Efficient Memory Management

Wahid Redjeb[1,2]
wahid.redjeb@cern.ch

[1]CERN, European Organization for Nuclear Research, Meyrin, Switzerland
[2]RWTH Aachen University, III. Physikalisches Institut A, Aachen, Germany,

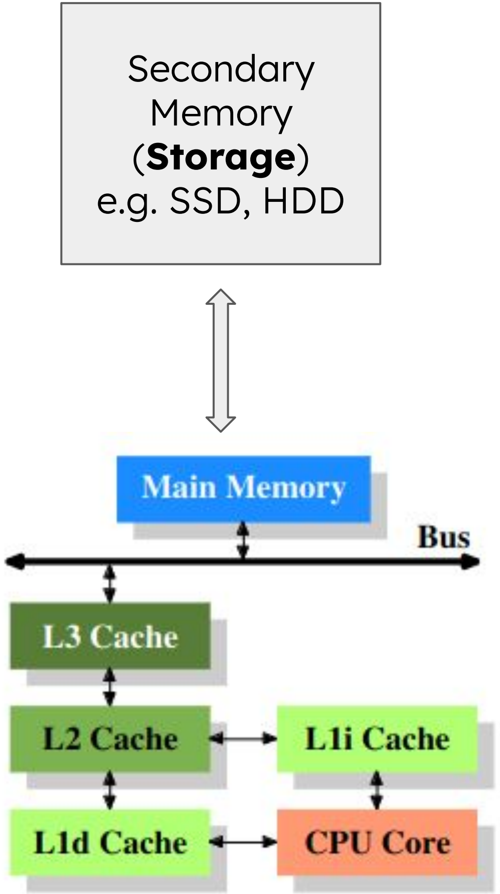
What is memory?

- In general, memory refers to the storage a program uses to write and read data

- Memory is usually managed through **virtual memory OS**
  - Map different hardware to address spaces
    - RAM
    - GPU memory
    - HBM
    - Disk space: swap or mmap files

Different types of memory

- Secondary Memory (SSD, HDD) [variable storage]
- Main Memory (RAM) [usually tens of GBs]
- 3 levels of cache
    - Small [32/64kB] separate L1 (I+D) caches for each core.
    - Medium [256kB - 6MB] combined L2 cache, perhaps shared among some cores.
    - Large [4 - 20MB] combined L3 cache shared between all cores



Different types of memory - Latency

lower latency
higher bandwidth

memory	latency	bandwidth	capacity	cost
L1 cache	2 ns	100 TB/s	64 kB / core	
L2 cache	6 ns	50 TB/s	512 kB / core	
L3 cache	20 ns	(?) 10 TB/s	4 MB / core	1-2 $/MB
HBM RAM	200 ns	2 TB/s	up to 80 GB / device	20-100 $/GB
DDR RAM	200 ns	20-200 GB/s	up to 64 GB / core	3-4 $/GB
SSD	50-100 us	5 GB/s	30 TB / drive	100-200 $/TB
HDD	2 ms	300 MB/s	30 TB / drive	10-20 $/TB

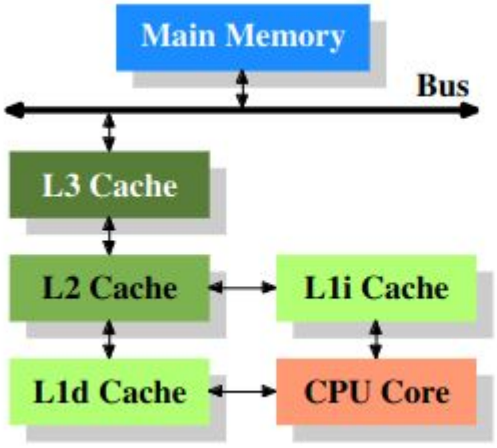
based on the performance of an AMD Rome EPYC CPU, NVIDIA A100 GPU, and datacentre-grade SSDs and HDDs

lower cost
higher capacity

A.Bocci, CERN

Caches

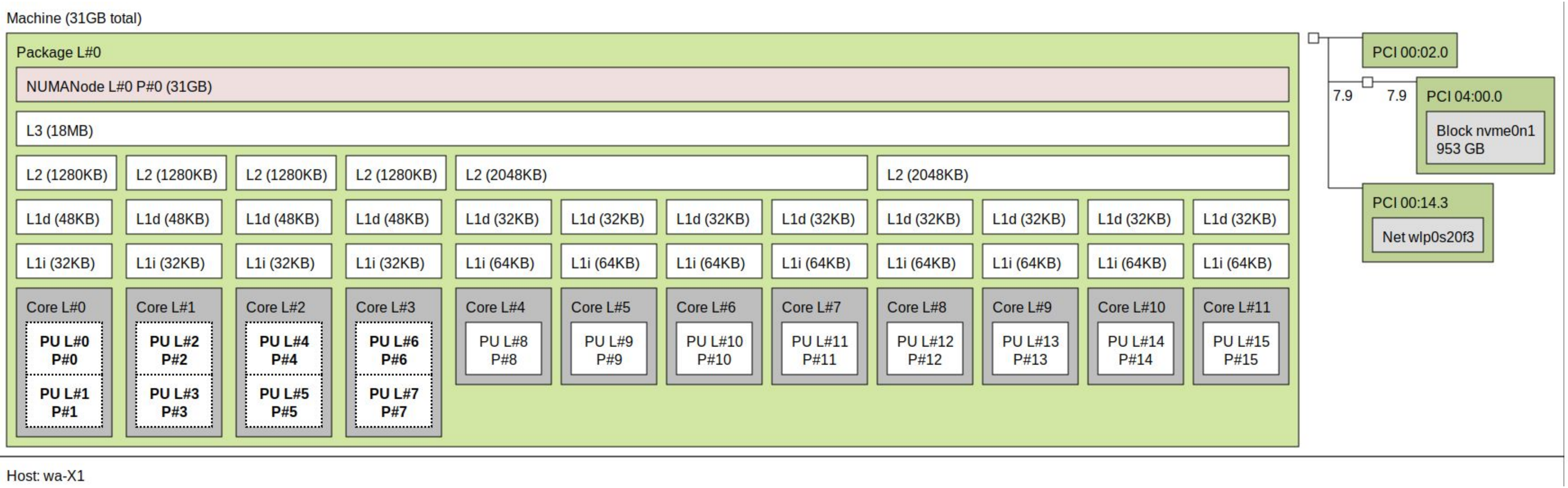
- CPU looks for data in L1 -> L2 -> L3 -> RAM
- Data area loaded in cache in unit of **cache lines**
  - **Usually 64bytes, but depends on architecture**
- Decision in which hierarchy level some data will stay depends on hardware
  - Memory controllers looks at **memory access patterns**
  - **Cache locality**
  - Cache lines might be promoted or demoted depending on these patterns
- Cache eviction policies
  - LRU (Last-recently-used)
  - FIFO (First-in-First-Out)
  - Random



