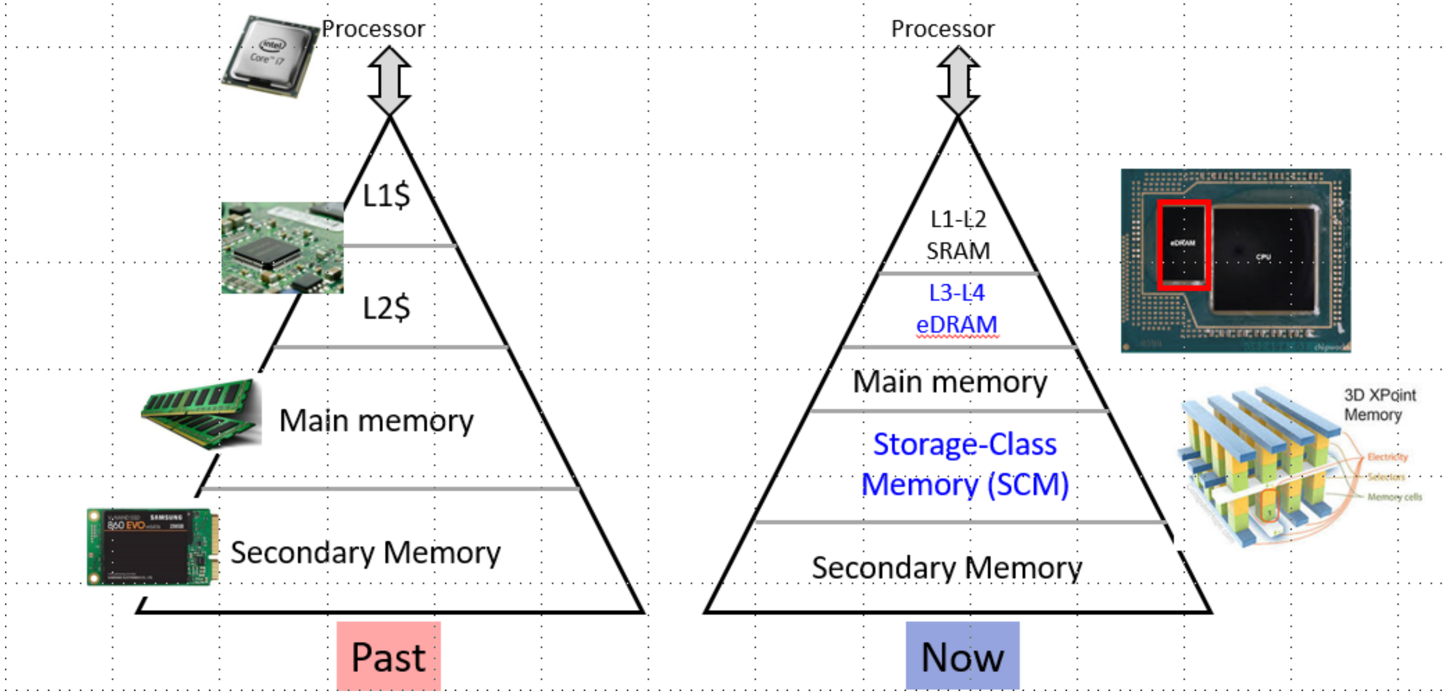
内存管理
Memory Management

现代C++基础
Modern C++ Basics

Jiaming Liang, undergraduate from Peking University

- Low-level Memory Management
- Smart Pointers
- Allocators
  - PMR

- The real structure of memory is quite complex…



Credit: Jie Zhang @ PKU Arch.

- However, OS has abstracted them as *virtual memory* by page table, so in most cases users can view memory as a large contiguous array.
  - When such abstraction causes performance bottleneck, programmers need to dig into that further;
    - C++ also has some utilities to solve some common problems.

Memory Management

Low-level Memory Management

Memory Management

- Low-level Memory Management
  - Object layout
  - `operator new/delete` in detail

Object layout

- Object will occupy a contiguous segment of memory that:
  - Starts at some address that matches some **alignment**;
  - And ends at some address that matches some **size**.
- A *complete object* may have many *subobjects* as members or elements (e.g. array or `class`).
  - `sizeof` reflects size of the type when it forms a complete object, which is always `>0`.
  - For example:

```
struct Empty {};
static_assert(sizeof(Empty) > 0);
```

- In most cases, subobjects just occupy memory in the same way:

```
struct Empty {};
struct NonEmpty
{
    int a;
    Empty e;
};
// Padding may exist so we use '>='
static_assert(sizeof(NonEmpty) >= sizeof(int) + sizeof(Empty));
```

Object layout

- However, some subobjects as class member can have 0 size…
  - Formally called "potentially-overlapping objects".

1. For a class, if it fulfills:
   - No non-static data members;
   - No virtual methods or virtual base class;
   - It's a base class.

   Then it's **allowed** to have 0 size.

