uses a thread block of $32 \times 8 \times 1 = 256$ threads per block, and have each thread calculate the derivative at 8 points, as opposed to a thread block of $4 * 64 * 1 = 256$ thread block, with each thread calculate the derivative at only 1 point.

Perfect coalescing is then regained.

GPU Computing with CUDA Lecture 3 - Efficient Shared Memory Use, Christopher Cooper

37.2. **Note on finite-difference methods on the shared memory of the device GPU, in particular, the pencil method, that attempts to improve upon the double loading of boundary "halo" cells (of the grid).** cf. Finite Difference Methods in CUDA C++, Part 1, by Dr. Mark Harris

Take a look at the code finite_difference.cu. The full code is there. In particular, consider how it launches blocks and threads in the kernel function (and call) `__global__ derivative_x`, `derivative_y`, `derivative_z`. `setDerivativeParameters` has the arrays containing `dim3` "instantiations" that have the grid and block dimensions, for x-,y-,z-derivatives and for "small" and "long pencils". Consider "small pencils" for now. The relevant code is as follows:

```
grid[0][0]  = dim3(my / sPencils, mz, 1);
block[0][0] = dim3(mx, sPencils, 1);

grid[0][1]  = dim3(my / lPencils, mz, 1);
block[0][1] = dim3(mx, sPencils, 1);

grid[1][0]  = dim3(mx / sPencils, mz, 1);
block[1][0] = dim3(sPencils, my, 1);

grid[1][1]  = dim3(mx / lPencils, mz, 1);
// we want to use the same number of threads as above,
// so when we use lPencils instead of sPencils in one
// dimension, we multiply the other by sPencils/lPencils
block[1][1] = dim3(lPencils, my * sPencils / lPencils, 1);

grid[2][0]  = dim3(mx / sPencils, my, 1);
block[2][0] = dim3(sPencils, mz, 1);

grid[2][1]  = dim3(mx / lPencils, my, 1);
block[2][1] = dim3(lPencils, mz * sPencils / lPencils, 1);
```

Let $N_i \equiv$ total number of cells in the grid in the ith direction, $i = x, y, z$. N_i corresponds to `m*` in the code, e.g. N_x is `mx`.

Note that in this code, what seems to be attempted is calculating the derivatives of a 3-dimensional grid, but using only 2-dimensions on the memory of the device GPU. In my experience, with the NVIDIA GeForce GTX 980 Ti, the maximum number of threads per block in the z-direction and the maximum number of blocks that can be launched in the z-direction is severely limited compared to the x and y directions (use `cudaGetDeviceProperties`, or run the code queryb.cu; I find

(85)

```
 Max threads per block: 1024
Max thread dimensions: (1024, 1024, 64)
Max grid dimensions:   (2147483647, 65535, 65535)
```

).

Let M_i be the number of threads on a block in the ith direction. Let N_i^{threads} be the total number of threads in the ith direction on the grid, i.e. the number of threads in the ith grid-direction. $i = x, y$. This is *not* the desired grid dimension N_i.

Surely, for a desired grid of size $N_x \times N_y \times N_z \equiv N_x N_y N_z$, then a total of $N_x N_y N_z$ threads are to be computed.

Denote $s_{\text{pencil}} \in \mathbb{Z}^+$ to be `sPencils`; example value is $s_{\text{pencil}} = 4$. Likewise, denote $l_{\text{pencil}} \in \mathbb{Z}^+$ to be `lPencils`; example value is $l_{\text{pencil}} = 32 > s_{\text{pencil}} = 4$.

Then, for instance the small pencil case, for the x-derivative, we have

$$\begin{aligned} &\text{grid dimensions } (N_y/s_{\text{pencil}}, N_z, 1) \\ &\text{block dimensions } (N_x, s_{\text{pencil}}, 1) \end{aligned} \tag{86}$$

Then the total number of threads launched in each direction, x and y, is

$$\begin{aligned} N_x^{\text{threads}} &= \frac{N_y}{s_{\text{pencil}}} N_x \\ N_y^{\text{threads}} &= \frac{N_z}{s_{\text{pencil}}} \end{aligned}$$

While it is true that the total number of threads computed matches our desired grid:

$$N_x^{\text{threads}} \cdot N_y^{\text{threads}} = \frac{N_y}{s_{\text{pencil}}} N_x N_z s_{\text{pencil}} = N_x N_y N_z$$

take a look at the block dimensions that were demanded in Eq. **??**, $(N_x, s_{\text{pencil}}, 1)$. The total number of threads to be launched in this block is $N_x \cdot s_{\text{pencil}}$. Suppose $N_x = 1920$. Then easily $N_x \cdot s_{\text{pencil}} >$ allowed maximum number of threads per block. In my case, this number is 1024.

Likewise for the case of x-direction, but with long pencils. The blocks and threads to be launched on the grid and blocks for the kernel function (`derivative_x`) is

$$\begin{aligned} &\text{grid dimensions } (N_y/l_{\text{pencil}}, N_z, 1) \\ &\text{block dimensions } (N_x, s_{\text{pencil}}, 1) \end{aligned} \tag{87}$$

The total number of threads to be launched in each block is also $N_x \cdot s_{\text{pencil}}$ and for large N_x, this could easily exceed the maximum number of threads per block allowed.

Also, be aware that the shared memory declaration is

```
__shared__ float s_f[sPencils][mx+8]
```

N_x (i.e. `mx`) can be large and we're requiring a 2-dim. array of size $(N_x + 8) * s_{\text{pencil}}$ of floats, for each block. As, from Code listing 85, much more blocks can be launched than threads on a block, and so trying to launch more blocks could possibly be a better solution.

38. Mapping scalar (data) to colors; Data Visualization

Links I found useful:

Taku Komura has good lectures on visualization with computers; it was heavily based on using VTK, but I found the principles and overview he gave to be helpful: here's Lecture 6 Scalar Algorithms: Colour Mapping. (Komura's teaching in the UK, hence spelling "colour")

Good article on practical implementation of a rainbow: https://www.particleincell.com/2014/colormap/, i.e. Converting Scalars to RGB Colormap.

I will formulate the problem mathematically (and clearly).

What we want is this, a bijection:

$$\mathbb{R} \longrightarrow [0,255]^3 \equiv RGB$$

$$[0,1) \longrightarrow RGB$$

$$U \subset \mathbb{R} \longrightarrow [0,1) \longrightarrow RGB \tag{88}$$

$$x \in U \longmapsto f = \tfrac{x-\text{minval}}{\text{maxval - minval}} \longmapsto (r,g,b)$$

with

$$\begin{aligned} \text{minval} &:= \min_{x\in U} x \\ \text{maxval} &:= \max x \end{aligned} \text{ and}$$

$$r,g,b \in [0,\dots,255] \subset \mathbb{Z}$$

Let n = number of "mapping segments" or "segments" (matplotlib terminology). e.g. $n = 5$.

Consider $\frac{1}{n}$, e.g. $\frac{1}{n} = \frac{1}{5} = 0.20$.

Let

$$y := \lfloor 255(nf - \lfloor nf \rfloor) \rfloor \in [0,255] \tag{89}$$

Then

$$(r,g,b) = \begin{cases} (255,y,0) & \text{if } \lfloor nf \rfloor = 0 \text{ or } 0 \le nf < 1 \\ (255-y,255,0) & \text{if } \lfloor nf \rfloor = 1 \text{ or } 1 \le nf < 2 \\ (0,255,y) & \text{if } \lfloor nf \rfloor = 2 \text{ or } 2 \le nf < 3 \\ (0,255-y,255) & \text{if } \lfloor nf \rfloor = 3 \text{ or } 3 \le nf < 4 \\ (y,0,255) & \text{if } \lfloor nf \rfloor = 4 \text{ or } 4 \le nf < 5 \\ (255,0,255) & \text{if } \lfloor nf \rfloor = 5 \end{cases} \tag{90}$$

To understand how this color mapping is implemented in `matplotlib`, take a look at

`class matplotlib.colors.LinearSegmentedColormap(name, segmentdata, N=256, gamma=1.0)` of colors - Matplotlib 1.5.3 documentation ,

and look at

row i: x y0 y1

/

/

row i+1: x y0 y1

http://scipy.github.io/old-wiki/pages/Cookbook/Matplotlib/Show_colormaps and http://stackoverflow.com/questions/16834861/create-own-colormap-using-matplotlib-and-plot-color-scale, for more examples of creating color maps, color bars.

39. On Griebel, Dornseifer, and Neunhoeffer's *Numerical Simulation in Fluid Dynamics: A Practical Introduction*

Griebel, Dornseifer, and Neunhoeffer (1997) [23]

See also Software of Research group of Prof. Dr. M. Griebel, Institute für Numerische Simulation http://wissrech.ins.uni-bonn.de/research/software/

39.1. **Boundary conditions.** cf. Sec. 2.1. The Mathematical Mode: The Navier-Stokes Equations, Ch. 2 The Mathematical Description of Flows, Griebel, Dornseifer, and Neunhoeffer (1997) [23], pp. 12-13

Let

$\varphi_n \equiv$ component of velocity orthogonal to boundary (in exterior normal direction)

$\varphi_t \equiv$ component of velocity parallel to boundary (in tangential direction)

derivatives in normal direction:

$$\frac{\partial \varphi_n}{\partial n}$$
$$\frac{\partial \varphi_t}{\partial n}$$

Let fixed boundary $\Gamma := \partial\Omega$. Consider

(1) *No-slip condition.*
No fluid penetrates boundary.
fluid is at rest there; i.e.

$$\begin{aligned} \varphi_n(x,y) &= 0 \\ \varphi_t(x,y) &= 0 \end{aligned} \tag{91}$$

(2) *Free-slip condition.*
No fluid penetrates boundary.
Contrary to no-slip condition,
there's no frictional losses at boundary, i.e.

$$\begin{aligned} \varphi_n(x,y) &= 0 \\ \frac{\partial \varphi_t}{\partial n}(x,y) &= 0 \end{aligned} \tag{92}$$

Free-slip condition often imposed along line or plane of symmetry in problem, thereby reducing size of domain, where flow needs to be comprised by half.

39.2. **Specific problems and related boundary conditions, boundary specifications.**

39.2.1. *Plate an an angle to the inflow.* For

$$\begin{aligned} y_0 &:= a_0 L_y \\ y_1 &:= a_1 L_y \end{aligned}$$

for $0 \le a_0 < a_1 \le 1$, e.g. $a_0 = \frac{2}{5}$, $a_1 = \frac{3}{5}$.

Consider an inclined plate to be an obstacle

$$\sum_{i=y_0+1}^{y_1-1} \sum_{j=i-1}^{i+1} \texttt{FLAG}_{ij}$$

Consider $y = Ax + b$, $y, x, A, b \in \mathbb{R}$, and consider

$$\begin{aligned} y &\ge Ax + b_0 \\ y &\le Ax + b_1 \end{aligned}$$

and so for the 2 points "at the bottom edge of this inclined plate",

$$(y_0+1, y_0+1-1) = (y_0+1, y_0)$$
$$(y_1-1, y_1-1-1) = (y_1-1, y_1-2)$$
$$y_0 = A(y_0+1)+b_0$$
$$y_1-2 = A(y_1-1)+b_0$$
$$y_1-2-y_0 = A(y_1-y_0)-2A$$
$$A = \frac{y_1-y_0-2}{y_1-y_0-2} = 1$$

$b_0 = -1$.

So $y \geq x-1$.

Likewise $y \leq x+1$ (plug in values).

So in parallel, consider $\forall\, i_x = \{y_0+1, y_0+2, \dots y_1-1\}$, access memory values at $j = i_x-1, i_x, i_x+1$.

Even though the "striding", or stride, for accessing $j = i_x-1, i_x, i_x+1$ is (L_x+2), each of the threads for $\forall\, i_x = \{y_0+1, y_0+2, \dots y_1-1\}$ will actually be concurrent.

39.3. **Shared memory tiling scheme applied to the staggered grid; i.e. shared memory tiling scheme for only the "inner cells", excluding halo of radius 1 boundary "cells".** I will first review the shared memory tiling scheme over the entire grid of absolute size $Lx \times Ly$.