Different types of memory

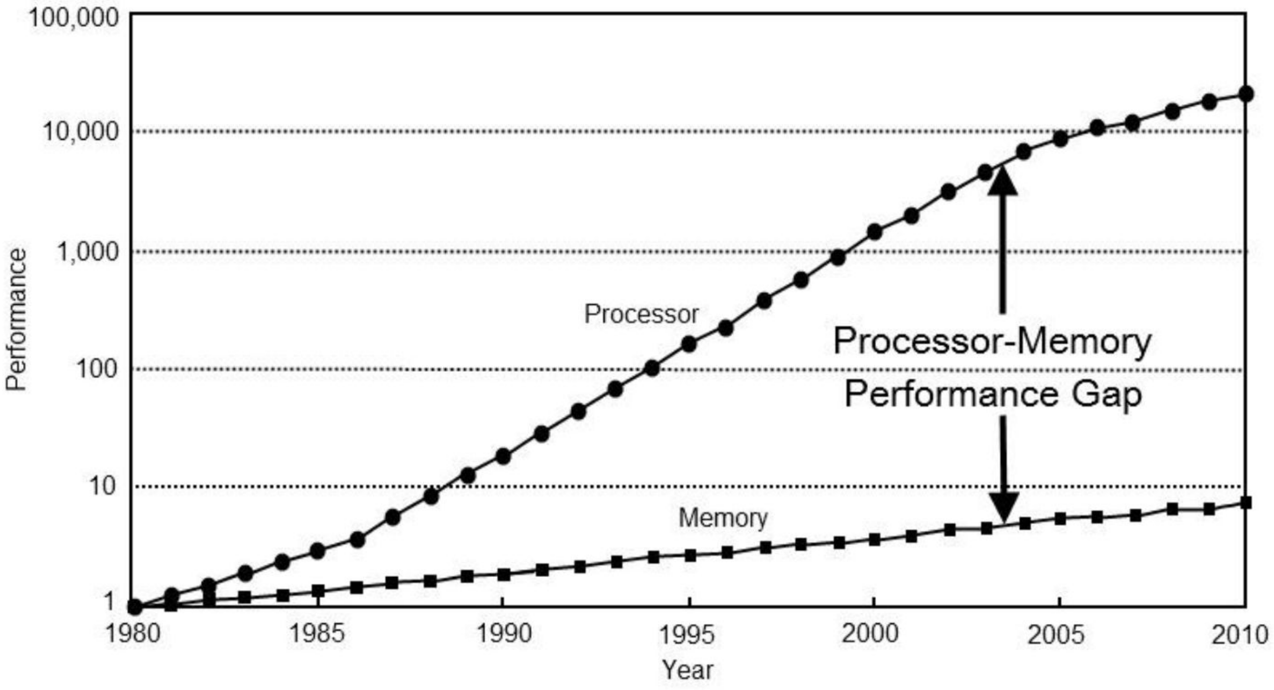
Have a look at your system

- `lscpu`
- `lstopo`



Why are we interested in memory?

- Most of the memory is *very* **slow** compared to CPU operations



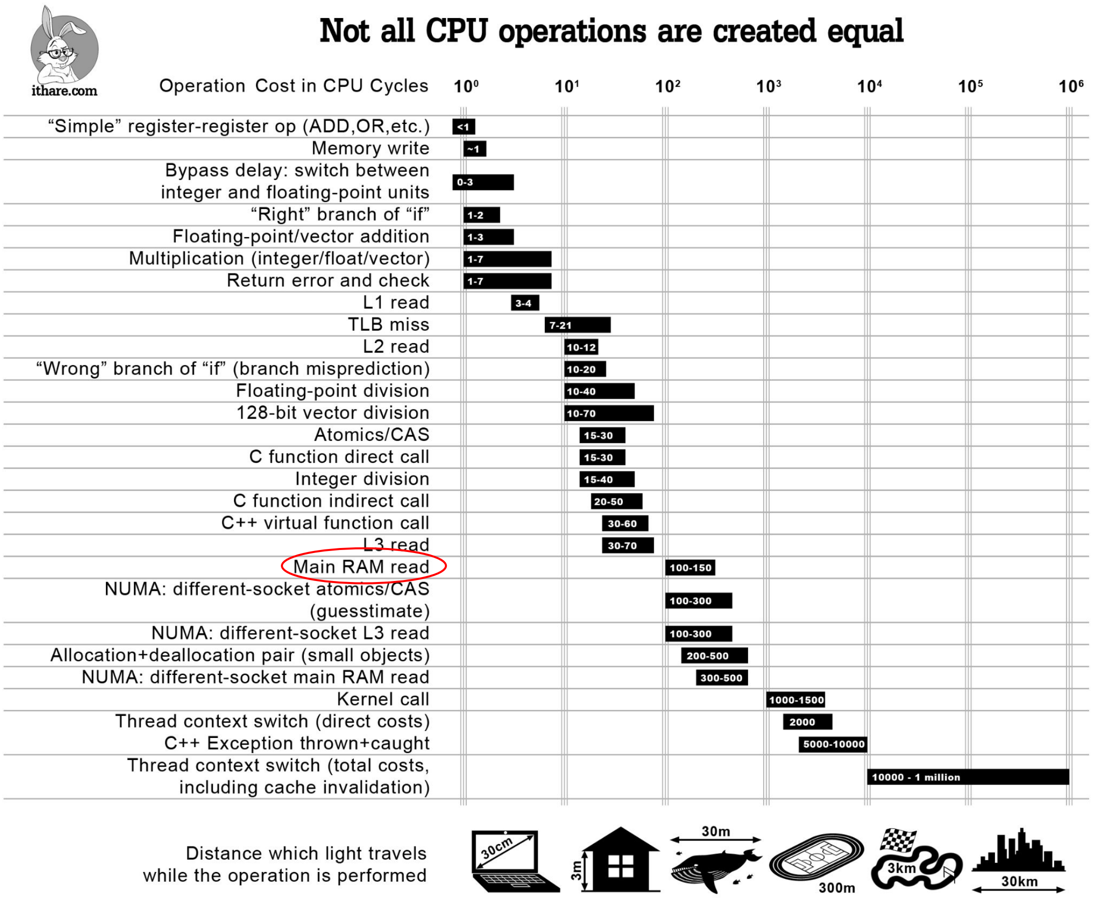
Why are we interested in memory?

Everything here is better than reading from main memory

When writing efficient code, the most important thing to address is memory

But there's no general rule, the best solution to adopt depends on your **data**

- **Know your data**



Data oriented design

- Data temporal locality
  - Exploit data that has just been read or written to memory
  - Exploit data that is “hot” in the processor cache
- Data spatial locality
  - Fully exploit cache line: work on adjacent data!
  - Avoid pointers chasing if possible
    - Pointers to pointers to pointers …
  - AoS → SoA
- Hide memory latency
  - Prefetch data in advance while working on previous data
  - Keep the processor busy while more data is fetched
  - Common strategy on GPU
- If possible avoid dynamic allocations
  - Remember: understand your data
  - Custom allocators
- Avoid high level abstraction

BASICS

Size of Data Types

Size of a type corresponds to the number of bytes needed to store an object of that type

- Use `sizeof()` operator to get the size of your type
  - Try it yourself with some common types
  - `char`, `int`, `float`, `double`, `int *`, `std::vector<double>`, `std::vector<int>`
- Define your own Class / Struct with different members and get the size of your class
  - Try to change the order of the members
  - Try to add a bool to your members

Size of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

Size of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

13 bytes

Size of Data Types

```cpp
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

13 bytes

```cpp
sizeof(MyStruct) -> 24
```

Size of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

13 bytes

Alignment of data types

- To have a more efficient memory access from the CPU data types are ***aligned***
- *Alignment is an integer value representing the number of bytes between successive addresses at which objects of this type can be allocated.*
    - Type with alignment of 4 can be allocated only every 4 bytes
- The valid alignment values are **non-negative integral powers of two.**
- The operator **`alignof`**`()` gives you the alignment of a type
- You can request stricter alignment using **`alignas`**`()` specifier
- The alignment of any class object is given by the largest of the alignment of its members

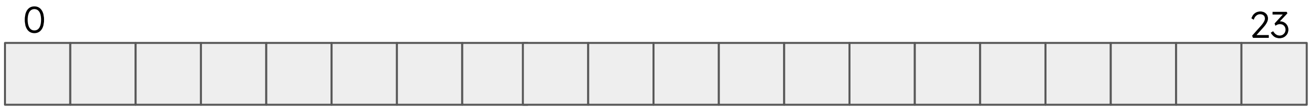
Alignment of Data Types

```cpp
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

```cpp
alignof(int) -> 4
alignof(double) -> 8
```

```cpp
alignof(bool) -> 1
```

13 bytes



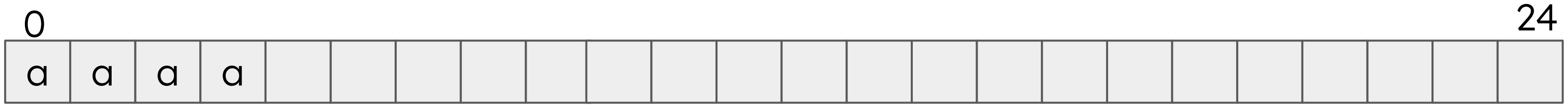
Alignment of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

```
alignof(int) -> 4
alignof(double) -> 8
```

```
alignof(bool) -> 1
```