   Moreover, it's **forced** to have 0 size if
   - The derived class is a standard-layout class.

```cpp
struct Empty {};
struct NonEmpty : Empty
{ // Standard-layout
    int a;
};
static_assert(sizeof(NonEmpty) == sizeof(int));
```

- Also called "*Empty Base (Class) Optimization*" (EBO/EBCO).

Object layout

- So now we can understand `static_cast` / `reinterpret_cast`…

- For `static_cast`, besides inheritance-related pointer conversion, it also processes `void*`.
  - You can convert any **object** pointer to `void*` (this is also implicit conversion).
  - You can also convert explicitly `void*` to any object pointer.
  - **BUT**, this requires the underlying object of type `U` and the converted pointer `T*` (ignoring cv) to have the relationship (called *pointer-interconvertible*) as:
    - `T == U`.
    - `U` is a union type, while `T` is type of its member (though using it still needs this member to be in its lifetime).
    - `U` is standard-layout, while `T` is type of its **first** member or its base class.
    - Or all vice versa/transitivity (i.e. you can swap `T` and `U` above; after all, "inter").

In lecture 5, *Lifetime & Type Safety*.

FYI, this can be checked by [`std::is_pointer_interconvertible_with_class`]() and [`std::is_pointer_interconvertible_base_of`]() since C++20.

Object layout

- Empty base will be collapsed so conversion is safe.

```cpp
struct Empty {};
struct NonEmpty : Empty
{ // standard-layout
    int a;
};

NonEmpty obj;
// ptr points to the base class of obj.
Empty* ptr = reinterpret_cast<Empty*>(&obj);
// ptr2 points to obj.a.
int* ptr2 = reinterpret_cast<int*>(&obj);

static_assert(std::is_pointer_interconvertible_with_class(&NonEmpty::a));
static_assert(std::is_pointer_interconvertible_base_of_v<Empty, NonEmpty>);
```

- A class is said to be **standard-layout**, if:
  - All non-static data members have the same accessibility and are also standard-layout.
    - This is because the layout of members that have different accessibility are unspecified (before C++23); e.g. as the sequence of declaration or first all public members and second all private members.
  - No virtual functions or non-standard-layout base classes.
  - The base class is not the type of the first member data.
  - There is at most one class in the inheritance hierarchy that has non-static member variable.
    - That's because layout of inheritance is not regulated.

Object layout

[1]: Strictly speaking, it should be "similar types", e.g. adding cv-qualifiers is allowed. See [conv.qual] for details.
[2]: Except for potentially non-unique object like string literals.

2. Since C++20, for a member subobject that is marked with attribute `[[no_unique_address]]`, it's **allowed** to have 0 size.
   - Particularly, msvc will ignore this attribute for backward compatibility; instead, it respects `[[msvc::no_unique_address]]`.
   - For example:

```cpp
struct Y
{
    int i;
    [[no_unique_address]] Empty e;
};
```

In gcc/msvc/clang, `sizeof(Y) == 4`.

   - Note: C++ regulates two objects of the same type[1] must have **distinct addresses**[2].
     - For example:

```cpp
struct Z
{
    char c;
    // e1 and e2 cannot share the same address because they have the
    // same type, even though they are marked with [[no_unique_address]].
    // However, either may share address with 'c'.
    [[no_unique_address]] Empty e1, e2;
};
```

All three compilers make `sizeof(Y) == 2`.

Object layout

```
struct W
{
    char c[2];
    // e1 and e2 cannot have the same address, but one of
    // them can share with c[0] and the other with c[1]:
    [[no_unique_address]] Empty e1, e2;
};
```

- Theoretically, this can be optimized as `sizeof(W) == 2`; however, all three compilers make `sizeof(W) == 3`.

- And again, we can understand in standard layout…

> - A class is said to be **standard-layout**, if:
>   - All non-static data members have the same accessibility and are also standard-layout.
>   - This is because the layout of members that have different accessibility are unspecified (before C++23); e.g. as the sequence of declaration or first all public members and second all private members.
>   - No virtual functions or non-standard-layout base classes.
>   - The base class is not the type of the first member data.
>   - There is at most one class in the inheritance hierarchy that has non-static member variable.
>   - That's because layout of inheritance is not regulated.

Object layout

- Now EBCO doesn't guarantee to happen:

```
struct Empty {};
struct NonEmpty : Empty
{ // Not standard-layout
    Empty e;
    int a;
};

NonEmpty obj;
// ptr doesn't necessarily point to e.
Empty* ptr = reinterpret_cast<Empty*>(&obj);
```

  - In ABI, base class may be put first;
  - As subject of base class must be distinguished from the first member, then base class may be not really "empty".
  - And a non-empty base leads to non-standard-layout*.

*: there may be some defects in current definitions. See SO question.