For

$$k_x := i_x + j_x M_x \in \{0, 1 \dots N_x * M_x - 1\}$$
$$k_y := i_y + j_y M_y \in \{0, 1 \dots N_y * N_y - 1\}$$

Let

$$S_x := M_x + 2r \in \mathbb{Z}^+$$
$$S_y := M_y + 2r \in \mathbb{Z}^+$$

Then

$$\forall\, i \in \{i = i_x, i_x+M_x, \dots | i_x \leq i < S_x\}$$
$$\forall\, j \in \{j = i_y, i_y+M_y \dots | i_y \leq j < S_y\},$$

$$l_x := i-r+M_x j_x \in \{-r, -r+1, \cdots -r+M_x-1\} + M_x j_x$$
$$l_y := j-r+M_y j_y \in \{-r, -r+1, \cdots -r+M_y-1\} + M_y j_y$$

and then load input data into shared memory:

$$s_{\text{in}}[i+jS_x] := f^{(c)}(l_x + l_y L_x) \text{ with } \begin{matrix} 0 \leq i < S_x & 0 \leq l_x < L_x \\ 0 \leq j < S_y & 0 \leq l_y < L_y \end{matrix} \tag{93}$$

If $k_x \geq L_x$ or $k_y \geq L_y$, then end or return (check condition).

The actual stencil calculation is performed as follows:

$\forall\, \nu_y = 0, 1, \dots W-1$

$k_y^{\text{st}} := s_y + \nu_y - r$

$\forall\, \nu_x := 0, 1, \dots W-1$

$k_x^{\text{st}} := s_x + \nu_x - r$

$$g^{(c)}(k) = \sum_{\nu_y=0}^{W-1} \sum_{\nu_x=0}^{W-1} c_{\nu=\nu_x+W\nu_y} s_{\text{in}}[k_x^{\text{st}} + k_y^{\text{st}} S_x] \tag{94}$$

with $c_{\nu=\nu_x+W\nu_y} \equiv c(\nu_x, \nu_y)$ and $k := k_x + L_x * k_y$.

Now, we want to deal with applying the shared memory tiling scheme on a staggered grid, so threads aren't launched on the entire (staggered) grid, but only on inner cells.

Examine each step of the loading input data into shared memory procedure above: for $i_x \equiv$ `threadIdx.x`, $i_y \equiv$ `threadIdx.y`,

$$i_x = 0, 1 \dots M_x - 1$$
$$i_y = 0, 1 \dots M_y - 1$$

and for the `for` loops,

$$i \in \{i_x, i_x+M_x, \dots | i_x \leq i < S_x := M_x + 2r\}$$
$$i \in \{0, M_x\}, \{1, M_x+1\}, \dots \{2r-1, M_x+2r-1\}, \{2r\}, \{2r+1\}, \dots \{M_x-1\},$$

e.g. for $r = 1$,

$$i \in \{0, M_x\}, \{1, M_x+1\}, \{2\}, \{3\}, \dots \{M_x-1\},$$

and similarly,

$$j \in \{i_y, i_y+M_y, \dots | i_y \leq j < S_y := M_y + 2r\}$$
$$j \in \{0, M_y\}, \{1, M_y+1\}, \dots \{2r-1, M_y+2r-1\}, \{2r\}, \{2r+1\}, \dots \{M_y-1\},$$

Then, l_x, l_y are defined:

$$l_x := i-r+M_x j_x \in$$
$$\{-r+M_x j_x, -r+M_x(j_x+1)\}, \{1-r+M_x j_x, 1-r+M_x(j_x+1)\}, \dots \{r-1+M_x j_x, r-1+M_x(j_x+1)\} \dots$$
$$\dots \{r+M_x j_x\}, \{r+1+M_x j_x\}, \dots \{-1-r+M_x(j_x+1)\} \tag{95}$$

$$l_y := j-r+M_y j_y \in$$
$$\{-r+M_y j_y, -r+M_y(j_y+1)\}, \{1-r+M_y j_y, 1-r+M_y(j_y+1)\}, \dots \{r-1+M_y j_y, r-1+M_y(j_y+1)\} \dots$$
$$\dots \{r+M_y j_y\}, \{r+1+M_y j_y\}, \dots \{-1-r+M_y(j_y+1)\} \tag{96}$$

Now consider staggered grid:

$$k_x \in \{0, 1 \dots N_x M_x - 1\} \text{ but where } L_x < N_x M_x \leq L_x + M_x - 1$$
$$k_y \in \{0, 1 \dots N_y M_y - 1\} \text{ but where } L_y < N_y M_y \leq L_y + M_y - 1$$

but the "absolute" indices we want to loop over, for $\|verb\| \in \mathbb{Z}^{(l_x+2)\times(L_y+2)}$ are only `FLAG`'s "inner cells".

$$i := 1, 2 \dots L_x$$
$$j := 1, 2 \dots L_y$$

Note the 1-to-1 correspondence:

$$i = k_x + 1$$
$$j = k_y + 1$$

To me, at least, I'm bewildered with how to proceed, as this is a difficult problem, so I will begin to check corner cases and easy, simple cases.

Let $r = 1$.

Let $j_x = j_y = 0$.

Then for

$$i \in \{0, M_x\}, \{1, M_x+1\}, \{2\}, \{3\}, \dots \{M_x-1\}$$
$$j \in \{0, M_y\}, \{1, M_y+1\}, \{2\}, \{3\}, \dots \{M_y-1\}$$

and so consider

$$l_x = i+1-r+M_x j_x = i+M_x j_x$$
$$l_y = j+1-r+M_y j_y = j+M_y j_y$$

Clearly,for $j_x = j_y = 0$, the correct cells are loaded from input $f(l_x + (L_x+2)l_y) \xleftarrow{\text{flatten}} f(l_x, l_y)$ (note the "special", i.e. particular "striding" or stride of L_x+2 for a staggered grid).

For $J_x = 1$, $j_y = 0$,

$$l_x = i + M_x \in \{M_x, 2M_x\}, \{M_x + 1, 2M_x + 1\}, \{2 + M_x\}, \{3 + M_x\}, \dots \{2M_x - 1\}$$

which accesses $l_x = M_x$, which lies within the "radius" for "halo" cells of the "next" thread block over. By induction, we set

$$l_x := i + M_x j_x$$
$$l_y := j + M_y j_y$$

which correctly loads the input array.

Next, we have to do the actual stencil computation.

For "filterwidth" $W = 2r + 1$, consider $\nu_x = 0, 1 \dots W - 1$. For $r = 1$, $\nu_x, \nu_y = 0, 1, 2$

$$\nu_y = 0, 1 \dots W - 1$$

We want to consider when $\nu_x, \nu_y = \{0, 2\} = \{0, W - 1\}$, $\forall\, s_x = i_x + r$, $s_y = i_y + r$. So consider

$$s_{\text{in}}(s_x + 0 - r, s_y + 1 - r), \qquad s_{\text{in}}(s_x + 2 - r, s_y + 1 - r)$$
$$s_{\text{in}}(s_x + 1 - r, s_y + 0 - r), \qquad s_{\text{in}}(s_x + 1 - r, s_y + 2 - r)$$

and corresponding flags for

```
B_W  B_O
B_S  B_N
```

respectively (Ost is east in Deutsche, German, hence the "O").

cf. 2.2.2. Conservation of Momentum, Ch. 2 The Mathematical Description of Flows, Griebel, Dornseifer, and Neunhoeffer (1997) [23], pp. 16

For incompressible fluids,

$$\rho(\mathbf{x}, t) = \rho_\infty = \text{ const}$$

$$\frac{\partial \mathbf{u}}{\partial t} + u^j \frac{\partial \mathbf{u}}{\partial x^j} + \frac{1}{\rho_\infty} \text{grad} p = \frac{\mu}{\rho_\infty} \Delta \mathbf{u} + \mathbf{g} \tag{97}$$

with

dynamic viscosity μ

kinematic viscosity $\nu = \dfrac{\mu}{\rho_\infty}$

39.3.1. *Dynamic Similarity of Flows.* cf. 2.3 Dynamic Similarity of Flows, Ch. 2 The Mathematical Description of Flows, Griebel, Dornseifer, and Neunhoeffer (1997) [23], pp. 17

For incompressible flows,

$$\begin{aligned} (x^i)^* &:= \frac{x^i}{L} \\ t^* &:= \frac{u_\infty t}{L} \\ (u^i)^* &:= \frac{u^i}{u_\infty} \\ p^* &:= \frac{p - p_\infty}{\rho_\infty u_\infty^2} \end{aligned} \tag{98}$$

$$\frac{\partial \mathbf{u}^*}{\partial t^*} \left(\frac{u_\infty}{L/u_\infty} \right) + \frac{u_\infty^2}{L} (u^j)^* \frac{\partial \mathbf{u}^*}{\partial (x^j)^*} + \frac{1}{\rho_\infty} \frac{\rho_\infty u_\infty^2}{L} \text{grad}^* p^* = \frac{\mu}{\rho_\infty} \frac{1}{L^2} u_\infty \Delta^* (\mathbf{u})^* + \mathbf{g} \Longrightarrow$$

$$\frac{\partial \mathbf{u}^*}{\partial t^*} + (u^j)^* \frac{\partial \mathbf{u}^*}{\partial (x^j)^*} + \text{grad}^* p^* = \frac{\mu}{\rho_\infty u_\infty L} \Delta^* \mathbf{u}^* + \frac{L}{u_\infty^2} \mathbf{g} \tag{99}$$

$$\begin{aligned} \text{Re} &:= \frac{\rho_\infty u_\infty L}{\mu} && \text{(Reynolds number)} \\ \text{Fr} &:= \frac{u_\infty}{\sqrt{L \|\mathbf{g}\|}} && \text{(Froude number)} \end{aligned} \tag{100}$$

Compare this Eq. 99, derived with dynamic similarity, to Eq. (2.2a) of Griebel, Dornseifer, and Neunhoeffer (1997) [23], and my version of the dimensionless momentum conservation equation:

$$\frac{\partial \mathbf{u}}{\partial t} + u^j \frac{\partial \mathbf{u}}{\partial x^j} + \text{grad} p = \frac{1}{\text{Re}} \Delta \mathbf{u} + \mathbf{g}^* \qquad \text{(incompressible)} \tag{101}$$

However, Griebel, Dornseifer, and Neunhoeffer (1997) [23] writes it as

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial u^j \mathbf{u}}{\partial x^j} + \text{grad} p &= \frac{1}{\text{Re}} \Delta \mathbf{u} + \mathbf{g}^* \\ \text{div}(\mathbf{u}) &= 0 \end{aligned} \tag{102}$$

as, separately, a *momentum equation* and *continuity equation*. Griebel, Dornseifer, and Neunhoeffer (1997) [23] ends up implementing this version and so I needed to keep in mind the continuity equation condition.

39.4. **Discretization of the Navier-Stokes Equations.** cf. 3.1.2 Discretization of the Navier-Stokes Equations, Ch. 3 The Numerical Treatment of the Navier-Stokes Equations, Griebel, Dornseifer, and Neunhoeffer (1997) [23], pp. 26

39.4.1. *Gridding (revisited); staggered grid.* Consider again C_{ij}^2 a 2-(cubic) simplex.

$$\begin{matrix} i = 1, 2, \dots L_x \\ j = 1, 2, \dots L_y \end{matrix} \xrightarrow{\text{Griebel, et.al's notation}} \begin{matrix} i = 1, 2, \dots i_{\max} \\ j = 1, 2, \dots j_{\max} \end{matrix}$$

Assume rectangles of all same size, width δx, length δy.