13 bytes



Alignment of Data Types

```cpp
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

```cpp
alignof(int) -> 4
alignof(double) -> 8
```

```cpp
alignof(bool) -> 1
```

13 bytes

0 … 24

a	a	a	a					b	b	b	b	b	b	b	b								

Alignment of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

```
alignof(int) -> 4
alignof(double) -> 8
```

```
alignof(bool) -> 1
```

13 bytes

0 24

a	a	a	a					b	b	b	b	b	b	b	b	c							

Alignment of Data Types

```
struct MyStruct {
    int a; //4 bytes
    double b; //8 bytes
    bool c; // 1 byte
};
```

```
alignof(int) -> 4
alignof(double) -> 8
```

```
alignof(bool) -> 1
```

13 bytes

0 ... 24

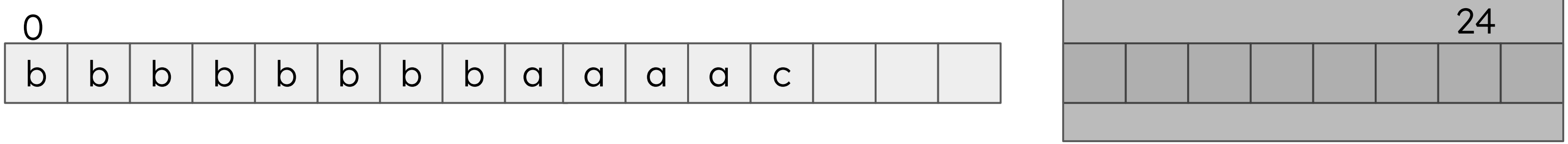
a	a	a	a					b	b	b	b	b	b	b	b	c							

Additional padding is required to properly align each data member!
Let's optimize this

Alignment of Data Types

```
struct MyStruct {
    double b; // 8bytes
    int a; //4 bytes
    bool c; // 1 byte
};
```

13 bytes



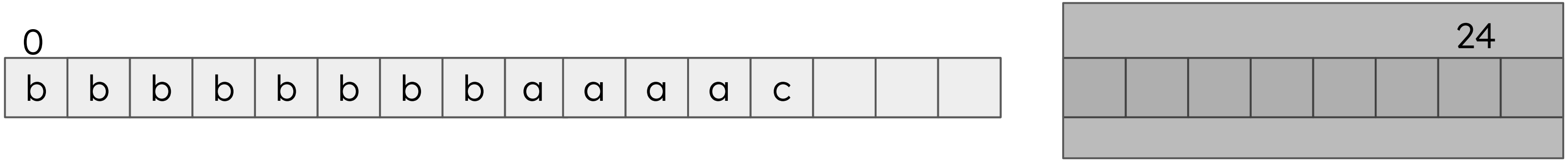
Struct is much more packed now, we are already saving 8 bytes

Alignment of Data Types - Optimize memory design