Layout Compatible*

- **This part is optional**.
- Finally we fix our claim before:

> - Similarly, for union type, it's **illegal** to access an object that's not in its lifetime (it's only allowed in C)!
>   - Here `u.a` is in its lifetime, while `u.b` is not.
>   - You should use `std::memcpy` or `std::bit_cast` since C++20 to make them bitwise equivalent.

- Rigorously, when types have *common initial sequence*, it's legal to access out of lifetime:

```cpp
struct T1 { int a, b; };
struct T2 { int c; double d; };
union U { T1 t1; T2 t2; };
int f() {
  U u = { { 1, 2 } };   // active member is t1
  return u.t2.c;        // OK, as if u.t1.a were nominated
}
```

Layout Compatible*

- Formally, we say two types are layout compatible if:
  - Naïve cases:
    - They are of the same type, ignoring cv qualifier; or,
    - They are enumerations with the same underlying integer type.
  - Otherwise,
    1. They are both standard-layout; and,
    2. Their common initial sequence covers all members.
- where common initial sequence means the longest sequence of non-static data members and bit-fields in declaration order that:
  1. corresponding entities are layout-compatible; and,
  2. corresponding entities have the same alignment requirements; and,
  3. either both entities are bit-fields with the same width or neither is a bit-field.

Layout Compatible*

- For example:

```cpp
struct A { int a; char b; };
struct B { const int b1; volatile char b2; };
// A and B's common initial sequence is A.a, A.b and B.b1, B.b2

struct C { int c; unsigned : 0; char b; };
// A and C's common initial sequence is A.a and C.c

struct D { int d; char b : 4; };
// A and D's common initial sequence is A.a and D.d

struct E { unsigned int e; char b; };
// A and E's common initial sequence is empty
```

A and B are layout-compatible.

- Since C++20, you can use `std::is_layout_compatible`* and `std::is_corresponding_member` to check it.

```cpp
struct T1 { int a, b; };
struct T2 { int c; double d; };
struct T3 { int a, b; };

static_assert(std::is_corresponding_member(&T1::a, &T2::c));
static_assert(!std::is_corresponding_member(&T1::b, &T2::d));
static_assert(std::is_layout_compatible_v<T1, T3>);
```

*: Strictly speaking, `std::is_layout_compatible` will tolerate non-struct-type, while the standard only regulates struct-type.

Alignment

- To maximize efficiency, data should be aligned properly.
  - For example, on some platform:

```
// 00 // long long, char, int can live here
// 01 // char
// 02 // char
// 03 // char
// 04 // char, int can live here
// 05 // char
// 06 // char
// 07 // char
// 08 // long long, char, int can live here
```

- In C++, it can be checked by `alignof(T)`;
  - Platform-dependent, return `std::size_t`, quite like `sizeof`.

```
std::println("{} {} {}", alignof(char), alignof(int), alignof(long long));
```

```
Program returned: 0
  1 4 8
```

*Or using type traits `std::alignment_of`.

Alignment

- When wrapping data in class, every object will be aligned to its own alignment, leading to padding.
  - For example:

```
struct S {  // begins at:
    char a; // 0
    // 3 padding bytes to match alignof(i)
    int i;  // 4
    char b; // 8
    // 3 padding bytes to match alignof(j)
    int j;  // 12
    char c; // 16
    // 7 padding bytes to match alignof(l)
    long long l; // 24
    // Possible padding bytes to match alignof(S)
}; // sizeof(S): at least 32.
```

Each element in C array should be suitably aligned, thus `sizeof(X)` must be multiple of `alignof(X)`.

Alignment

- Naturally, all scalar types will have alignment not greater than `alignof(std::max_align_t)` (in `<cstddef>`).
  - And allocation will align to this alignment by default.
- However, sometimes you may want *over-aligned* data.
  - Then you can use `alignas(N)` to make alignment `N`.
    - Ignored when `N == 0`, compile error if `N` is not power of 2.
  - For example, to match OpenGL uniform layout:

```
struct BasicParams
{
    alignas(16) glm::vec3 cameraPos;
    int randOffset;

    alignas(16) glm::vec3 cameraForward, cameraRight, cameraUp;
    float g;
```

These three members are all aligned to 16.

Alignment

- Note 1: you can also use `alignas(T)` to have alignment same as `T`.
- Note 2: when using multiple `alignas`, the largest one will be selected.
  - So our previous code segment can be rewritten:

```
alignas(std::max(alignof(float), alignof(int))) std::byte arr[20];
float* ptr = reinterpret_cast<float*>(arr);
*ptr = 1.0f;
int* ptr2 = reinterpret_cast<int*>(arr);
// std::cout << *ptr2; // -> illegal
```

```
alignas(float) alignas(int) std::byte arr[20];
```

- Note 3: you can do pack expansion in `alignas`, which is same as `alignas(arg1) alignas(arg2) … alignas(argN)`.
  - i.e. select the largest alignment among N arguments.