To clarify (or drive home the point), cell (i, j), C_{ij}^2 occupies

$$[(i-1)\delta x, i\delta x] \times [(j-1)\delta y, j\delta y] \qquad \begin{matrix} \forall\, i = 1, 2, \dots L_x \\ j = 1, 2, \dots L_y \end{matrix}$$

cell centers:

$$x_{i-0.5, j-0.5} \equiv (x_{i-0.5, j-0.5}) = ((i - 0.5)\delta x, (j - 0.5)\delta y)$$

4 1-(cubic) simplices (edges)

$$C^1_{i, j-0.5} = \{i\delta x\} \times [(j - 0.5)\delta y, j\delta y] \qquad C^1_{i-0.5, j} = [(i - 0.5)\delta x, i\delta x] \times \{j\delta y\}$$
$$C^1_{i-1, j-0.5} = \{(i-1)\delta x\} \times [(j - 0.5)\delta y, j\delta y] \qquad C^1_{i-0.5, j-1} = [(i - 0.5)\delta x, i\delta x] \times \{(j-1)\delta y\}$$

(i, j) assigned to pressure at cell center, u^x velocity at right edge, u^y-velocity at upper edge of cell, i.e.

$$p : (i, j) \mapsto p_{i,j} \mapsto ((i - 0.5)\delta x, (j - 0.5)\delta y) \qquad \forall (i, j) \in \{1, 2, \dots L_x\} \times \{1, 2, \dots L_y\},\ \text{so } (i, j) \mapsto x_{ij} \in \Omega$$
$$u^x : (i, j) \mapsto u^x_{i,j} \mapsto (i\delta x, (j - 0.5)\delta y) \qquad \forall (i, j) \in \{1, 2, \dots L_x\} \times \{1, 2, \dots L_y\},\ \text{so } (i, j) \mapsto x_{ij} \in \Omega$$
$$u^y : (i, j) \mapsto u^y_{i,j} \mapsto ((i - 0.5)\delta x, j\delta y) \qquad \forall (i, j) \in \{1, 2, \dots L_x\} \times \{1, 2, \dots L_y\},\ \text{so } (i, j) \mapsto x_{ij} \in \Omega$$

where domain Ω cell (i, j), $i = 1 \dots L_x$. $[(i-1)\delta x, i\delta x] \times [(j-1)\delta y, j\delta y]$.
$j = 1 \dots L_y$

To reiterate,

$$\begin{matrix} & & (i - 0.5, j) & & \\ (i-1, j-0.5) & & (i - 0.5, j - 0.5) & & (i, j - 0.5) \\ & & (i - 0.5, j - 1) & & \end{matrix} \xleftarrow{\text{discretization}} ((i-1)\delta x, (j - 0.5)\delta y) \qquad \begin{matrix} & ((i - 0.5)\delta x, j\delta y) & \\ ((i - 0.5)\delta x, (j - 0.5)\delta y) & & (i\delta x, (j - 0.5)\delta y) \\ & ((i - 0.5)\delta x, (j-1)\delta y) & \end{matrix}$$

39.4.2. *Discretization of the Spatial Derivatives; Treatment of the Spatial Derivatives.* cf. 3.1.1 Simple Discretization Formulas, Ch. 3 The Numerical Treatment of the Navier-Stokes Equations, Griebel, Dornseifer, and Neunhoeffer (1997) [23].

$$\begin{aligned}
&\text{forward difference} && \left[\frac{du}{dx}\right]^r_i := \frac{u(x_{i+1}) - u(x_i)}{\delta x}\\
&\text{backward difference} && \left[\frac{du}{dx}\right]^l_i := \frac{u(x_i) - u(x_{i-1})}{\delta x}\\
&\text{central difference} && \left[\frac{du}{dx}\right]^c_i := \frac{u(x_{i+1}) - u(x_{i-1})}{2\delta x}
\end{aligned}$$

39.4.3. *Stability problems, unphysical oscillations.* Upwind difference or upwinding

$$\left[\frac{du}{dx}\right]^{\text{up}}_i := \frac{(1+\epsilon)(u_i - u_{i-1}) + (1-\epsilon)(u_{i+1} - u_i)}{2\delta x} \text{ with } \epsilon := \text{sign}(k) \tag{103}$$

cf. Eq. (3.9) Griebel, Dornseifer, and Neunhoeffer (1997) [23]

where k is in $\frac{d^2u}{dx^2} + k\frac{du}{dx} = f$ in Ω

Consider the weighted average of both.

$$\begin{gathered}
\gamma \in [0,1]\\
\gamma \cdot \text{ upwind difference } + (1-\gamma) \cdot \text{ central difference}\\
\gamma \cdot \left[\frac{du}{dx}\right]^{\text{up}}_i + (1-\gamma)\left[\frac{du}{dx}\right]^c_i
\end{gathered}$$

donor-cell scheme. Consider $\frac{d(ku)}{dx}$, $k \in C^\infty(\mathbb{R})$.

e.g. $\mathbb{R} \xrightarrow{\text{discretization}} \mathbb{Z}$. Consider $x_i \equiv i\delta x$, $i \in \mathbb{Z}$.

Consider $k_l := k_{i-0.5}$

$k_r := k_{i+0.5}$

So then

$$\left[\frac{d(ku)}{dx}\right]^{dc}_i := \frac{k_r u_r - k_l u_l}{\delta x} \tag{104}$$

cf. Eq. (3.11) Griebel, Dornseifer, and Neunhoeffer (1997) [23]

And so defining

$$\begin{aligned}
u_r &:= \begin{cases} u_i & k_r > 0 \\ u_{i+1} & k_r < 0 \end{cases}\\
u_l &:= \begin{cases} u_{i-1} & k_l > 0 \\ u_i & k_l < 0 \end{cases}
\end{aligned} \tag{105}$$

can be rewritten as

$$\begin{aligned}
\left[\frac{d(ku)}{dx}\right]^{dc}_i &= \frac{1}{2\delta x}((k_r - |k_r|)u_{i+1} + (k_r + |k_r| - k_l + |k_l|)u_i + (-k_l - |k_l|)u_{i-1}) =\\
&= \frac{1}{2\delta x}(k_r(u_i + u_{i+1}) - k_l(u_{i-1} + u_i) + |k_r|(u_i - u_{i+1}) - |k_l|(u_{i-1} - u_i))
\end{aligned} \tag{106}$$

cf. Eq. (3.12), pp. 25, Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Consider terms such as

$$\frac{\partial(u^j \mathbf{u})}{\partial x^j}$$

Take the average:

$$\begin{aligned}
\left[\frac{\partial(u^x u^y)}{\partial y}\right]_{i,j} &:= \frac{1}{\delta y}\left(\frac{(u^y_{i,j} + u^y_{i+1,j})}{2}\frac{(u^x_{i,j} + u^x_{i,j+1})}{2} - \frac{(u^y_{i,j-1} + u^y_{i+1,j-1})}{2}\frac{(u^x_{i,j-1} + u^x_{i,j})}{2}\right)\\
\left[\frac{\partial((u^x)^2)}{\partial x}\right]_{i,j} &:= \frac{1}{\delta x}\left(\left(\frac{u_{i,j} + u_{i+1,j}}{2}\right)^2 - \left(\frac{u_{i-1,j} + u_{i,j}}{2}\right)^2\right)
\end{aligned} \tag{107}$$

To understand these formulas, **take a look at Fig. 3.6. on pp. 28** of Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Continuity equation.

$$\text{div}\mathbf{u} = 0 \tag{108}$$

cf. Eq. (2.2c)

$$\text{div}\mathbf{u} = 0 \xrightarrow{discretization} \begin{aligned}\left[\frac{\partial u}{\partial x}\right]_{i,j} &:= \frac{u_{i,j} - u_{i-1,j}}{\delta x}\\ \left[\frac{\partial u^y}{\partial y}\right]_{i,j} &:= \frac{u^y_{i,j} - u^y_{i,j-1}}{\delta y}\end{aligned}$$

Also note this result that'll be used for the Poisson equation describing pressure p:

$$\boxed{\text{divgrad}p = \Delta p = \text{div}\left(-\frac{\partial(u^j\mathbf{u})}{\partial x^j} + \frac{1}{\text{Re}}\Delta\mathbf{u} + \mathbf{g}^*\right)} \tag{109}$$

$$\xrightarrow{\text{div}\mathbf{u}} \text{div}\left(\frac{\partial \mathbf{u}}{\partial t}\right) = \frac{\partial}{\partial t}\text{div}\mathbf{u} = 0$$

was used.

39.4.4. *F,G terms (which include central difference, c, and donor-cell, dc, methods).* Page 29 of cf. 3.1.2 Discretization of the Navier-Stokes Equations, Ch. 3 The Numerical Treatment of the Navier-Stokes Equations, Griebel, Dornseifer, and Neunhoeffer (1997) [23] is *gold* for understanding the implementation Griebel, Dornseifer, and Neunhoeffer had used and how to implement the computation or calculationg of F, G.

$\forall$ cell (i,j), $i = 1 \dots i_{\max} - 1$, $j = 1 \dots j_{\max}$, for u at the midpoint of the *right edge* of the cell, then from Eq. (3.19)a,

$$\begin{aligned}
\left[\frac{\partial(u^2)}{\partial x}\right]_{i,j} &:= \frac{1}{\delta x}\left(\left(\frac{u_{ij} + u_{i+1j}}{2}\right)^2 - \left(\frac{u_{i-1j} + u_{ij}}{2}\right)^2\right) +\\
&+\gamma\frac{1}{\delta x}\left(\frac{|u_{ij} + u_{i+1j}|}{2}\frac{(u_{ij} - u_{i+1j})}{2} - \frac{|u_{i-1j} + u_{ij}|}{2}\frac{(u_{i-1j} - u_{ij})}{2}\right)
\end{aligned} \tag{110}$$

Indeed, in this case

$$\begin{aligned}
k_r &:= \frac{u_{ij} + u_{i+1j}}{2}\\
k_l &:= \frac{u_{i-1j} + u_{ij}}{2}
\end{aligned}$$

and Eq. 106 tells us (cf. Eq. 3.12 of Griebel, Dornseifer, and Neunhoeffer (1997) [23]), which again is

$$\left[\frac{d(ku)}{dx}\right]^{dc}_i = \frac{1}{2\delta x}(k_r(u_i + u_{i+1}) - k_l(u_{i-1} + u_i) + |k_r|(u_i - u_{i+1}) - |k_l|(u_{i-1} - u_i))$$

and so in this case

$$\left[\frac{\partial(u^2)}{\partial x}\right]^{dc}_{ij} = \frac{1}{\delta x}\left[\left(\frac{u_{ij}+u_{i+1j}}{2}\right)\left(\frac{u_{ij}+u_{i+1j}}{2}\right) - \left(\frac{u_{i-1j}+u_{ij}}{2}\right)\left(\frac{u_{i-1j}+u_{ij}}{2}\right) + \frac{|u_{ij}+u_{i+1j}|}{2}\frac{(u_{ij}-u_{i+1j})}{2} + \frac{|u_{i-1j}+u_{ij}|}{2}\frac{(u_{i+1j}-u_{ij})}{2}\right] =$$
$$= \left[\frac{\partial(u^2)}{\partial x^2}\right]^{c}_{ij} + \frac{1}{\delta x}\left(\frac{|u_{ij}+u_{i+1j}|}{2}\frac{(u_{ij}-u_{i+1j})}{2} + \frac{|u_{i-1j}+u_{ij}|}{2}\frac{(u_{i+1j}-u_{ij})}{2}\right)$$

Consider Fig. 3.6 on pp. 28 of Griebel, Dornseifer, and Neunhoeffer (1997) [23], "Values required for the discretization of the u-momentum equation.