```cpp
struct MyStruct {
    double b; // 8bytes
    int a; //4 bytes
    bool c; // 1 byte
};
```

- Put data members in decreasing size order
- Group data members based on their size and alignment
  - Dedicate some time to understand if you are introducing padding and if you can avoid it
- Group data members based on their usage
  - Better to have data members that are used together within a single cache line!
    - Cache line usually are 64bytes.

13 bytes



Struct is much more packed now, we are already saving 8 bytes

Exercise

- Create a Class or struct for a Particle with the following members
  - 1 `const std::string` to hold the particle's name;
  - 3 `double`s for the x, y, z velocities
  - 3 `bool`s to mark if there has been a collision along the x, y z directions
  - 1 `float` for the mass
  - 1 `float` for the energy
  - 3 `double`s for the px, py, pz coordinates
  - 1 `const int` for the particle's id
  - 1 `static int` to keep track of the total number of objects
- What is the best order for your members?

Memory operations - Allocation

- `void* std::malloc(std::size_t size);`
  - Allocates size bytes of uninitialized storage.
  - If successful returns pointer to the beginning of newly allocated memory
  - On failure returns a null pointer
  - Suitable alignment for any scalar type
  - **Nothing is initialized**, just raw memory
  - Requires manual freeing of the memory
- `void* std::calloc(std::size_t num, std::size_t size);`
  - Allocate memory for an array of `num` objects of size `size`
  - Initialized it to all bits zero
- `void* std::aligned_alloc(std::size_t alignment, std::size_t size);`
  - Allocate a block of memory of at least size bytes
  - The memory buffer is aligned to at least alignment bytes
    - Useful in SIMD to avoid Cache False Sharing
      - Require memory aligned to a cache line (64bytes usually)

Memory operations - Freeing memory

- `std::free(void* ptr);`
  - Frees allocated memory block by malloc(), calloc() aligned_alloc()
  - The content of the memory is not erased!
    - Any object in the memory is not destroyed!
  - The free operation returns the memory to the system

Memory operations - Constructing objects

- Remember, `std::malloc()`, `std::calloc()`, `std::aligned_alloc()` return raw, uninitialized memory
- `T* new T(args…);`
  - Allocates and creates object `T`
- `T* new(ptr) T{args…};`
  - `ptr` is some memory previously allocated
  - Constructs an object of type `T` using its constructor `T::T(args…)`
    - The object is created in the allocated memory at `ptr`
- `T* new(ptr) T[N]{args…};`
  - `ptr` is some memory previously allocated
  - Constructs `N` object of type `T` using its constructor `T::T(args…)`
    - The object is created in the allocated memory at `ptr`

Memory operations - Destroy objects

- Before freeing the memory (`std::free()`), you have to destroy the created objects

- `std::destroy_at(T* ptr);`
  - Calls destructor of object of type `T` at the memory address `ptr`
  - Equivalent to `ptr->~T();`
- `std::destroy_n(T* ptr, std::size_t n);`
  - Calls destructor of `n` objects of type `T` starting at the memory address `ptr`
- `std::destroy(T* first, T* last);`
  - Calls destructor of the objects of type `T` in the range `[first, last]`

(False Sharing)

- False sharing is a performance-degrading usage pattern that happens in multi-threaded application
- If two cores are accessing different elements that are in the same cache line
  - Each core has it’s own copy of the cache line
- Core0 reads the value X from the cache line
  - It marks the cache line as **exclusive**
- Core1 reads the value Y from the its copy of the same cache line
  - Both core mark the cache line as **shared**
- Core0 decides to write in address space of X
  - Marks its cache line as **updated**