Alignment

```
alignas(1) int a = 2;
```
❌

- Note 4: over-align only: if `alignas` is weaker than its natural alignment (i.e. alignment without `alignas`), compile error.
  - Some compilers will ignore or only warn.
- Note 5: alignment is NOT part of the type, so you cannot alias it in `using` or `typedef`.

```
struct C {
    long long x;
    int y;
};

using T = alignas(16) C;
```
❌

Attributes are added after `struct`.

```
struct alignas(16) C {
    long long x;
    int y;
};
```
✅

```
struct C {
    long long x;
    int y;
};

struct D
{
    alignas(16) C c;
};
```
✅

- Note 6: function parameter and exception parameter are not allowed to use `alignas`.

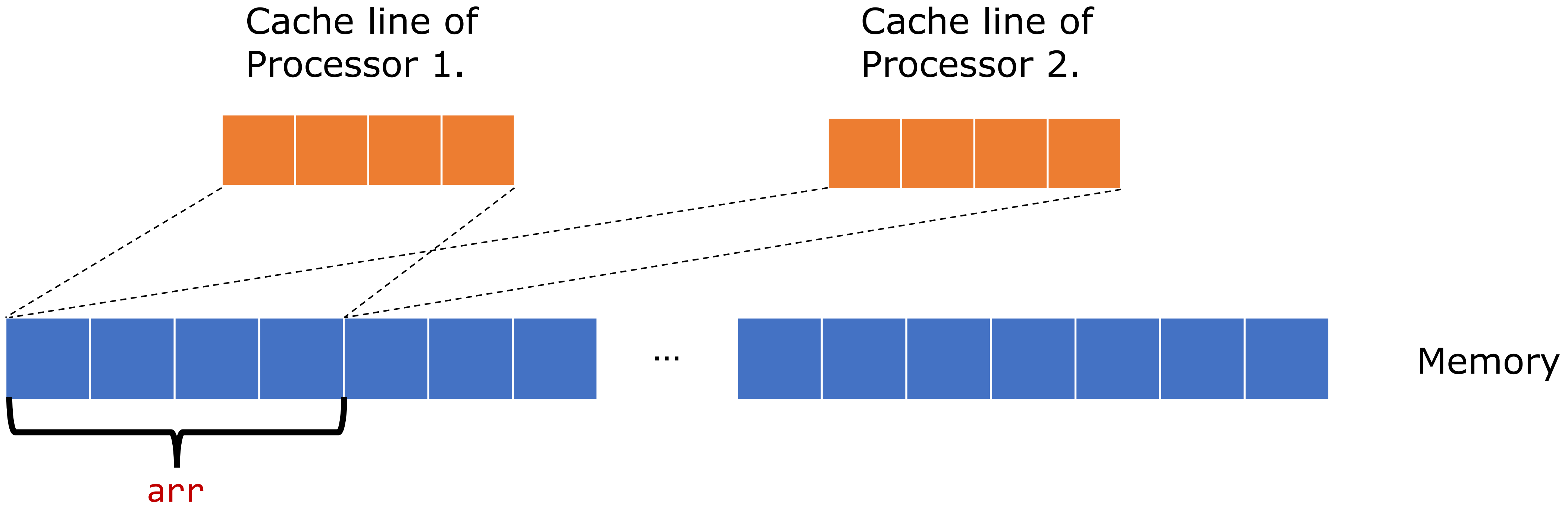
False Sharing

- Practical example: false sharing
  - From abstraction, when different threads operate on different data, parallelism will be maximized since no lock is needed.

```cpp
// Here to prevent compiler optimization to collapse
// We use atomic<int> instead of int.
std::atomic<int> arr[4];
void Work(int idx) { arr[idx]++; }
```

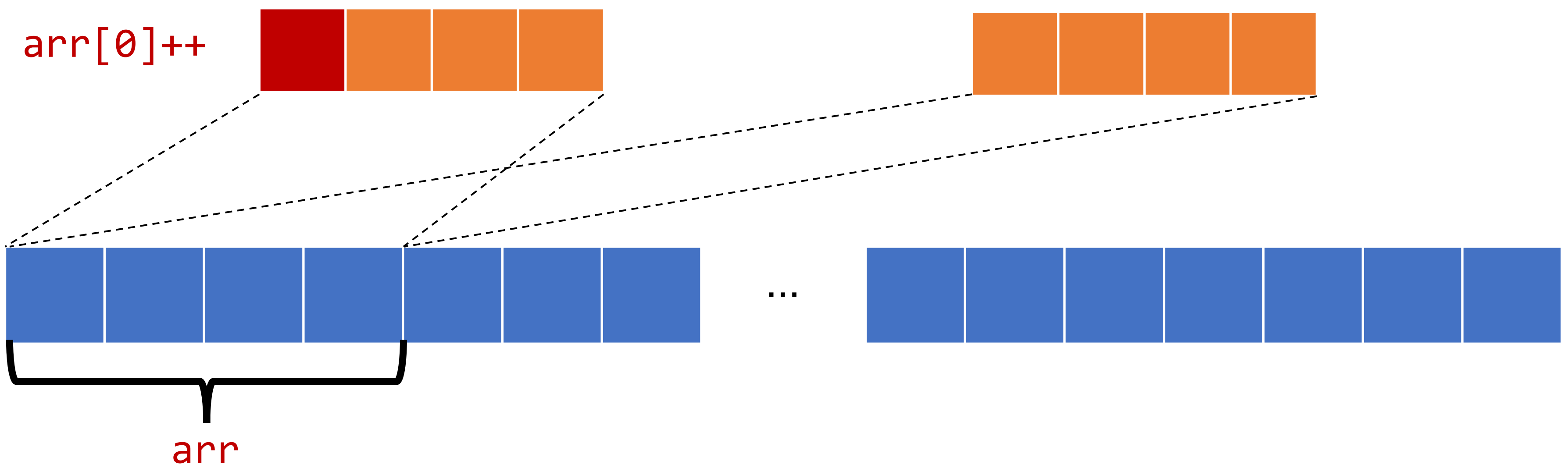
- However, due to limitation of computer architecture, such abstraction is wrong…
  - Cache on different processors has to obey coherence protocol like MESI.
  - To put it simply, when write happens on a cache line, it’ll inform other processors whose cache also own this line to make it invalid.
  - And invalid line needs to be reloaded, leading to inefficiency.