$$\begin{array}{ccccc} & & u_{ij+1} & & \\ & v_{ij} & & v_{i+1j} & \\ u_{i-1j} & & u_{ij} & & u_{i+1j} \\ & v_{ij-1} & & v_{i+1j-1} & \\ & & u_{ij-1} & & \end{array} \quad \Longrightarrow \quad \begin{array}{ccc} & u^x_{ij+1} & \\ u^x_{i-1j} & \mathbf{u}_{ij} & \mathbf{u}_{i+1j} \\ & \mathbf{u}_{ij-1} & u^y_{i+1j-1} \end{array}$$

$$\begin{array}{ccccc} & & v_{ij+1} & & \\ & u_{i-1j+1} & & u_{ij+1} & \\ v_{i-1j} & & v_{ij} & & v_{i+1j} \\ & u_{i-1j} & & u_{ij} & \\ & & v_{ij-1} & & \end{array} \quad \Longrightarrow \quad \begin{array}{ccc} u^x_{i-1j+1} & \mathbf{u}_{ij+1} & \\ \mathbf{u}_{i-1j} & \mathbf{u}_{ij} & u^y_{i+1j} \\ & u^y_{ij-1} & \end{array}$$

$$F := (u^x)^n + \delta t\left[\frac{1}{\mathrm{Re}}\Delta u^x - \frac{\partial(u^j u^x)}{\partial x^j} + (g^*)^x\right] \tag{111}$$
$$G := (u^y)^n + \delta t\left[\frac{1}{\mathrm{Re}}\Delta u^y - \frac{\partial(u^j u^y)}{\partial x^j} + (g^*)^y\right]$$

So. *Compute $u^{(n+1)}$, $v^{(n+1)}$ according to (3.34), (3.35).* cf. pp. 34, 3.2.2 The Discrete Momentum Equations, Griebel, Dornseifer, and Neunhoeffer (1997) [23].

$$u^{(n+1)}_{i,j} = F^{(n)}_{i,j} - \frac{\delta t}{\delta x}(p^{(n+1)}_{i+1,j} - p^{(n+1)}_{i,j}) \qquad \begin{array}{l} i = 1 \ldots i_{\max} - 1 \\ j = 1 \ldots j_{\max} \end{array} \tag{112}$$

cf. (3.34) Griebel, Dornseifer, and Neunhoeffer (1997) [23]

$$v^{(n+1)}_{i,j} = G^{(n)}_{i,j} - \frac{\delta t}{\delta y}(p^{(n+1)}_{i,j+1} - p^{(n+1)}_{i,j}) \qquad \begin{array}{l} i = 1 \ldots i_{\max} \\ j = 1 \ldots j_{\max} - 1 \end{array} \tag{113}$$

cf. (3.34) Griebel, Dornseifer, and Neunhoeffer (1997) [23]
with F, G, cf. (3.29), and with

F discretized at right edge of cell (i, j)
G discretized at upper edge of cell (i, j)

which is

$$F_{i,j} := u_{i,j} + \delta t\left(\frac{1}{\mathrm{Re}}\left(\left[\frac{\partial^2 u}{\partial x^2}\right]_{i,j} + \left[\frac{\partial^2 v}{\partial y^2}\right]_{i,j}\right) - \left[\frac{\partial(u^2)}{\partial x}\right]_{i,j} - \left[\frac{\partial(uv)}{\partial y}\right]_{i,j} + g_x\right) \qquad \begin{array}{l} i = 1 \ldots i_{\max} - 1 \\ j = 1 \ldots j_{\max} \end{array} \tag{114}$$

cf. Eq. (3.36) Griebel, Dornseifer, and Neunhoeffer (1997) [23], and

$$G_{i,j} := v_{i,j} + \delta t\left(\frac{1}{\mathrm{Re}}\left(\left[\frac{\partial^2 v}{\partial x^2}\right]_{i,j} + \left[\frac{\partial^2 v}{\partial y^2}\right]_{i,j}\right) - \left[\frac{\partial(uv)}{\partial x}\right]_{i,j} - \left[\frac{\partial(v^2)}{\partial y}\right]_{i,j} + g_y\right) \qquad \begin{array}{l} i = 1 \ldots i_{\max} \\ j = 1 \ldots j_{\max} - 1 \end{array} \tag{115}$$

cf. Eq. (3.37) Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Consider where the set of cells C_{ij} intersect or overlap that are needed for $u^{n+1}_{i,j}$ and $v^{(n+1)}_{i,j}$.

$$\{(i,j)|^{i = 1 \ldots i_{\max} - 1}_{j = 1 \ldots j_{\max} - 1}\} \Longleftrightarrow u^{(n+1)}_{i,j} = ((u^x)^{(n+1)}_{i,j}, (u^y)^{(n+1)}_{i,j})$$

For Eq. (3.36), (3.37) of Griebel, Dornseifer, and Neunhoeffer (1997) [23], which are quoted for the implementation or algorithm,

$$F_{i,j} := u_{i,j} + \delta t\left(\frac{1}{\mathrm{Re}}\left(\left[\frac{\partial^2 u}{\partial x^2}\right]_{i,j} + \left[\frac{\partial^2 v}{\partial y^2}\right]_{i,j}\right) - \left[\frac{\partial(u^2)}{\partial x}\right]_{i,j} - \left[\frac{\partial(uv)}{\partial y}\right]_{i,j} + g_x\right) \qquad \begin{array}{l} i = 1 \ldots i_{\max} - 1 \\ j = 1 \ldots j_{\max} \end{array} \tag{116}$$

cf. Eq. (3.36) Griebel, Dornseifer, and Neunhoeffer (1997) [23], and

$$G_{i,j} := v_{i,j} + \delta t\left(\frac{1}{\mathrm{Re}}\left(\left[\frac{\partial^2 v}{\partial x^2}\right]_{i,j} + \left[\frac{\partial^2 v}{\partial y^2}\right]_{i,j}\right) - \left[\frac{\partial(uv)}{\partial x}\right]_{i,j} - \left[\frac{\partial(v^2)}{\partial y}\right]_{i,j} + g_y\right) \qquad \begin{array}{l} i = 1 \ldots i_{\max} \\ j = 1 \ldots j_{\max} - 1 \end{array} \tag{117}$$

cf. (3.37) Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Eq. (3.38) of Griebel, Dornseifer, and Neunhoeffer (1997) [23] is essentially the discretization of the Poisson equation (that takes advantage of the fact that we're dealing with the incompressible case):

$$\Delta p = \frac{1}{\delta t}\mathrm{div}((F,G)) \xrightarrow{\text{discretization}} \tag{118}$$

$$\frac{p^{(n+1)}_{i+1,j} - 2p^{(n+1)}_{i,j} + p^{(n+1)}_{i-1,j}}{(\delta x)^2} + \frac{p^{(n+1)}_{i,j+1} - 2p^{(n+1)}_{i,j} + p^{(n+1)}_{i,j-1}}{(\delta y)^2} = \frac{1}{\delta t}\left(\frac{F^{(n)}_{i,j} - F^{(n)}_{i-1,j}}{\delta x} + \frac{G^{(n)}_{i,j} - G^{(n)}_{i,j-1}}{\delta y}\right) \qquad \begin{array}{l} i = 1 \ldots i_{\max} \\ j = 1 \ldots j_{\max} \end{array} \tag{119}$$

cf. Eq. (3.38) of Griebel, Dornseifer, and Neunhoeffer (1997) [23], pp. 35, 3.2.3. The Poisson Equation for the Pressure.

39.4.5. *Poisson equation for pressure (for this incompressible case), and towards SOR (successive Over Relaxation) method.*

$$\frac{\epsilon^E_i(p^{(n+1)}_{i+1,j} - p^{(n+1)}_{i,j}) - \epsilon^W_i(p^{(n+1)}_{i,j} - p^{(n+1)}_{i-1,j})}{(\delta x)^2} + \frac{\epsilon^N_j(p^{(n+1)}_{i,j+1} - p^{(n+1)}_{i,j}) - \epsilon^S_j(p^{(n+1)}_{i,j} - p^{(n+1)}_{i,j-1})}{(\delta y)^2} =$$
$$= \frac{1}{\delta t}\left(\frac{F^{(n)}_{i,j} - F^{(n)}_{i-1,j}}{\delta x} + \frac{G^{(n)}_{i,j} - G^{(n)}_{i,j-1}}{\delta y}\right) \tag{120}$$
$$i = 1 \ldots i_{\max}$$
$$j = 1 \ldots j_{\max}$$

where

$$\epsilon^W_i := \begin{cases} 0 & i = 1 \\ 1 & i > 1 \end{cases}$$
$$\epsilon^E_i := \begin{cases} 1 & i < i_{\max} \\ 0 & i = i_{\max} \end{cases}$$
$$\epsilon^S_j := \begin{cases} 0 & j = 1 \\ 1 & j > 1 \end{cases}$$
$$\epsilon^N_j := \begin{cases} 1 & j < j_{\max} \\ 0 & j = j_{\max} \end{cases}$$

cf. Eq. (3.43) Griebel, Dornseifer, and Neunhoeffer (1997) [23].

e.g., for $1 < i < i_{\max}$, and $1 < j < j_{\max}$,

$$\frac{(p^{(n+1)}_{i+1,j} - p^{(n+1)}_{i,j}) - (p^{(n+1)}_{i,j} - p^{(n+1)}_{i-1,j})}{(\delta x)^2} + \frac{(p^{(n+1)}_{i,j+1} - p^{(n+1)}_{i,j}) - (p^{(n+1)}_{i,j} - p^{(n+1)}_{i,j-1})}{(\delta y)^2} = \\ = \frac{1}{\delta t}\left(\frac{F^{(n)}_{i,j} - F^{(n)}_{i-1,j}}{\delta x} + \frac{G^{(n)}_{i,j} - G^{(n)}_{i,j-1}}{\delta y}\right)$$

Indeed it takes into account the discretized boundary conditions, which was obtained from the so-called *Chorin projection method* (which is based on the Hodge-Helmholtz decomposition).

$$\begin{aligned} p_{0,j} &= p_{i,j}; \qquad & p_{imax+1,j} &= p_{imax,j}, \qquad & j &= 1 \dots j_{\max} \\ p_{i,0} &= p_{i,1}; \qquad & p_{i,jmax+1} &= p_{i,jmax}, \qquad & i &= 1 \dots i_{\max} \end{aligned} \tag{121}$$

cf. Eq. (3.41) Griebel, Dornseifer, and Neunhoeffer (1997) [23].

39.4.6. *Successive Over Relaxation (SOR) method.* $\forall\, it = 1 \dots it_{\max}$

$i = 1 \dots i_{\max}$,

$j = 1 \dots j_{\max}$

$$p^{it+1}_{i,j} := (1-\omega)p^{it}_{i,j} + \\ + \frac{\omega}{\left(\frac{\epsilon^E_i + \epsilon^W_i}{(\delta x)^2} + \frac{\epsilon^N_j + \epsilon^S_j}{(\delta y)^2}\right)} \cdot \left(\frac{\epsilon^E_i p^{it}_{i+1,j} + \epsilon^W_i p^{it+1}_{i-1,j}}{(\delta x)^2} + \frac{\epsilon^N_j p^{it}_{i,j+1} + \epsilon^S_j p^{it+1}_{i,j-1}}{(\delta y)^2} - \mathrm{rhs}_{ij}\right) \tag{122}$$

cf. Eq. (3.44) Griebel, Dornseifer, and Neunhoeffer (1997) [23].