    - It has to send a message to Core1 saying it has updated the cache line
- Core1 marks its cache line as **invalid**
  - Has to re-read the cache line from main memory
- Core0 has to immediately return the result back to main memory

This process for keeping caches in coherence can be extremely expensive!

Optimize Memory Access

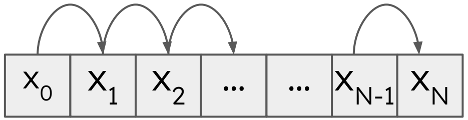
Two main principles:

- Exploit **time locality**
  - If a program accesses one memory address, there is a good chance that it will access the same address again after a short amount of time.
    - E.g loops (variable `sum` continuously updated)
- Exploit **spatial locality**
  - If a program accesses one memory address, there is a good chance that it will also access other nearby addresses.

Note: Data Structure and Memory Access are two faces of the same coin. You should design them together!

Sequential Memory Access

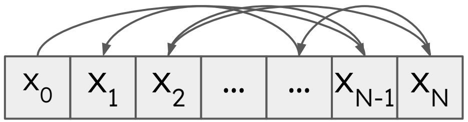
- Consecutive element access
- Good cache locality
- Good memory bandwidth
- Each cycle can read consecutive memory area
    - Cached Memory Access
- Good use of prefetcher



Perfect memory access pattern for CPUs!

Random Memory Access

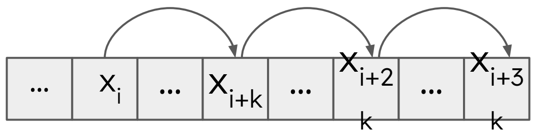
- Elements are accessed in random order
- Cache locality not ensured anymore
- Bad memory bandwidth
- Impossible to prefetch data
- Prefetcher not used



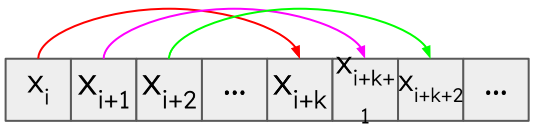
Never use this!

Strided Memory Access

- Elements are accessed at fixed intervals
- Good use of prefetcher
  - Pattern easy to predict



- Very common pattern on GPU
  - Stride size = Grid Size
  - Coalesced memory access
    - Good cache locality and bandwidth

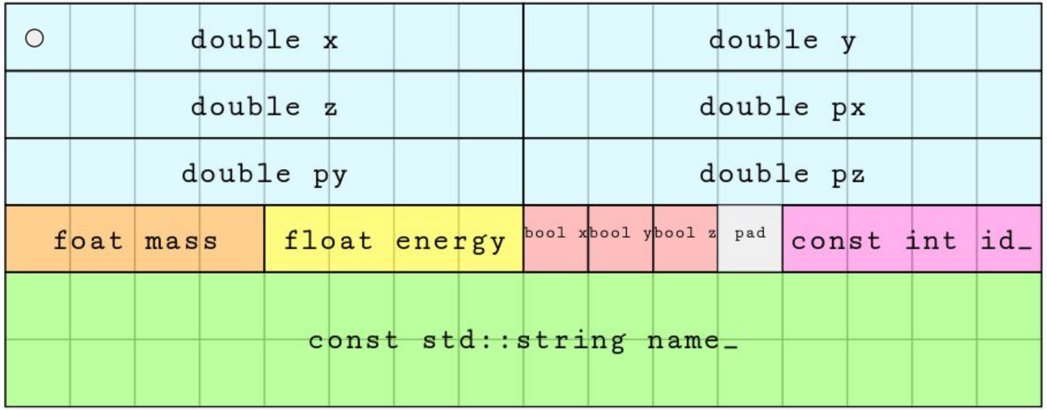


Memory Access - Data Structures

- The way you access memory is not only driven by the algorithm, but it strongly depends on how you designed your datastructure
- Let’s investigate our GoodParticle datastructure

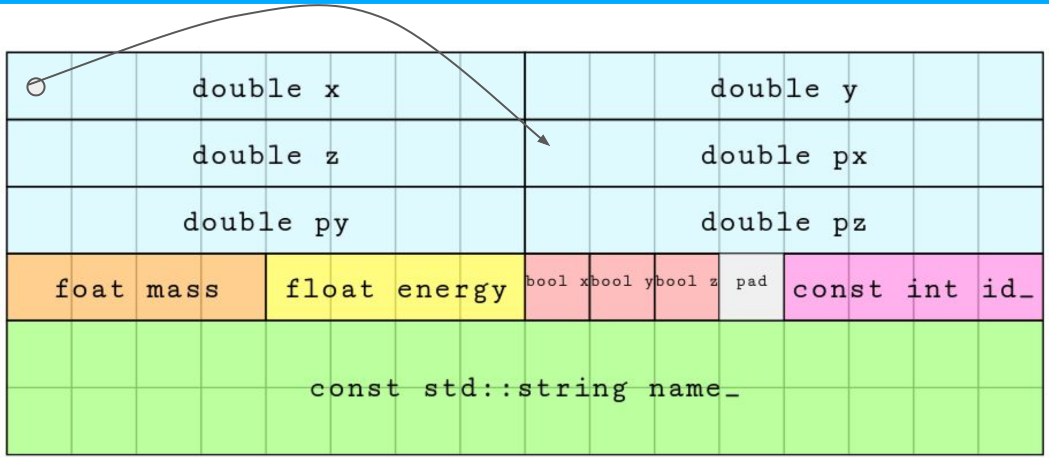
- Write a function to initialize a collection of N GoodParticles
  - Assign some value to each member of GoodParticle
  - Pick a x_max value
  - And a time value t
- Write a function that takes as input the collection, and x_max
- Iterate over the elements of this collection and for each element:
  - Update the position x → x = x + px / mass * t
  - If x < 0 or x > x_max → set hit_x to true
    - Else, set it to false and change the sign of px

GoodParticle memory access



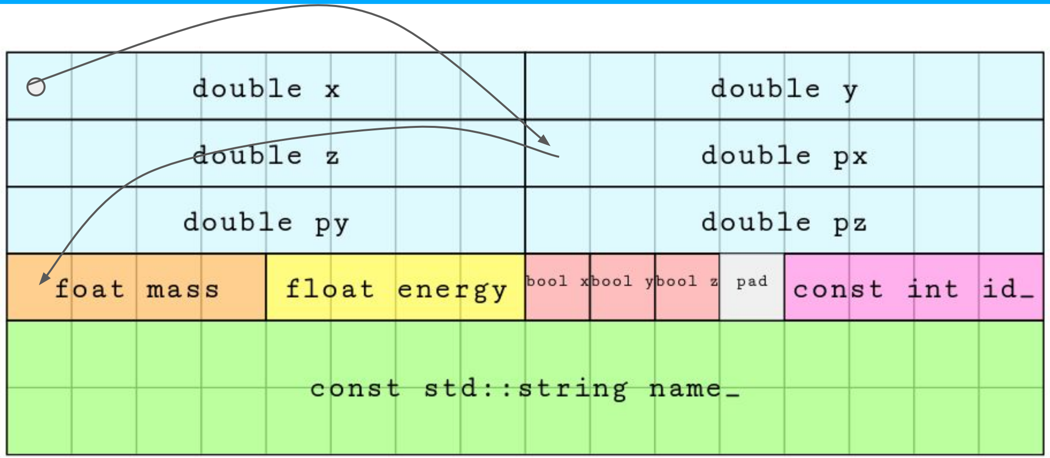
x +=

GoodParticle memory access



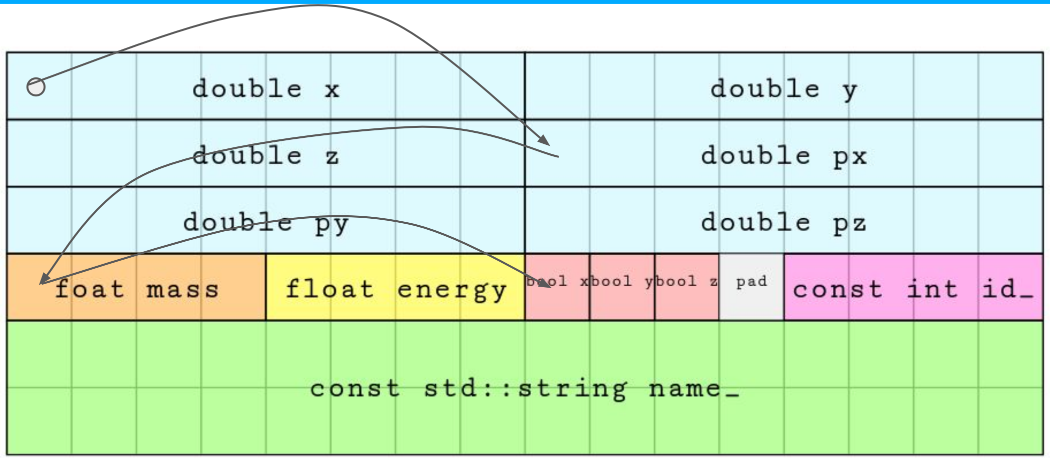
x += px

GoodParticle memory access



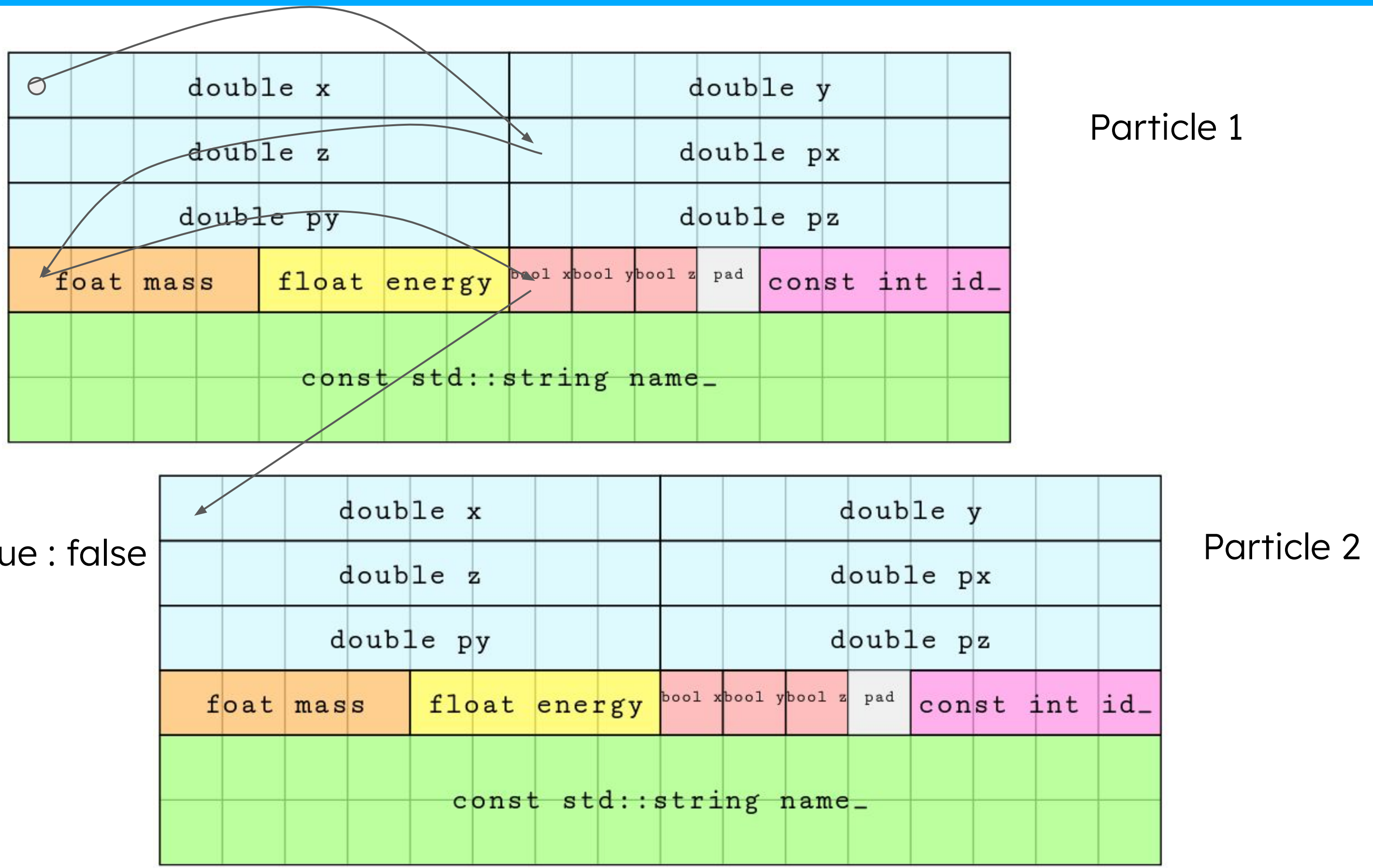
x += px/m * t

GoodParticle memory access



p.x += p.px/p.m * t
p.hit_x = statement? true : false

GoodParticle memory access



p.x += p.px/p.m * t
p.hit_x = statement? true : false

Next iteration

GoodParticle memory access

- Our problem needs only some members of our class GoodParticle
    - We are paying the price of loading the full object for accessing its members
    - `sizeof(GoodParticle) = 96bytes`
    - `sizeof(double`$_{x}$`) + sizeof(double`$_{px}$`) + sizeof(double`$_{hit\_x}$`) + sizeof(float`$_{mass}$`) = 21bytes`
        - We are using only 22% of what we are reading!
- Our `std::vector<GoodParticle>` is commonly called Array of Struct
    - Very common dastracture coming from Object Oriented Programming (OOP)
        - Self contained objects
            - Bad cache locality and bad memory bandwidth
        - Commonly used because it easy to represent the reality
            - Not so good for manipulating data in some scenario
- In principe we would like to have a data structure that allow us to use only what we need in a specific piece of code

Array of Structs vs Struct of Arrays