False Sharing

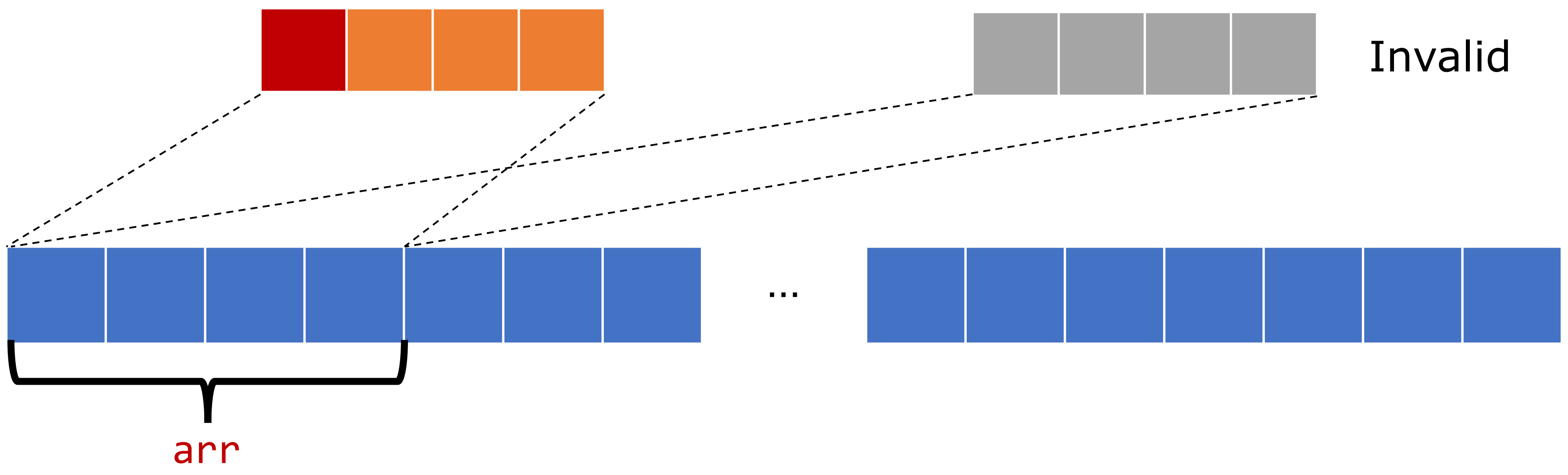


Illustrative animation for false sharing.
(Details may vary for different architectures)