39.4.7. *Implementation, routines, algorithms for incompressible Navier-Stokes equations solver.* Consider the serial version of the incompressible Navier-Stokes equations solver with finite difference:

- $t := 0$, $n := 0$
- initial values of u, v, p i.e. $\mathbf{u}, p$
- While $t < t_{\text{end}}$
  - select δt (according to (3.50) if stepsize control is used)
  - set boundary values for u, v
  - Compute $F^{(n)}$ and $G^{(n)}$ according to (3.36), (3.37)
  - Compute RHS of pressure Eq. (3.38)
  - Set $it := 0$
  - While $it < it_{\max}$ and $\|r^{it}\| > \text{eps}$ (resp., $\|r^{it}\| > \text{eps}\ \|p^0\|$)
    - Perform an SOR cycle according to (3.44)
    - Compute the residual norm for the pressure equation, $\|r^{it}\|$
    - $it := it + 1$
  - Compute $u^{(n+1)}$ and $v^{(n+1)}$ according to (3.34), (3.35)
  - $t := t + \delta t$
  - $n := n + 1$

Algorithm 1. Base version (or serial version), pp. 40, of Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Compare this with the parallelized version:

- $t := 0$, $n := 0$
- initial values of u, v, p i.e. $\mathbf{u}, p$
- While $t < t_{\text{end}}$
  - select δt (according to (3.50) if stepsize control is used)
  - set boundary values for u, v
  - Compute $F^{(n)}$ and $G^{(n)}$ according to (3.36), (3.37)
  - Compute RHS of pressure Eq. (3.38)
  - Set $it := 0$
  - While $it < it_{\max}$ and $\|r^{it}\| > \text{eps}$ (resp., $\|r^{it}\| > \text{eps}\ \|p^0\|$)
    - Perform an SOR cycle according to (3.44)
    - Exchange the pressure values in the boundary strips
    - Compute the partial residual and send these to the master process
    - *Master process* computes the residual norm of the pressure equation $\|r^{it}\|$ and broadcasts it to all processes
    - $it := it + 1$
  - Compute (update) $u^{(n+1)}$ and $v^{(n+1)}$ according to (3.34), (3.35)
  - $t := t + \delta t$
  - $n := n + 1$

Algorithm 3. Parallel version, pp. 115 of Griebel, Dornseifer, and Neunhoeffer (1997) [23]

40. Solving Poisson equation's by the preconditioned conjugate gradient method

The Poisson equation comes into play when determining (solving for) the pressure p in the incompressible Navier-Stokes equations for fluid flow. Let us review how this comes about.

Recall the full incompressible Navier-Stokes equations, which comes from momentum conservation, and mass conservation, respectively:

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial u^j \mathbf{u}}{\partial x^j} + \mathrm{grad} p &= \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^* \\ \mathrm{div}\mathbf{u} &= 0 \end{aligned} \tag{123}$$

Using the incompressibility condition

$$\mathrm{div}\mathbf{u} = 0$$

then exchanging the partial derivative over time with divergence div (we should be able to do this if there is a foliation over time $t \in \mathbb{R}$, i.e. "time-slices" of (spatial) Riemannian manifold N):

$$\mathrm{div}\frac{\partial \mathbf{u}}{\partial t} = \frac{\partial}{\partial t}\mathrm{div}\mathbf{u} = 0 \tag{124}$$

and so applying the divergence div to Eq. 123, we obtain

$$\mathrm{div}\mathrm{grad} p = \Delta p = \mathrm{div}\left(\frac{-\partial(u^j \mathbf{u})}{\partial x^j} + \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^*\right) \tag{125}$$

Now, considering the right-hand side (RHS) of Eq. 125, we effectively add 0 via the incompressible condition $\mathrm{div}\mathbf{u} = 0$:

$$\mathrm{div}\left(\frac{-\partial(u^j \mathbf{u})}{\partial x^j} + \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^*\right) = \mathrm{div}\left(\frac{-\partial(u^j \mathbf{u})}{\partial x^j} + \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^*\right) + \mathrm{div}\frac{\partial \mathbf{u}}{\partial t} = \mathrm{div}\left(\frac{-\partial(u^j \mathbf{u})}{\partial x^j} + \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^* + \frac{\partial \mathbf{u}}{\partial t}\right) = \\ =: \mathrm{div}\left(\frac{\partial}{\partial t}(F, G)\right) = \frac{\partial}{\partial t}\mathrm{div}(F, G)$$

where we've defined $F, G \in C^\infty(\mathbb{R} \times N)$:

$$\begin{aligned} F &= F(t, x) \in C^\infty(\mathbb{R} \times N) \\ G &= G(t, x) \in C^\infty(\mathbb{R} \times N) \\ &\text{where} \\ \frac{-\partial(u^j \mathbf{u})}{\partial x^j} + \frac{1}{\mathrm{Re}}\Delta \mathbf{u} + \mathbf{g}^* + \frac{\partial \mathbf{u}}{\partial t} &=: \frac{\partial}{\partial t}(F, G) \end{aligned} \tag{126}$$

to be components of this time-dependent vector field over N. I've only defined 2 components F, G for the 2-dimensional case, but easily, arbitrarily finite number of components can be defined.

This F, G is the "continuous" version of the discretized F, G in Griebel, Dornseifer, and Neunhoeffer (1997) [23].

Thus, we see how the Poisson equation comes into play for incompressible Navier-Stokes equations, in order to solve for the pressure p:

$$\Delta p = \frac{\partial}{\partial t}\mathrm{div}(F, G) \tag{127}$$

Let's compare this equation with the general form of the Poisson equation, for $p, f \in C^\infty(N)$, for (Riemannian) manifold N (e.g. $N = \mathbb{R}^2, \mathbb{R}^3$):

$$\mathrm{divgrad} p = \Delta p = f \tag{128}$$

Let's drop the time-dependence (because we sought to consider the quantities of p and f at a specific point in time t, and not their changes in time, we can do this) and focus upon solving the Poisson equation..

For $p \in C^\infty(N)$, consider how the (second order) central (finite) difference discretization of the Laplacian leads to a so-called "sparse" matrix.

M. Ament, G. Knittel, D. Weiskopf, W. Straßer. *A Parallel Preconditioned Conjugate Gradient Solver for the Poisson Problem on a Multi-GPU Platform* (2010) http://www.vis.uni-stuttgart.de/~amentmo/docs/ament-pcgip-PDP-2010.pdf

40.0.1. *Lid-Driven Cavity.* cf. 5.1. Lid-Driven Cavity. Ch. 5 Example Applications, Griebel, Dornseifer, and Neunhoeffer (1997) [23].

On $S_N \subset \partial\Omega \subset \Omega \subset \mathbb{R}^2$ ("northern" N boundary of domain Ω, which is a submanifold of Ω or of $\mathbb{R}^2$,

$$\begin{aligned} u^x(x) &= \overline{u} \\ u^y(x) &= 0 \end{aligned} \qquad \forall\, x \in S_N \tag{129}$$

For the *discretization*, consider the staggered grid and how u^x or i.e. the x-component of velocity field $\mathbf{u}$, u, is at the center of each edge (or i.e. face), and by convention, the cell index (i, j) corresponds to the "right" face. So

$$\begin{array}{ccc} u^x_{i-1,j_{\max}+1} & & u^x_{i,j_{\max}+1} \\ & u^x_{i-\frac{1}{2},j_{\max}+\frac{1}{2}} & \\ u^x_{i-1,j_{\max}} & & u^x_{i,j_{\max}} \end{array}$$

and so

$$\begin{aligned} u^x_{i,j_{\max}+\frac{1}{2}} &= \frac{u^x_{i,j_{\max}+1} + u^x_{i,j_{\max}}}{2} = \overline{u} \text{ or} \\ u^x_{i,j_{\max}+1} &= 2\overline{u} - u^x_{i-1,j_{\max}}\,; \qquad i = 1 \dots i_{\max} \end{aligned} \tag{130}$$

Part 10. **Finite Element; Finite Element Method, Finite Element Analysis; Finite Element Exterior Calculus**

cf. Lecture 11 "Method of Weighted Residuals", Darmofal (2005) [21]

cf. Ch. 23 of "Electrostatics via Finite Elements" from Landau, Páez, and Bordeianu (2015) [26].

$$\phi_i(x) = \begin{cases} 0 & \text{for } x < x_{i-1} \text{ or } x > x_{i+1} \\ \frac{x - x_{i-1}}{h_{i-1}} & \text{for } x_{i-1} \le x \le x_i \\ \frac{x_{i+1} - x}{h_i} & \text{for } x_i \le x \le x_{i+1} \end{cases} \tag{131}$$

Consider

$$*\mathbf{d}\phi(x) \in \Omega^{d-1}(N)$$

e.g. $d = 3$, $d - 1 = 2$.

Consider $\omega \in \Omega^p(N)$, in general, $f \in C^\infty(N)$, in general.

$$\int_\Omega d(f\omega) = \int_\Omega d(f\omega_{i_1 \dots i_p} dx^{i_1} \wedge \dots \wedge dx^{i_p}) = \int_\Omega \frac{\partial(f\omega_{i_1 \dots i_p})}{\partial x^d} dx^j \wedge dx^{i_1} \wedge \dots \wedge dx^{i_p} = \int_\Omega (d\omega) f + \int_\Omega (df) \wedge \omega = \\ = \int_{\partial\Omega} f\omega$$

$\dim\Omega = p + 1$.

$$\implies \int_\Omega f(d\omega) + \int_\Omega (df) \wedge \omega = \int_{\partial\Omega} f\omega \tag{132}$$

So if $\Omega = -*\mathbf{d}U$; $U \in C^\infty(N)$, $E = -dU \in \Omega^1(N) = \Gamma(T^*N)$, $f = \Phi$

$$-\int_\Omega (\mathbf{d}*\mathbf{d})U(x)\Phi(x) = -\int_{\partial\Omega} (*\mathbf{d}U)\Phi(x) + \int_\Omega \mathbf{d}\Phi \wedge *\mathbf{d}U = \int_\Omega 4\pi\rho(x)\Phi \mathrm{vol}^d$$

Writing out the components, which is true in full generality,

$$\mathbf{d}U = \frac{\partial U}{\partial x^j} dx^j$$

$$*\mathbf{d}U = \frac{\sqrt{g}}{(d-1)!} \epsilon_{i_1, i_2 \dots i_d} g^{ij} \frac{\partial U}{\partial x^j} dx^{i_2} \wedge \dots \wedge dx^{i_d}$$

Then, insightfully,

$$\mathbf{d}\Phi \wedge *\mathbf{d}U = \frac{\partial \Phi}{\partial x^{i_1}} g^{i_1 j} \frac{\partial U}{\partial x^j} \mathrm{vol}^d \equiv \langle \mathrm{grad}\Phi, \mathrm{grad}U \rangle$$

$$\implies -\int_{\partial\Omega} (*\mathbf{d}U)\Phi(x) + \int_\Omega \mathbf{d}\Phi \wedge *\mathbf{d}U = -\int_{\partial\Omega} (*\mathbf{d}U)\Phi(x) + \int_\Omega \langle \mathrm{grad}\Phi, \mathrm{grad}U \rangle = \int_\Omega 4\pi\rho(x)\Phi \mathrm{vol}^d \tag{133}$$

$\forall\, l = 0, 1, \dots N - 1$

$$\int_\Omega 4\pi\rho(x)\phi_l(x)\mathrm{vol}^d = \int_{x_{i-1}}^{x_i} 4\pi\rho(x)\frac{x - x_{i-1}}{h_{i-1}} dx + \int_{x_i}^{x_{i+1}} 4\pi\rho(x)\frac{x_{i+1} - x}{h_i} dx$$

cf. Lecture 11, Method of Weighted Residuals, `lect11.pdf`, Darmofal (2005) [21].

From heat equation,

$$\mathrm{div}(k\mathrm{grad}T) = -q \tag{134}$$

or in components

$$\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^k} \left(\sqrt{g} k \frac{\partial T}{\partial x^l} g^{kl} \right) = -q$$

In 1-dim.,

$$\frac{\partial}{\partial x^k} \left(k \frac{\partial T}{\partial x^k} \right) = -q$$

over $\Omega = \left[\frac{-L}{2}, \frac{L}{2}\right]$.

e.g. Ex. 11.1 (Steady heat diffusion),
Suppose $L = 2$,
thermal conductivity $k = 1$,
heat source $q(x) = 50e^x$.
$T(\pm 1) = 100$.