```
struct Particle {
    double x;
    double y;
    double z;
    …
};

std::vector<Particle> particles;
```

- All data fields for each element are stored together in a contiguous block of memory.
- Cache locality might be loss if not all the elements are used

```
struct ParticleSoA {
    std::vector<double> x;
    std::vector<double> y;
    std::vector<double> z;
    …
};

ParticleSoA particles;
```

- Each data field of all elements is stored in separate arrays.
- This layout is beneficial when you need to perform operations on some fields for all elements concurrently

AoS vs SoA

- Take the last exercise
  - Implement an SoA version of GoodParticle
  - Add two more functions, one for initializing the SoA collection and one to perform the operation previously discussed
- Try to time it
  - Try to use compiler optimization (-O1 -O2 -O3)
  - What happens?
- What memory access pattern are we using now?
- Is your data structure interface that different?

AoS vs SoA

- Sequential access pattern on each member of our object!
- Use only what you need
  - You can pass to your function only the members you are going to use

```
int N = 100;
std::vector<GoodParticleAoS> particles(N);
```

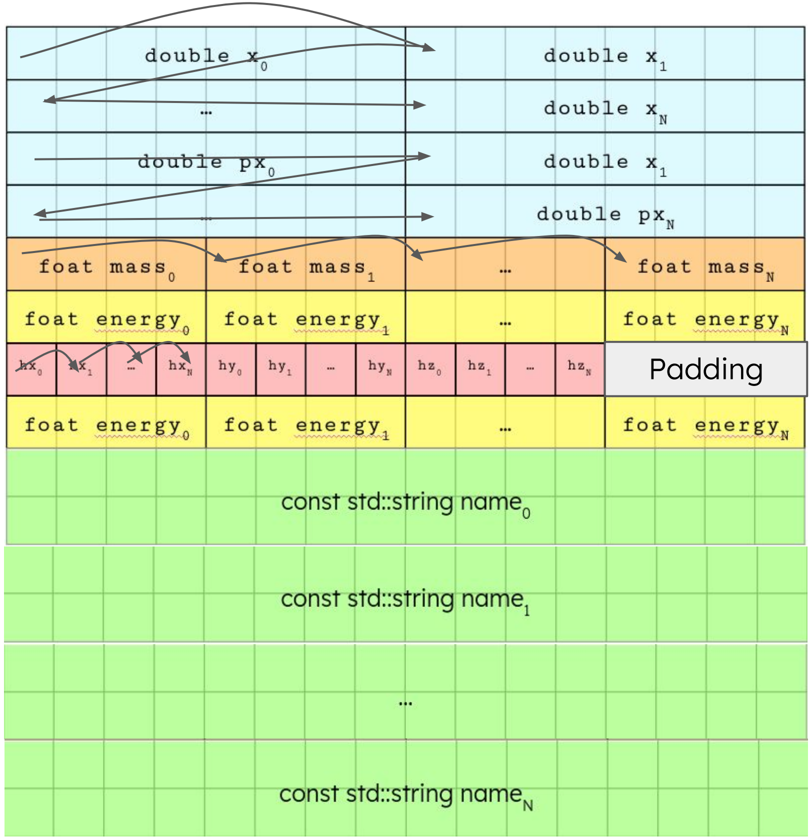
96 bytes * 100 = 9600 bytes

9600 bytes / 64 bytes/cacheline = **150 cache lines**

```
ParticleSoA particles(N);
```

21 bytes * 100 = 2100 bytes

2100 / 64 bytes/cacheline = **33 cache lines!**



More on SoA

- So far our SoA uses std::vector, which is useful to be able to resize our datastructure
- However, resizing is quite expensive
- Better to have fixed sized SoA
  - If you don’t know your exact size, better to put a Max Value
    - Knowing the size (and alignment) at compile time helps the compiler to optimize your code
      - Especially true for vectorization!
- Moreover, you can use single memory buffers to allocate and deallocate memory in one go, or to transfer it to accelerators
  - And you could also reuse the same memory!

Exercise

- Modify your ParticleSoA struct such that:
  - Contains a single memory buffer and a single size
  - Contains M pointers pointing to the beginning of each “column”
  - Explicit constructor that takes the number of particle you want to allocate
    - Allocates the needed memory with a single operation
    - Set each pointer to the beginning of the column
    - Remember alignment!