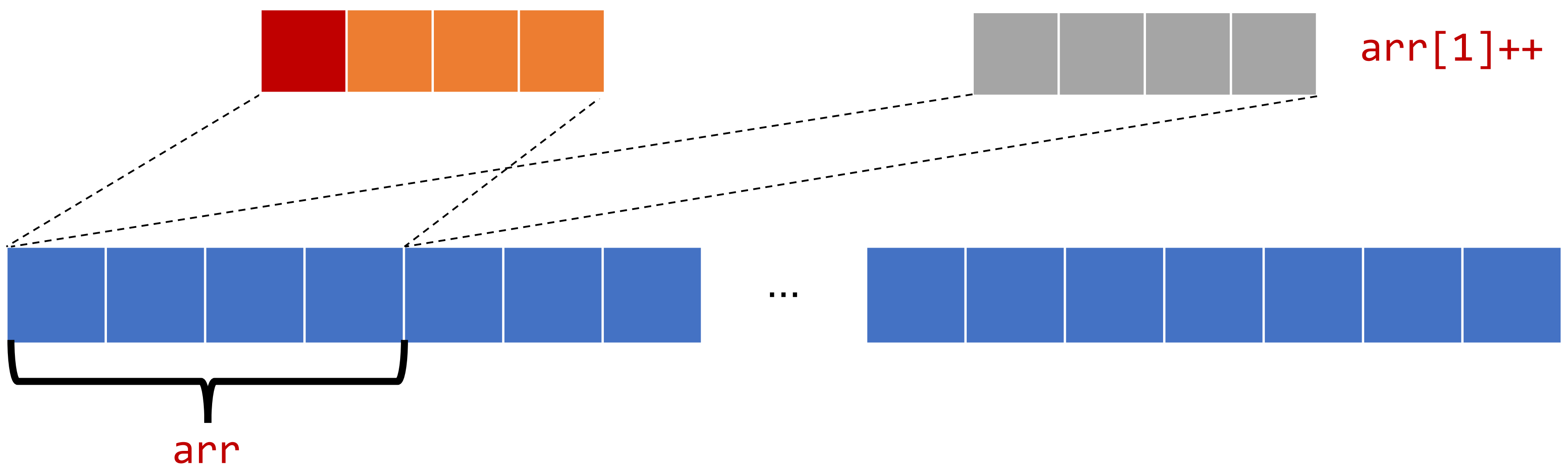
False Sharing



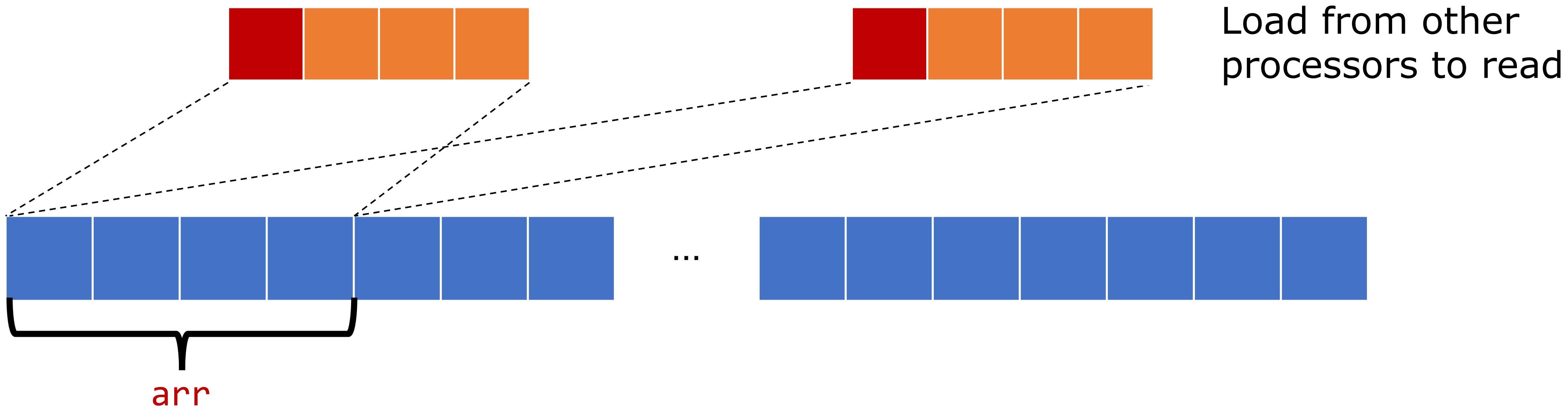
False Sharing



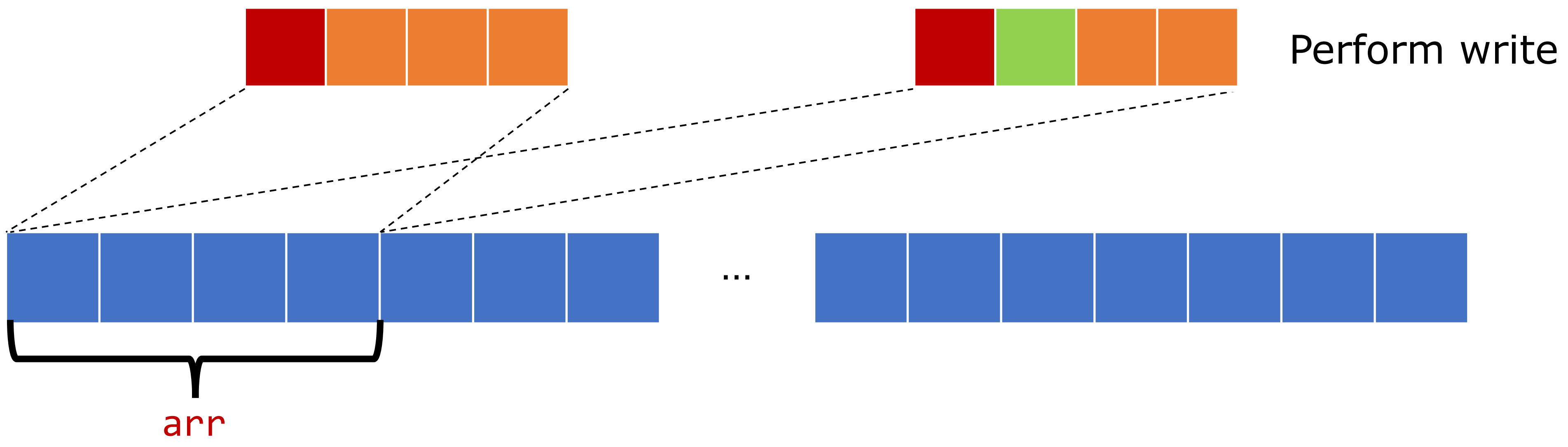
False Sharing



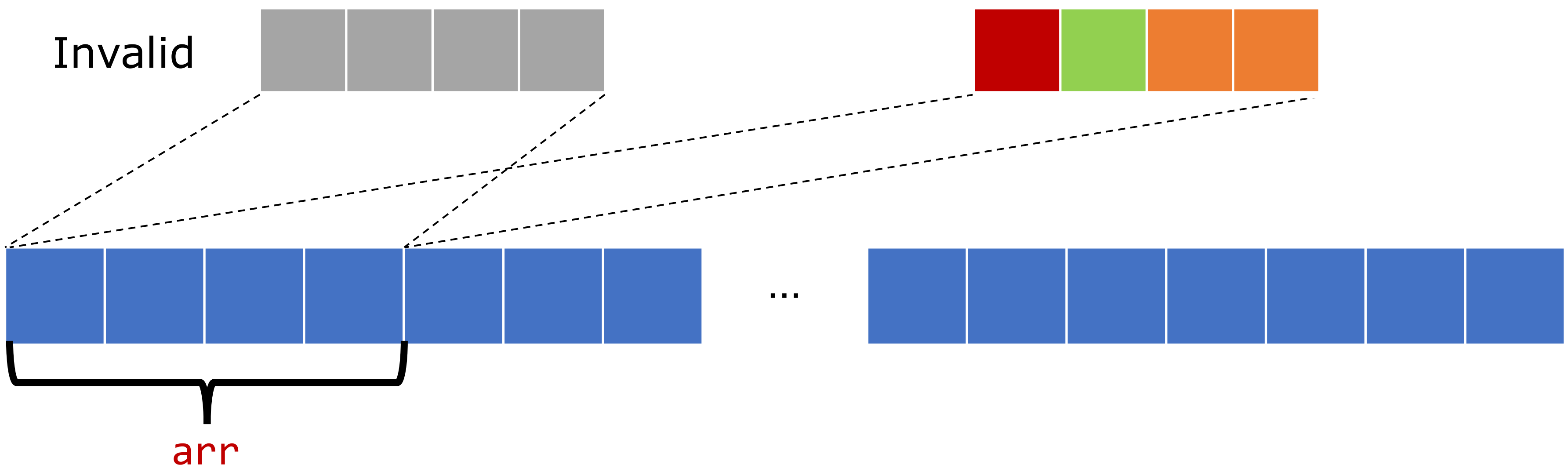
False Sharing



False Sharing



False Sharing



False Sharing

- So when writes in different threads are on the same cache line, every write will happen exclusively as if having a lock.
  - This leads to false parallelism, degrading the performance.
- Solution: make threads access data on different cache lines!
  - C++17 provides constant `std::hardware_destructive_interference_size` in `<new>`.
  - For example:

```
struct OveralignedInt
{
    alignas(std::hardware_destructive_interference_size) std::atomic<int> elem;
};
// alignas(N) T arr[4] won't align every element, but just arr[0].
// To align every element, we need to wrap inside a struct.
OveralignedInt arr[4];
void Work(int idx) { arr[idx].elem++; }
```

False Sharing

- On the other hand, for a single thread, we hope accessed data to lie on the same cache line to minimize pollution.
  - For example:

Improperly aligned,
use two cache lines.

Properly aligned,
use single cache line.

  - To force data to lie on the same cache line, we can align the head as cache line head.
  - C++17 thus introduces `std::hardware_constructive_interference_size` for that.

False Sharing

- For example:

```
struct alignas(hardware_constructive_interference_size)
OneCacheLiner // occupies one cache line
{
    std::atomic_uint64_t x{};
    std::atomic_uint64_t y{};
}
oneCacheLiner;

struct TwoCacheLiner // occupies two cache lines
{
    alignas(hardware_destructive_interference_size) std::atomic_uint64_t x{};
    alignas(hardware_destructive_interference_size) std::atomic_uint64_t y{};
}
twoCacheLiner;
```

- Question: aren't `std::hardware_destructive_interference_size` and `std::hardware_constructive_interference_size` just same as cache line size?
  - Why do we need two constants to represent them?