Integrate twice:

$$kT'' = -50e^x$$
$$T' = \frac{-50e^x}{k} + A_0$$
$$T = \frac{-50e^x}{k} + A_0 x + B_0$$

$$\frac{-50e}{k} + A_0 + B_0 = 100$$
$$\frac{-50e^{-1}}{k} - A_0 + B_0 = 100$$
$$B_0 = 100 + \frac{50}{k}\cosh 1$$
$$A_0 = \frac{50}{k}\sinh 1$$

$$T = \frac{-50e^x}{k} + \frac{50}{k}\sinh 1 x + 100 + \frac{50}{k}\cosh 1$$

A common approach is to approximate the solution with a series of weighted functions.
cf. 11.2. The Method of Weighted Residuals of Darmofal (2005) [21], Lecture 11, `lect11.pdf`.
Consider

$$\int_{-1}^{1} w(x)R(\overline{T}, x)dx$$

Choose N weight functions $w_j(x)$ for $1 \leq j \leq N$,
and setting N weighted residuals to 0

$$R_j(\overline{T}) = \int_{-1}^{1} w_j(x)R(\overline{T}, x)dx \equiv \text{ weighted residual for } w_j \tag{135}$$

determine appropriate weight functions,
Galerkin method is to set weight functions euqal to functions used to approximate solution

$$w_j(x) = \phi_j(x) \qquad \text{(Galerkin)}$$

e.g. 1-dim. heat diffusion:

$$w_1(x) = (1-x)(1+x)$$
$$w_2(x) = x(1-x)(1+x)$$

$$R_1(\overline{T}) = \int_{-1}^{1} w_1(x)R(\overline{T}, x)dx = \int_{-1}^{1} (1-x)(1+x)(-2\alpha_1 - 6\alpha_2 x + 50e^x)dx = \frac{-8}{3}\alpha_1 + 200e^{-1} = 0$$
$$R_2(\overline{T}) = \int_{-1}^{1} w_2(x)R(\overline{T}, x)dx = \int_{-1}^{1} x(1-x)(1+x)(-2\alpha_1 - 6\alpha_2 x + 50e^x)dx = \frac{-8}{3}\alpha_2 + 100e^{-1} - 1200e^{-1} = 0$$

and so, I would say (EY : 20170204)

$$R_j(\overline{T}) = 0 \mapsto \alpha_j \tag{136}$$

cf. 12.3. 1-D Linear Elements and the Nodal Basis of Darmofal (2005) [21], Lecture 12, The Finite Element Method for One-Dimensional Diffusion `lect12.pdf`.

40.0.2. *nodal basis.* N elements, $N+1$ degrees of freedom.

$$\widetilde{T}(x) = \sum_{i=1}^{N+1} a_i\phi_i(x) \text{ or } \widetilde{T}(x) = \sum_{i=0}^{N} a_i\phi_i(x)$$
$$\text{if } \widetilde{T}(x) = \sum_{i=1}^{N+1} \widetilde{T}(x_i)\phi_i(x) \text{ or } \widetilde{T}(x) = \sum_{i=0}^{N} \widetilde{T}(x_i)\phi_i(x)$$

iff

$$\widetilde{T}(x_j) = \sum_{i=0}^{N} a_i\phi_i(x_j) = \sum_{i=0}^{N} a_i\delta_{ij} = a_j$$

Now

$$\phi_i(x) = \begin{cases} 0 & \text{for } x < x_{i-1} \\ \frac{x - x_{i-1}}{\Delta x_{i-1}} & \text{for } x_{i-1} < x < x_i \\ \frac{x_{i+1} - x}{\Delta x_i} & \text{for } x_i < x < x_{i+1} \\ 0 & \text{for } x > x_{i+1} \end{cases}$$

But, if we only have "nodal points" or grid points, or given points on $\mathbb{R}$ to evaluate $\widetilde{T}$ of $\{x_0, x_1, \dots x_N\}$ (for $N+1$ points), then how can we define the following basis functions?

$$\phi_0(x) = \begin{cases} \frac{x_1 - x}{\Delta x_0} & \text{for } x_0 < x < x_1 \\ 0 & \text{for } x > x_1 \end{cases}$$

$$\phi_1(x) = \begin{cases} 0 & \text{for } x < x_0 \\ \frac{x - x_0}{\Delta x_0} & \text{for } x_0 < x < x_1 \\ \frac{x_2 - x}{\Delta x_1} & \text{for } x_1 < x < x_2 \\ 0 & \text{for } x > x_2 \end{cases}$$

$$\phi_N(x) = \begin{cases} 0 & \text{for } x < x_{N-1} \\ \frac{x - x_{N-1}}{\Delta x_{N-1}} & \text{for } x_{N-1} < x < x_N \\ \frac{x_{N+1} - x}{\Delta x_N} & \text{for } x_N < x < x_{N+1} \\ 0 & \text{for } x > x_{N+1} \end{cases}$$

$$\phi_{N+1}(x) = \begin{cases} 0 & \text{for } x < x_N \\ \frac{x - x_N}{\Delta x_N} & \text{for } x_N < x < x_{N+1} \end{cases}$$

Consider defining the residual $R = R(\widetilde{T}, x)$ as

$$R(\widetilde{T}, x) := \text{div}(k\text{grad}T) + q$$

and consider multiplying it by so-called "weights", Φ, which is set to be ϕ_j's.

$$\int_\Omega d*(kdT)\Phi = \int_\Omega -q\Phi\text{vol}^d \Longrightarrow \int_\Omega (d*(kdT) + q\text{vol}^d)\Phi = 0$$
$$\Longrightarrow \int_{\partial\Omega} *(kdT)\Phi 0 \int_\Omega \langle \text{grad}\Phi, k\text{grad}T\rangle + \int_\Omega q\Phi\text{vol}^d$$

If $\Phi = \phi_j(x)$, s.t. $\phi_j(x) = 0$ on $\partial\Omega$,

$$-\int_\Omega \langle \text{grad}\phi_j, k\text{grad}T\rangle + \int_{\mathcal{O}_j} q\phi_j\text{vol}^d = 0$$

e.g. $\mathcal{O}_j = [x_{j-1}, x_{j+1}]$

if

$$\phi_j(x) = \begin{cases} 0 & \text{for } x < x_{i-1} \\ \frac{x - x_{i-1}}{\Delta x_{i-1}} & \text{for } x_{i-1} < x < x_i \\ \frac{x_{i+1} - x}{\Delta x_i} & \text{for } x_i < x < x_{i+1} \\ 0 & \text{for } x > x_{i+1} \end{cases}$$

$$\int_{x_{i-1}}^{x_i} \frac{k \text{grad} T}{\Delta x_{i-1}} \text{vol}^d - \int_{x_i}^{x_{i+1}} \frac{k \text{grad} T}{\Delta x_i} \text{vol}^d + \int_{\mathcal{O}_j} q \phi_j \text{vol}^d = 0$$

Suppose $\widetilde{T}(x) = \sum_{i=1}^{N+1} a_i \phi_k(x)$,

$$\int_{x_{i-1}}^{x_i} \frac{k}{\Delta x_{i-1}^2}(a_i - a_{i-1}) - \int_{x_i}^{x_{i+1}} \frac{k}{\Delta x_i^2}(-a_i + a_{i+1}) + \int_{\mathcal{O}_j} q \phi_j \text{vol}^d = 0 =$$

$$= \frac{a_i - a_{i-1}}{\Delta x_{i-1}^2} \int_{x_{i-1}}^{x_i} k - \frac{a_{i+1} - a_i}{\Delta x_i^2} \int_{x_i}^{x_{i+1}} k + \int_{\mathcal{O}_j} q \phi_j \text{vol}^d = 0 = k \frac{a_i - a_{i-1}}{\Delta x_{i-1}} - k \frac{a_{i+1} - a_i}{\Delta x_i} + \int_{\mathcal{O}_j} q \phi_j \text{vol}^d = 0$$

If $q = 50e^x$, and consider that

$$\int 50 e^x \phi_j \text{vol}^d = \frac{50}{\Delta x_{i-1}}(x e^x - e^x - x_{i-1} e^x)$$

Part 11. CUB, NCCL

41. CUB

41.1. **Striped arrangement.** Recall our notation

$$i_x \in \{0, 1, \dots M_x - 1\} \subset \mathbb{Z} \iff \texttt{threadIdx.x}$$

with M_x corresponding to `blockDim.x`

Consider

$$F : \{0, 1, \dots M_x - 1\} \subset \mathbb{Z} \to \mathbb{K}^d$$
$$F(i_x) = (F^{(0)}(i_x), F^{(d)}(i_x) \dots F^{(d)}(i_x))$$

with $d \iff$ items per thread, or `ITEMS_PER_THREAD`.

5.3 Flexible data arrangement across threads, Striped arrangement mentions a *logical stride* $S_x \in \mathbb{Z}^+ \iff$ `BLOCK_THREADS`.

It goes on to say (CUB documentation) "thread i owns items $(i), (i +$ `block-threads`$), \dots (i +$ `block-threads` $*$ (`items-per-thread` $- 1)$).

Denote this `block-threads` $*$ `items-per-thread`, stripped arrangement, as $S_x d$.

What is this `BLOCK_THREADS`? The thread block size in threads, from template parameters for the cub LoadDirectStriped documentation. It wasn't obvious to me at first, until reading the documentation on its template parameters.

So

$$S_x = M_x$$

Is this enforced somehow? What happens if $S_x > 1024$, the hardware limit on number of threads on a single block?

So we are asked to consider

$$i_x, M_x, d \mapsto \{i_x, i_x + M_x, \dots i_x + M_x l^x, \dots i_x + M_x(d-1)\}_{l^x = 0, 1, \dots d-1}$$

Consider the "maximal" case:

$$i_x = M_x - 1, M_x, d \mapsto \{M_x - 1, 2M_x - 1, \dots M_x d - 1\}$$

So we have $M_x d$ total values to consider for this single (thread) block.

In reality, what's really going on is the flatten functor, necessitated by just how C/C++ doesn't have multi-dimensional arrays.

$$F : \{0, 1, \dots M_x - 1\} \subset \mathbb{Z} \to \mathbb{K}^d \mapsto F : \{0, 1 \dots M_x d - 1\} \subset \mathbb{Z} \to \mathbb{K}^1 \tag{137}$$
$$F \in L(\{0, 1 \dots M_x - 1\} \subset \mathbb{K}^d) \to L(\{0, 1 \dots M_x d - 1\}, \mathbb{K})$$

with $M_x \leq 1024$ (hardware enforced)

42. NCCL - Optimized primitives for collective multi-GPU communication

43. Theano's scan

cf. theano scan tutorial

43.1. **Example 3: Reusing outputs from the previous iterations.** For input $\begin{matrix} X : \mathbb{Z}^+ \to R - \text{Module} \\ t \mapsto X(t) \end{matrix}$, e.g. R-Module such as $\mathbb{R}^d, V, \text{Mat}_{\mathbb{R}}(N_1, N_2), \tau_s^r(V)$, and a function f that acts at each iteration, i.e. $\forall\, t = 0, 1, \dots T$,

$$f : R - \text{Module} \times R - \text{Module} \to R - \text{Module}$$
$$(X_1, X_0) \mapsto X_1 + X_0$$

, then we want to express

$$f(X(t), X(t-1)) = X(t) + X(t-1) \qquad \forall\, t = 0, 1 \dots T - 1,$$

In the end, we should get

$$X \in (R - \text{Module})^T \xrightarrow{\text{scan}} Y \in (R - \text{Module})^T$$

$Y \equiv$ `output`

In summary, the dictionary between the mathematics and the Python theano code for scan seems to be the following:

If $k = 1$, $\forall\, t = 0, 1, \dots T - 1$,

(138)

$$F : R - \text{Module} \times R - \text{Module} \to R - \text{Module} \iff \text{Python function (object) or Python lambda expression } \mapsto \texttt{scan(fn=)}$$
$$F(X(t), X(t-1)) \mapsto X(t)$$
$$(X(0), X(1), \dots X(T-1)) \in (R - \text{Module})^T \iff \texttt{scan(sequences=)}$$
$$X(-1) \in R - \text{Module} \iff \texttt{scan(outputs\_info=[ ])}$$

43.2. **Example 4 : Reusing outputs from multiple past iterations.** $\forall\, t = 0, 1, \dots T - 1$, $T \iff$ `n_steps`$= T$,

Consider

$$\begin{matrix} F : (R - \text{Module})^k \to R - \text{Module} \\ F(X(t-k), X(t-(k-1)), \dots X(t-1)) = X(t) \end{matrix} \iff \texttt{fn} \in \text{Python function (object)} \tag{139}$$

If $k = 1$, we'll need to be given $X(0) \in R - \text{Module}$. Perhaps consider $\forall\, t = -k, -(k-1), \cdots -1, 0, 1 \dots T - 1$ ("in full").