```
g++ -Wall -Wextra -fsanitize=address your_program.cc
```

To check if gcc is happy with your alignment!

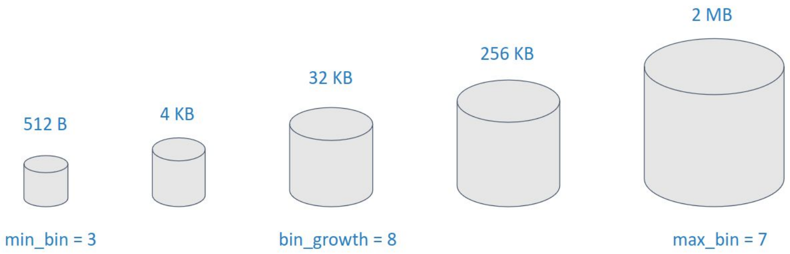
More exercises: Caching Allocator

- Allocating and deallocating can be very expensive
- We can try to reduce the impact of the allocations and deallocation by reusing some allocated memory

- Write a class representing an allocator
  - Should have an allocate(), deallocate() and free() methods
  - Let’s take inspiration from the CUB caching allocator
  - Next slide for more details

More exercises: Caching Allocator

- Idea: reuse memory already preallocated but not used
- Let's decide to only allocate memory in fixed size blocks

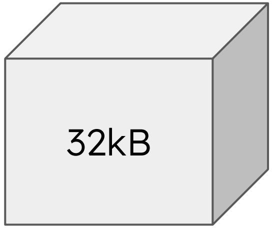


- Everytime I ask for some memory the allocator should decide the minimum block it has to allocate.
  - For example if I ask for 24kB of memory it would allocate 32kB
- Once the memory is not used anymore, we don't release the memory, but instead we keep the memory in a pool
  - If another allocation fits this 32kB of memory, the same block will be reused
  - Otherwise, we create another block

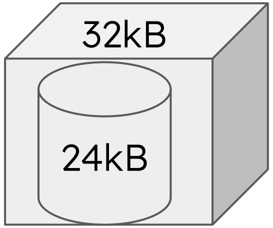
More exercises: Caching Allocator

Ask allocation

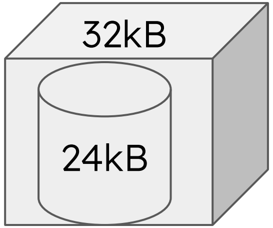
Allocate a big enough block



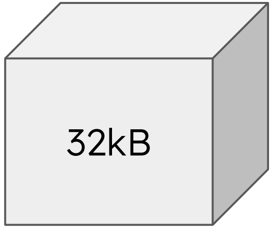
Assign block for request allocation



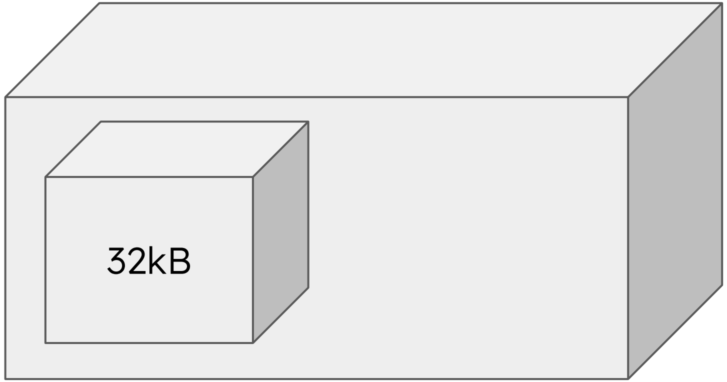
Ask for deallocation



deallocating

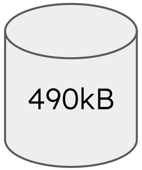


Caching allocated block

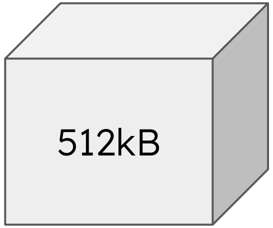


More exercises: Caching Allocator

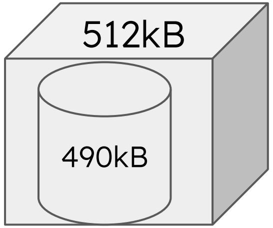
Ask for another allocation



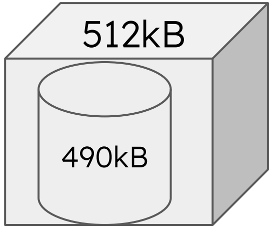
Allocate a big enough block



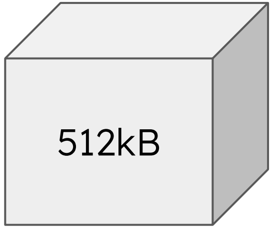
Assign block for request allocation



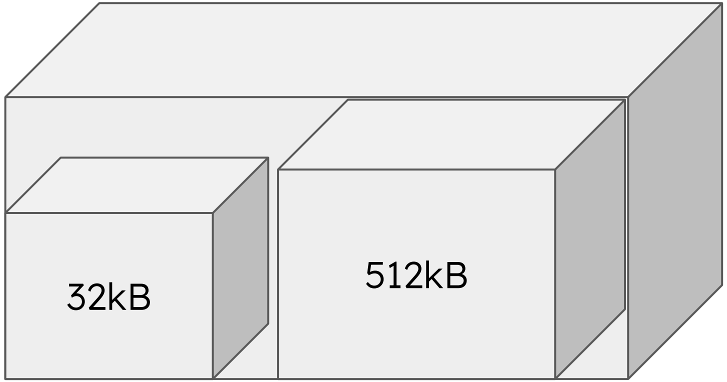
Ask for de allocation



deallocating



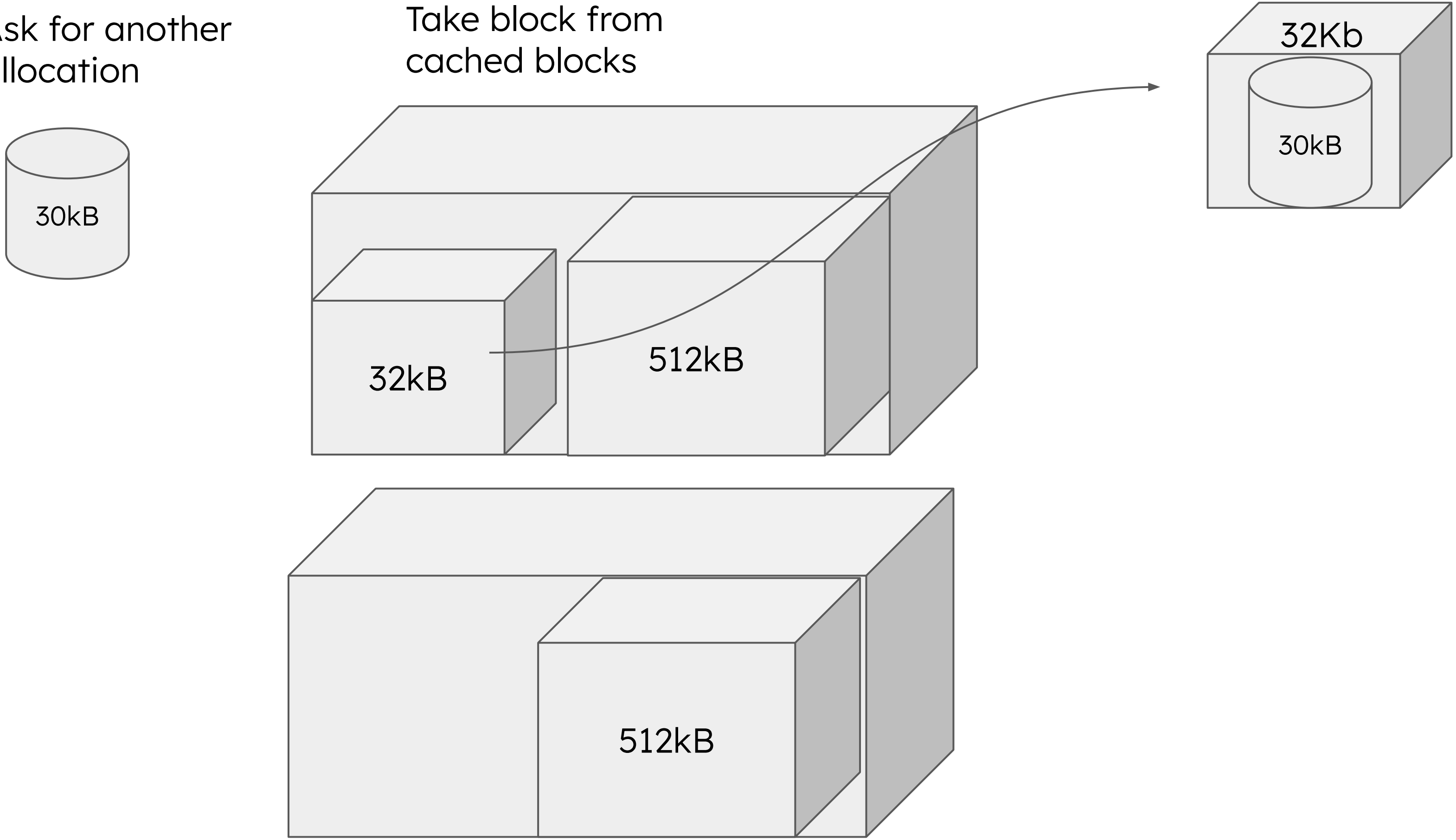
Caching allocated block



More exercises: Caching Allocator

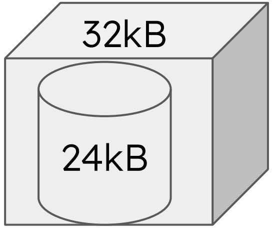
Ask for another allocation

Take block from cached blocks

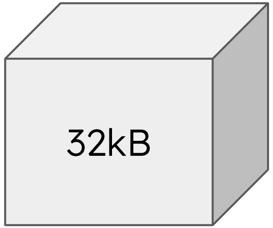


More exercises: Caching Allocator

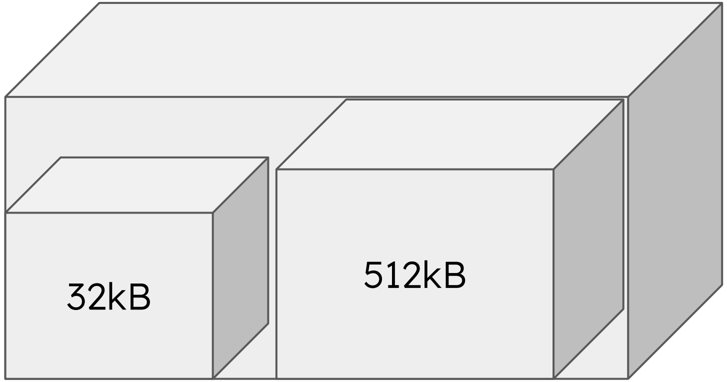
Ask for deallocation



deallocating

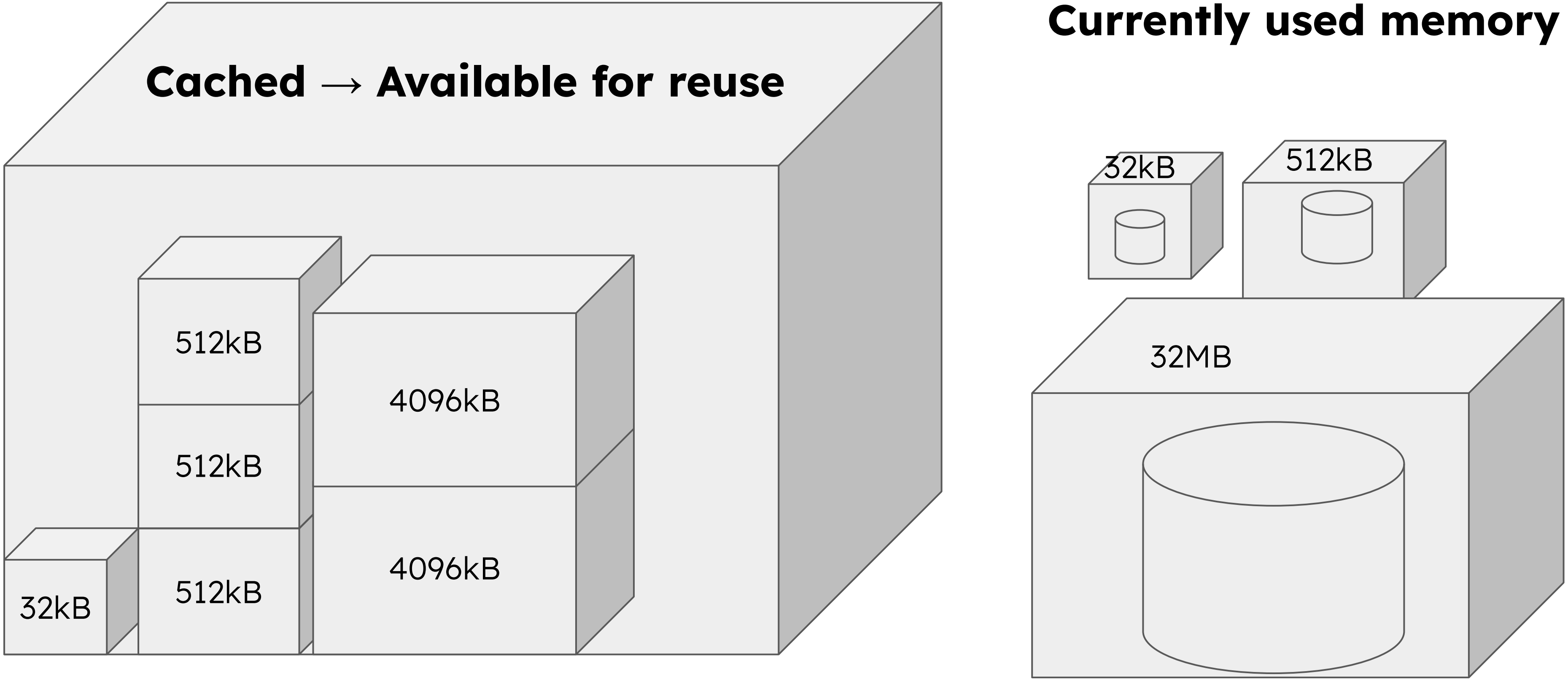


Caching allocated block



More exercises: Caching Allocator

- Possible scenario after some iterations



More exercises: Caching Allocator

- At this point I hope the illustration helped …

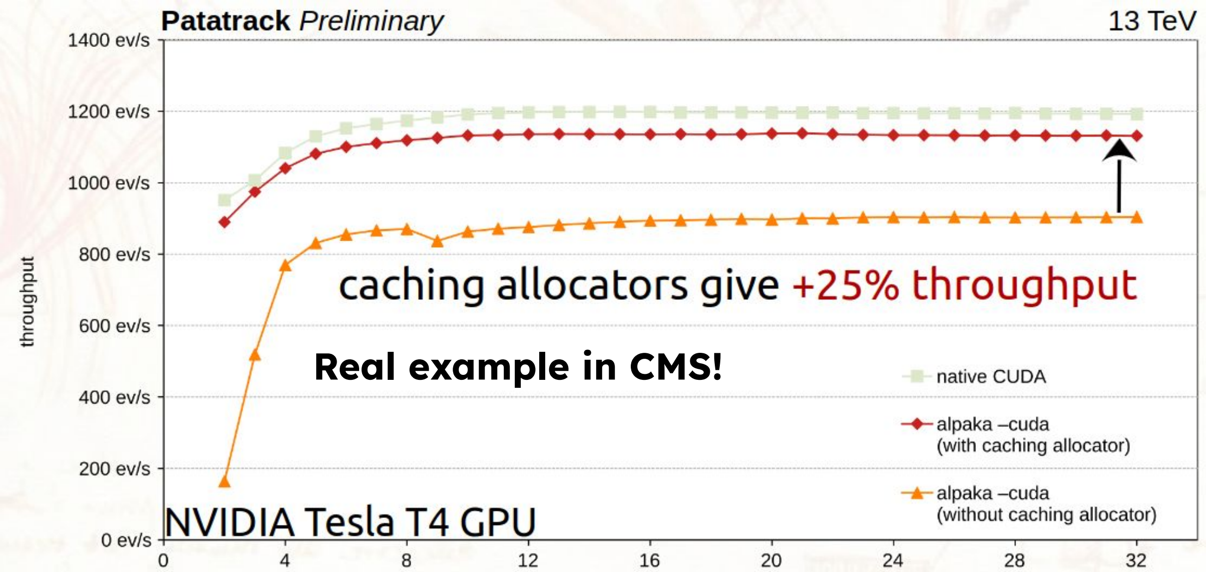
- Write an allocator that allocates blocks of memory fitting the requested size (blocks of memory of power of 2)
  - The allocator should have a min_number_bin and max_number_bin, max allocation size
  - Bin growth (8^bin_number)
  - If requested allocation is bigger than max_number_bin, allocate space normally
  - If requested size is bigger than max_allocation_size, return bad alloc
  - Remember alignment!
  - Use your allocator to allocate members in your ParticleSoA structure!
  - Try to measure performance!