For $k > 1$, we'll need to be given (or declare) $\{X(-k), X(-(k-1)), \dots X(-1)\}$.

So for $k = 1$, $X(-1) \in R - \text{Module}$ needed $\iff$ e.g. `T.scalar()` if R-Module $= \mathbb{R}$.

for $k > 1$, $(X(-k), X(-(k-1)), \dots X(-1)) \in (R - \text{Module})^k \iff$ e.g. `T.vector()`, into `''initial''` of a Python `dict`, if R-Module $= \mathbb{R}$.

Also, for $k > 1$,

$$(-k, -(k-1), \cdots -1) \iff \texttt{taps} = [-k, -(k-1), \cdots -1] (\text{a Python } \texttt{list})$$

scan, essentially, does this:

$$(X(-k), X(-(k-1)), \dots X(-1)) \mapsto (X(0), X(1) \dots X(T-1)) \tag{140}$$
$$F(X(t-k), X(t-(k-1)) \dots X(t-1)) = X(t), \qquad \forall\, t = 0, 1, \dots T - 1$$

given $F : (R - \text{Module})^k \to R - \text{Module}$, with $T =$`n_steps`.

Stroustrup [6]

Part 12. Test-Driven Development

cf. Test-driven development, wikipedia
Programmers also apply concept to improving and debugging legacy code developed with older techniques.
Test-driven development cycle: following sequence based on book **Test-Driven Development by Example**.

(1) **Add a test**. Each new feature begins with writing a test. Write test that defines a function or improvements of a function, which should be very succinct. To write a test, developer must clearly understand the feature's specification and requirements. Developer can accomplish this through *use cases* or *user stories*, can write test in whatever testing framework appropriate to software environment.
    This is differentiating feature of test-driven development vs. writing unit tests *after* code is written.
(2) **Run all tests and see if new test fails**. This validates that the *test harness* is working correctly, shows that new test doesn't pass without requiring new code, because required behavior already exists, and rules out possibility that new test is flawed and will always pass. New test should fail for the expected reason. This step increases developer's confidence in the new test.
(3) **Write the code** At this point, only purpose of written code is to pass the test.
(4) **Run tests**
(5) **Refactor code** Growing code must be cleaned up regularly during test-driven development. Duplication removed, Object, class, module, variable, and method names should clearly represent their current purpose and use. Split method bodies to improve maintainability.
(6) **Repeat**, starting with another new test.

Part 13. Optimization

44. Differential Geometry review

I looked up Non-convex optimization for analyzing big data with Prof. Dr. Martin Kleinsteuber for TUM and came across Absil, Mahony, and Sepulchre [28].
cf. Absil, Mahony, and Sepulchre [28], 3.1.5. The manifolds $\mathbb{R}^{n\times p}$ and $\mathbb{R}_*^{n\times p}$.

On $\mathbb{R}^{n\times p}$, define chart $\varphi : \mathbb{R}^{n\times p} \to \mathbb{R}^{np}$

$$\varphi : X \mapsto \text{vec}(X)$$

where $\text{vec}(X)$ "denotes the vector obtained by stacking the columns of X below one another." Absil, Mahony, and Sepulchre [28].
$\forall\, X \in \mathbb{R}^{n\times p}$, let $(i,j) \in \{0,1,\dots n-1\} \times \{0,1,\dots p-1\}$ (or $\{1\dots n\} \times \{1\dots p\}$) for 0-based counting or 1-based counting, respectively.
$X(i,j) \in \mathbb{R}$ is the entry at the ith row, jth column.
I had introduced a flatten functor before. Consider $(i,j) \mapsto k := i + nj \in \{0,1,\dots np-1\}$ (or $k = i + n(j-1) \in \{1\dots np\}$) for **column-major ordering**.
Equip manifold $\mathbb{R}^{n\times p}$, with $(\mathbb{R}^{n\times p}, \varphi)$ chart atlas, with inner product:

$$\langle Z_1, Z_2\rangle := \text{vec}(Z_1)^T \text{vec}(Z_2) = \text{tr}(Z_1^T Z_2) \tag{141}$$

since $\text{tr}(Z_1^T Z_2) = \sum_{k=0}^{p-1}() Z_1^T Z_2)_{kk} = \sum_{k=0}^{p-1} (Z_1^T)_{kl}(Z_2)_{lk} = \sum_{k=0}^{p-1}(Z_1)_{lk}(Z_2)_{lk}$.
Norm induced by inner product is **Frobenius norm**:

$$\|Z\|_F^2 = \langle Z, Z\rangle = \text{tr}(Z^T Z) \tag{142}$$

i.e. sum of squares of elements of Z.
Observe manifold topology of $\mathbb{R}^{n\times p}$ is equivalent to its canonical topology is Euclidean space.
Let $\mathbb{R}_*^{n\times p} \equiv$ set of all $n \times p$ matrices s.t. $p \le n$ and columns are linearly independent.
Consider $(\mathbb{R}_*^{n\times p})^c$. Let $X \in (\mathbb{R}_*^{n\times p})^c$.
Then $\det X = 0$, by def. of $\mathbb{R}_*^{n\times p})^c$. Then $\det(X^T X) = \det X^T \det X = 0$.
Since $\mathbb{R}_*^{n\times p})^c = \{X \in \mathbb{R}^{n\times p} | \det(X^T X) = 0\}$. Closed.
Then $\mathbb{R}_*^{n\times p}$ is an open subset of $\mathbb{R}^{n\times p}$ and open submanifold.
Its differentiable structure is generated by chart

$$\varphi : \mathbb{R}_*^{n\times p} \to \mathbb{R}^{np}$$
$$X \mapsto \text{vec}(X) = \text{flatten}(X)$$

noncompact Stiefel manifold $\mathbb{R}_*^{n\times p}$ of full-rank $n \times p$ matrices ($p \le n$ thus necessarily)

44.1. **Immersions and Submersions.** cf. 3.2.1 Immersions and submersions of Absil, Mahony, and Sepulchre [28].
Let smooth $F : M_1 \to M_2$, $\dim M_1 = d_1$, $\dim M_2 = d_2$.
Given $x \in M_1$, rank of F at x is $\dim \text{range}(D\widehat{F}(\varphi_1(x))[\cdot] : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2})$, i.e.

$$DF(x) : T_x M_1 \to T_x M_2$$

F is **immersion** if rank of $F = d_1$, $\forall\, x \in \text{domain}(F) = M_1$, i.e. $\dim(\text{range}(DF(x))) = d_1$ (hence $d_1 \le d_2$).
F is **submersion** if rank of $F = d_2$, $\forall\, x \in \text{domain}(F) = M_1$, i.e. $\dim(\text{range}(DF(x))) = d_2$ (hence $d_1 \ge d_2$).
F immersion iff $\forall\, x \in M_1$, $\exists\, (U, u^i), (V, v^i)$, $x \in U$, with coordinate representation called *canonical immersion*

$$(u^1, \dots u^{d_1}) \mapsto (u^1 \dots u^{d_1}, 0 \dots 0)$$

F submersion iff $\forall\, x \in M_1$, $\exists\, (U, u^i), (V, v^i)$, $x \in U$, with coordinate representation called *canonical submersion*

$$(u^1, \dots u^{d_1}) \mapsto (u^1 \dots u^{d_2})$$

Regular value $y \in M_2$ of F, if rank of $F = d_2$, $\forall\, x \in F^{-1}(y)$.

Part 14. Finite State Machines, State Machines; Embedded Systems

45. Discrete Dynamics; Finite-State Machines

cf. Ch. 3 Discrete Dynamics, Lee and Seshia (2016) [29] cf. 3.1 Discrete Systems, Lee and Seshia (2016) [29] e.g. consider, count number of cars in garage by counting number of cars that enter and leave. Each entry, departure, modeled as discrete event;
discrete event occurs at instance of time rather than over time. cont., integrator case: given input $x : \mathbb{R} \to \mathbb{R}$ (cont.-time signal) signal u going into up port of counter, on the other hand,

$$u : \mathbb{R} \to \{\text{absent}, \text{present}\} \equiv \{\{\}, \{*\}\}$$

means $\forall\, t \in \mathbb{R}$, $u(t) = \text{absent} = \{\}$, meaning there's no event at that time, or
$u(t) = \text{present} \equiv \{*\}$, meaning that there is. Counter.
when event present at up/down input port,
increments/decrement its count and produces on
output new value of the count
at all other times (when both inputs absent), produces no output (count output absent)

$$c : \mathbb{R} \to \{\text{ absetn }\} \cup \mathbb{Z} \equiv \{\{\}\} \cup \mathbb{Z}$$

Clearly, no need for this counter to do anything when input is absent.

45.0.1. *Discrete signals.* Discrete signals consist of sequence of instantaneous events in time
consider signal $e : \mathbb{R} \to \{ \text{ absent } \} \cup X$, $X \in \text{Obj}\mathbf{Sets}$.

Definition 12. *Signal e is* ***discrete signal*** *if, absent most of time, and, we can count, in order, times which it's present.* $\forall$ *time it's present, we have discrete event. Count events in order is important. Let* $T \subseteq \mathbb{R}$ *s.t.*

$$T = \{t \in \mathbb{R} | e(t) \neq \text{ absent} \equiv \{\{\}\}$$

Then e discrete if $\exists$ *1-to-1 function* $f : T \to \mathbb{N}$. *That's order preserving, i.e.* $\forall\, t_1, t_2 \in T$, *where* $t_1 \leq t_2$, *then* $f(t_1) \leq f(t_2)$
reactions of discrete system triggered by environment in which discrete system operates, ***event triggered****.*
when both inputs to counter are absent, no reaction occur.

45.0.2. *Modeling Actors as Functions.* integrator actor in Sec. 2.2, Fig. 3.2, modeled by

$$I_i : \mathbb{R}^{\mathbb{R}_+} \to \mathbb{R}^{\mathbb{R}_+}$$
$$I_i : L(\mathbb{R}, \mathbb{R}_+) \to L(\mathbb{R}, \mathbb{R}_+)$$

$y = I_i(x)$
where $i \equiv$ initial value of integration, x, y cont.-time signal e.g., if $i = 0$, $\forall\, t \in \mathbb{R}_+$, $x(t) = 1$, then

$$y(t) = i + \int_0^t x(\tau) d\tau = t$$

Similarly, counter in Fig. 3.1, modeled by

$$c_i : (\mathbb{R}_+ \to \{ \text{ absent, present} \})^P \to (\mathbb{R}_+ \to \{ \text{ absent } \} \cup \mathbb{Z})$$

where $P = \{\text{up, down }\} \equiv \{u, d\}$ = set of input ports. For example,

$$u : \mathbb{R}_+ \to \{ \text{ absent, present } \} \equiv \{\emptyset, \{*\}\}$$
$$u : t \mapsto u(t) \in \} \equiv \{\emptyset, \{*\}\}$$

Lee and Seshia (2016) [29] notation:

$$A^B \equiv \text{ set of all functions from } B \text{ to } A$$

A particular reaction will observe values of inputs at particular time t, and calculate output values for some time t. Suppose actor has input ports $P = \{p_1 \dots p_N\}$. Assume $\forall\, p \in P$, $V_p \in \text{Obj}\mathbf{Sets} \equiv$ values that maybe received on port p, when input present. $V_p \equiv$ type of port p. At reaction, $\forall\, p \in P$, let value of p, $p \in V_p \cup \{ \text{ absent } \}$ (Lee and Seshia seems to reuse notation, notation by context, here; but it's troublesome) valuation of inputs P is assignment of value in V_p, $\forall$ variable $p \in P$, or assertion that p absent. I think Lee and Seshia is terribly confused in the writing and lack of clarity is horrendous and pernicious.

45.1. **Discrete Signals.** Assume output is absent at times t where reaction doesn't occur.
e.g.

$$Q = \{ \text{ count } \}$$

$V_{\text{count}} = \mathbb{Z}$
at reaction, count assigned count of cars in garage.
So ports $P = \{p_1 \dots p_N\}$, output ports $Q = \{q_1 \dots q_M\}$, $\forall\, p \in P$, $V_p \in \mathbf{Set}$ or $\mathbb{R}$, $\forall\, q \in Q$, $V_q \in \mathbf{Set}$ or $\mathbb{R}$, at reaction,

$$y : P \to V_p \bigcup \{\emptyset\}$$
$$z : Q \to V_q \bigcup \{\emptyset\}$$

45.2. **Finite-State Machines.**

45.2.1. *Notion of a State.* cf. Sec. 3.2 The Notion of State, Lee, Seshia (2011) [**?**]
state, state of a system, formally, encoding of everything about past that has an affect on system's reaction to amount and future inputs
cf. Sec. 3.3 Finite-State Machines, Lee, Seshia (2011) [**?**]
Finite state machine; assume states $\equiv S \in \mathbf{FiniteSet}$

45.3. **Transitions.** cf. Sec. 3.3.1 Transitions, Lee, Seshia (2011) [**?**]
transition - directed edge, $\forall$ transition, label with "guard/action"
guard - determines whether transition maybe taken on reactions
action -specifies what outputs are produced on each reaction
guard is a **predicate** (boolean valued) s.t. true when transition should be taken.
e.g. example 3.4, FSM model for garage counter

$\neg$ $\sim$!	negation, not	$\neg A$ true, iff A false
$\wedge$ &	logical conjunction, and	if A and B both true, T, $A \wedge B$ true. Else, it's false.
$\vee$ + $\|\|$	logical (inclusion) disjunction, or	if A or B (or both) true, $A \vee B$ false, if both false, $A \vee B$ false
$\oplus$ $\veebar$	exclusive disjunction, xor	$A \oplus B$ true when either A or B, but not both are true.

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \oplus B$
T	T	F	T	T	F
T	F	F	F	T	T
F	T	T	F	T	T
F	F	T	F	F	F

action (in reference to Lee, Seshia (2011)), on a transition, action specifies resulting valuation on output ports when transition is taken.
"When assigning a value to an output port, we use notation name := value to distinguish assignment from a predicate, which would be written name = value.
Let
ports $P_{\text{in}} = \{p_1^{(\text{in})}, \dots p_{N_{\text{in}}^{(\text{in})}}\}$
output ports $P_{\text{out}} = \{p_1^{(\text{out})}, \dots p_{N_{\text{out}}^{(\text{out})}}\}$

At a reaction,

$$\begin{aligned} i &: P_{\text{in}} \to V_{P_{\text{in}}} \cup \{\emptyset\} \\ i &: p \mapsto i(p) \\ o &: P_{\text{out}} \to V_{P_{\text{out}}} \cup \{\emptyset\} \\ o &: p \mapsto o(p) \end{aligned} \tag{143}$$

Given

$$i(p_1^{(\text{in})}, \dots p_{N_{\text{in}}}^{(\text{in})}) \in (V_{P_{\text{in}}} \cup \{\emptyset\})^{P_{\text{in}}} \times S$$

where $S \equiv$ states of the system,

guard, g_{guard},

$$\begin{aligned} g_{\text{guard}} &: (V_{P_{\text{in}}} \cup \{\emptyset\})^{P_{\text{in}}} \times S \to \mathbb{Z}_2 = \{\text{true, false}\} \\ g_{\text{guard}}&(p_1^{(\text{in})}, \dots p_{N_{\text{in}}}^{(\text{in})}) \in \{\text{true, false}\} \end{aligned} \tag{144}$$

action a,

$$a : g_{\text{guard}}, e \mapsto o(p_1^{(\text{out})}, \dots p_{N_{\text{out}}}^{(\text{out})}) \tag{145}$$

where $e \in E$, $E \equiv$ set of all possible transitions.

cf. Sec. 3.3.3. Update Functions of Lee, Seshia (2011)
(States, Inputs, Outputs, update, initial State) $\equiv (S, I, O, u, s_0)$,
where $u : S \times I \to S \times O$, and update $\equiv$ transition function (synonymous)

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial u^j \mathbf{u}}{\partial x^j} + \text{grad}p = \frac{1}{\text{Re}}\Delta \mathbf{u} + \mathbf{g}^*$$
$$\text{div}(\mathbf{u}) = 0$$

$$\text{divgrad}p = \Delta p = \text{div}\left(-\frac{\partial (u^j \mathbf{u})}{\partial x^j} + \frac{1}{\text{Re}}\Delta \mathbf{u} + \mathbf{g}^*\right)$$

$$F := (u^x)^n + \delta t \left[\frac{1}{\text{Re}}\Delta u^x - \frac{\partial (u^j u^x)}{\partial x^j} + (g^*)^x\right]$$

$$G := (u^y)^n + \delta t \left[\frac{1}{\text{Re}}\Delta u^y - \frac{\partial (u^j u^y)}{\partial x^j} + (g^*)^y\right]$$

$$u_{i,j}^{(n+1)} = F_{i,j}^{(n)} - \frac{\delta t}{\delta x}(p_{i+1,j}^{(n+1)} - p_{i,j}^{(n+1)}) \qquad \begin{matrix} i = 1 \ldots i_{\max} - 1 \\ j = 1 \ldots j_{\max} \end{matrix}$$

$$v_{i,j}^{(n+1)} = G_{i,j}^{(n)} - \frac{\delta t}{\delta y}(p_{i,j+1}^{(n+1)} - p_{i,j}^{(n+1)}) \qquad \begin{matrix} i = 1 \ldots i_{\max} \\ j = 1 \ldots j_{\max} - 1 \end{matrix}$$

REFERENCES

[1] Zed A. Shaw. **Learn C the Hard Way: Practical Exercises on the Computational Subjects You Keep Avoiding (Like C)** (Zed Shaw's Hard Way Series) 1st Edition. Addison-Wesley Professional; 1 edition (September 14, 2015) ISBN-13: 978-0321884923.
[2] Eli Bendersky. Are pointers and arrays equivalent in C?
[3] Brian W. Kernighan, Dennis M. Ritchie. **C Programming Language**, 2nd Ed. 1988.
[4] Peter van der Linden. **Expert C Programming: Deep C Secrets** 1st Edition. Prentice Hall; 1st edition (June 24, 1994) ISBN-13: 978-0131774292
[5] Leo Ferres. "Memory management in C: The heap and the stack". Memory management in C: The heap and the stack
[6] Bjarne Stroustrup. **The C++ Programming Language**, 4th Edition. Addison-Wesley Professional; 4 edition (May 19, 2013). ISBN-13: 978-0321563842
[7] Carlo Ghezzi, Mehdi Jazayeri. **Programming Language Concepts** 3rd Edition. Wiley; 3 edition (June 23, 1997). ISBN-13: 978-0471104261
https://vowi.fsinf.at/images/7/72/TU_Wien-Programmiersprachen_VL_(Puntigam)_-_E-Book_SS08.pdf
[8] Scott Meyers. **Effective Modern C++: 42 Specific Ways to Improve Your Use of C++11 and C++14**. 1st Edition. O'Reilly Media; 1 edition (December 5, 2014) ISBN-13: 978-1491903995

[9] Trevor Hastie, Robert Tibshirani, Jerome Friedman. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Second Edition (Springer Series in Statistics) 2nd ed. 2009. Corr. 7th printing 2013 Edition. ISBN-13: 978-0387848570. https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf
[10] Jared Culbertson, Kirk Sturtz. *Bayesian machine learning via category theory.* arXiv:1312.1445 [math.CT]
[11] John Owens. David Luebki. *Intro to Parallel Programming. CS344.* **Udacity** http://arxiv.org/abs/1312.1445 Also, https://github.com/udacity/cs344
[12] John Cheng, Max Grossman, Ty McKercher. **Professional CUDA C Programming**. 1st Edition. Wrox; 1 edition (September 9, 2014). ISBN-13: 978-1118739327
[13] CS229 Stanford University. http://cs229.stanford.edu/materials.html
[14] Andrew Lavin, Scott Gray. *Fast Algorithms for Convolutional Neural Networks* arXiv:1509.09308 [cs.NE]
[15] Stanley B. Lippman, Josée Lajoie, Barbara E. Moo. **C++ Primer** (5th Edition). Addison-Wesley Professional; 5 edition (August 16, 2012) ISBN-13: 978-0321714114
[16] Richard Fitzpatrick. "Computational Physics." http://farside.ph.utexas.edu/teaching/329/329.pdf
[17] M. Hjorth-Jensen, **Computational Physics**, University of Oslo (2015) http://www.mn.uio.no/fysikk/english/people/aca/mhjensen/
[18] Prof. Dr. Michael Bader (*Lecturer*); Alexander Pöppl, Valeriy Khakhutskyy. (*Tutorials*). HPC - Algorithms and Applications - Winter 16. Winter 16/17. **TUM**. https://www5.in.tum.de/wiki/index.php/HPC_-_Algorithms_and_Applications_-_Winter_16
[19] Jason Sanders, Edward Kandrot. **CUDA by Example: An Introduction to General-Purpose GPU Programming** 1st Edition. Addison-Wesley Professional; 1 edition (July 29, 2010). ISBN-13: 978-0131387683
[20] M.E.J. Newman and G.T. Barkema. **Monte Carlo Methods in Statistical Physics**. Oxford University Press. 1999.
[21] David Darmofal. "16.901 Computational Methods in Aerospace Engineering, Spring 2005." (Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu (Accessed 12 Jun, 2016). License: Creative Commons BY-NC-SA
[22] Joel H. Ferziger and Milovan Peric. **Computational Methods for Fluid Dynamics**. Springer; 3rd edition (October 4, 2013). ISBN-13: 978-3540420743
I used the 2002 edition since that was the only copy I had available.
[23] Michael Griebel, Thomas Dornsheifer, Tilman Neunhoeffer. **Numerical Simulation in Fluid Dynamics: A Practical Introduction (Monographs on Mathematical Modeling and Computation)**. SIAM: Society for Industrial and Applied Mathematics (December 1997). ISBN-13: 978-0898713985 QA911.G718 1997
See also Software of Research group of Prof. Dr. M. Griebel, Institute für Numerische Simulation http://wissrech.ins.uni-bonn.de/research/software/
[24] Kyle E. Niemeyer, Chih-Jen Sung. *Accelerating reactive-flow simulations using graphics processing units.* 51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition 07-10 January 2013, Grapevine, Texas. American Institute of Aeronautics and Astronautics. AIAA 2013-0371
[25] Carlos Henrique Marchi, Roberta Suero, Luciano Kiyoshi Araki. *The Lid-Driven Square Cavity Flow: Numerical Solution with a 1024 x 1024 Grid.* **J. of the Braz. Soc. of Mich. Sci. & Eng. 186**, Vol. XXXI, No. 3, July-September 2009. ABCM
[26] Rubin H. Landau, Manuel J Páez, Cristian C. Bordeianu. **Computational Physics: Problem Solving with Python**. 3rd Edition. Wiley-VCH; 3 edition (September 8, 2015). ISBN-10: 3527413154 ISBN-13: 978-3527413157
[27] Vladimir I. Arnold, Valery V. Kozlov, Anatoly I. Neishtadt, E. Khukhro. **Mathematical Aspects of Classical and Celestial Mechanics (Encyclopaedia of Mathematical Sciences)** 3rd Edition. Encyclopaedia of Mathematical Sciences (Book 3). Springer; 3rd edition (November 14, 2006). ISBN-13: 978-3540282464
[28] P.-A. Absil, R. Mahony, and R. Sepulchre. **Optimization Algorithms on Matrix Manifolds**. Princeton University Press. 2008. ISBN 978-0-691-13298-3 https://press.princeton.edu/titles/8586.html
[29] Edward Ashford Lee, Sanjit Arunkumar Seshia. **Introduction to Embedded Systems: A Cyber-Physical Systems Approach**. (MIT Press) 2nd Edition. The MIT Press; (December 30, 2016). ISBN-13: 978-0262533812