Caching Allocator - Bonus

- At some point you will know how to deal with multi-threading using TBB
    - That means you will have to deal with race conditions!
- Can you make your allocator thread-safe?
- But possibilities are even more now, for example you can also decide to have an allocator for each thread or for each group of threads!

Caching Allocator - Bonus Bonus

- Stuff becomes more and more complex … now you have a GPU and you are the guru of GPU programming
  - You can manage both CPU and GPU memory with allocators!
  - I am not going to provide a solution for this exercise, but in case you are eager to try, you can have a look at the caching allocator used by CMS



From A.Bocci - ACAT2022

Memory Fragmentation

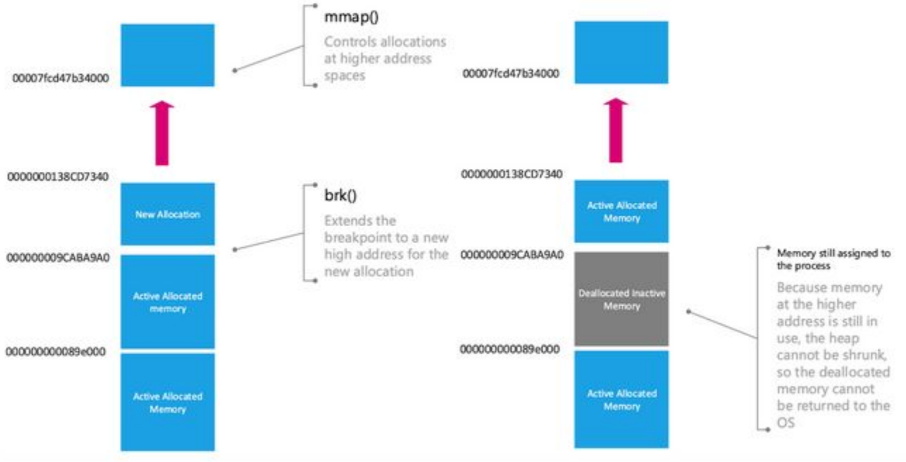
- **UNIX system uses the glibc memory allocator**

Allocate 1kB, 4kB, 2kB

Deallocate 1kB, 4kB

Allocate 2kB

Allocate 4kB → Unable



4kB are available, but not of contiguous memory
→ Memory Fragmentation

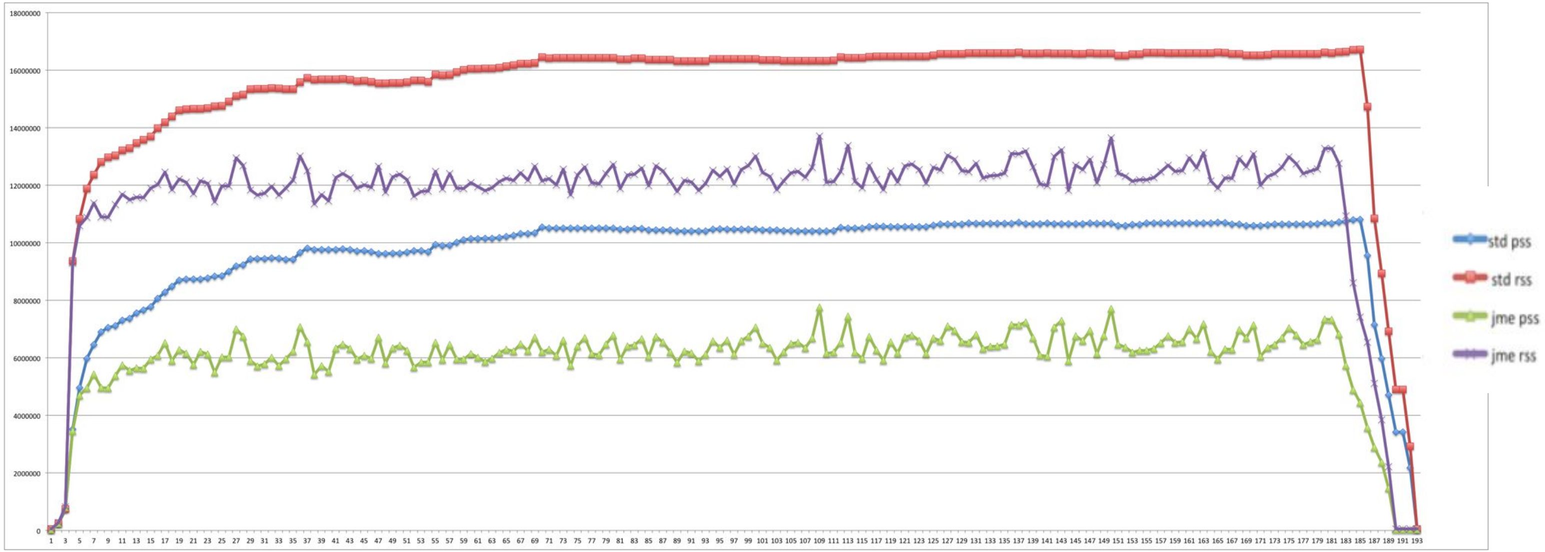
Jemalloc and TCMalloc

If your program allocates and deallocates objects with different life times, you get memory fragmentation and the process might not be able to return the memory to the OS

- Alternative allocators
  - Might give you better performance and reduce memory fragmentation
    - But detailed studies are necessary on the full application
- Jemalloc
  - Used by Mozilla Firefox, Facebook, …
  - Tries to avoid memory fragmentation
- TCMalloc
  - Developed by Google
  - Fast C implementation of malloc and new, multithreaded

Jemalloc on a real example

- Real scenario: CMS Software, multithreaded software
  - Multi-threading brings to even more memory fragmentation
- Jemalloc manages to reduce the peak memory, and as secondary effects reduces also the processing time!



Take Away Message

- Memory is what keeps you away from running code efficiently
- Keep memory always in mind when you are developing your software
- Remember to understand your hardware and map what you are programming on it
- Investigate your data before developing your data structure and try to understand the memory footprint and how to better access the memory
- Profile profile profile
    - perf, **valgrind**, intel VTune

Reference

- Thanks Andrea Bocci for all the inputs and help in preparing the lecture!

- Reducing memory footprint using jemalloc
  - https://twiki.cern.ch/twiki/bin/view/LCG/VIJemalloc
- What Every Programmer Should Know About Memory
  - https://akkadia.org/drepper/cpumemory.pdf
- What Programmers Should Know About Memory Allocation - S. Al Bahra, H. Sowa, P. Khuong - CppCon 2019
  - https://www.youtube.com/watch?v=gYfd25Bdmws&t
- CppCon 2014: Mike Acton "Data-Oriented Design and C++"
  - https://www.youtube.com/watch?v=rX0ItVEVjHc&t=2838s
- jemalloc

BONUS

Jemalloc example

- Here's a program with the aim of fragmenting the memory from Zac blog post
    - https://gist.github.com/ZacAttack/8c67b998c90afdb19c715dfe327112d2#file-heap-fragmentor-cpp
- Compile it and try to look at the