False Sharing

- Reason: in some architecture, destructive interference will be larger than a cache line…
  - For example, on Intel Sandy Bridge processor, it will do adjacent-line prefetching.
  - So when loading a cache line, the next cache line may or may not be substituted, leading `hardware_destructive_interference_size == 128` while `hardware_constructive_interference_size == 64`.

Supplementary

- Note 1: there exist several utilities for alignment in `<memory>`.
  1. `std::align`:

```
void* align( std::size_t alignment,
             std::size_t size,
             void*& ptr,
             std::size_t& space );
```

    - Assuming that we have a space that starts from `ptr` and has size `space`;
    - Now we want to allocate an object with `size` and `alignment` on the space;
      - Assuming that it can be allocated on `new_ptr` on `new_space` (i.e. suitably aligned).
    - So `std::align` just modifies `ptr` to `new_ptr`, `space` to `new_space`, and returns `new_ptr`.
      - If space is too small, then nothing happens and `nullptr` is returned.

Supplementary

- For example:

```
class Buffer
{
    std::vector<std::byte> buffer_;
    std::size_t size_;
    void* ptr_;

public:
    Buffer(std::size_t size) : size_{ size }
    {
        buffer.resize(size);
        ptr_ = buffer.data();
    }

    template<typename T>
    void* Alloc()
    {
        auto addr = std::align(sizeof(T), alignof(T), ptr_, size_);
        ptr_ += sizeof(T);
        size_ -= sizeof(T);
        return addr;
    }
}
```

Supplementary

2. To maximize optimization, you can inform compiler that a pointer is aligned by `std::assume_aligned<N>(ptr)` since C++20.
   - It's UB if it's not aligned to `N`, quite like `[[assume]]`.
   - Since C++26 you can also add `std::is_sufficiently_aligned<N>(ptr)` to check precondition in debug mode.
   - For example:

```cpp
void Func(int* ptr)
{
    static constexpr std::size_t alignment = 64;
    assert(std::is_sufficiently_aligned<alignment>(ptr));
    std::assume_aligned<alignment>(ptr);
    // Then compilers may do optimization based on
    // assumption of 64 alignment.
}
```

Supplementary

- Note 2: since C++17, you can use trait `std::has_unique_object_representations` to check if same value representations of two objects lead to the same object representation.
  - For example, for `float`, two `NaN` are not distinguishable but may have different bits, so the trait returns `false`.
  - Particularly, for a `struct`, when there are padding bytes, then it definitely returns `false` since they are not part of value of `struct`.

  - This trait can be used to check whether it's correct for a type to be hashed as a byte array.

Memory Management

- Low-level Memory Management
  - Object layout
  - `operator new/delete` in detail

new/delete

- To combine allocation and construction, C++ uses **new-expression** to substitute `malloc` in C.
  - Roughly speaking, it calls two different functions:
  - Allocation new, which only allocates memory (quite like `malloc`).
  - Placement new, i.e. construct the object on memory.
- And similarly, **delete-expression** has two parts:
  - Destructor, i.e. destruct the object on memory.
  - Deallocation delete, which only deallocates memory (quite like `free`).
- C++ allows you to **override** (replace) the allocation new by `operator new` and the deallocation delete by `operator delete`.

new/delete

- Thus the most basic versions like `malloc` and `free` are as below:
  - You can override them in global scope (i.e. namespace `::`).

```
void* operator new  ( std::size_t count );
void* operator new[]( std::size_t count );
void operator delete  ( void* ptr ) noexcept;
void operator delete[]( void* ptr ) noexcept;
```

- Besides, you can provide class-specific allocation & deallocation:

```
void* T::operator new  ( std::size_t count );
void* T::operator new[]( std::size_t count );
void T::operator delete  ( void* ptr );
void T::operator delete[]( void* ptr );
```

  - Which is preferred than global override, and isn't required to be `noexcept`.
  - They are always static function, even if you don't add keyword `static`.

```cpp
void* operator new(std::size_t byteCnt)
{
    std::println("Called overrided operator new, size={}", byteCnt);
    if (auto ptr = malloc(byteCnt); ptr)
        return ptr;
    throw std::bad_alloc{};
}

void operator delete(void* ptr) noexcept
{
    std::println("Called overrided operator delete, ptr={}", ptr);
    free(ptr);
}

class Base
{
public:
    static void* operator new[](std::size_t byteCnt)
    {
        std::println("INSIDE CLASS: Called overrided operator new[], size={}", byteCnt);
        if (auto ptr = malloc(byteCnt); ptr)
            return ptr;
        throw std::bad_alloc{};
    }
```

Here we don't add `static`, but it's still static function. You can't use `this` here.

```cpp
    void operator delete[](void* ptr) noexcept
    {
        std::println("INSIDE CLASS: Called overrided operator delete[], ptr={}", ptr);
        free(ptr);
    }
};
```

NOTE: before P3107 (DR23), `std::println` will use `std::string` and thus may need to use `operator new/delete`, causing infinite recursion. MS-STL has implemented this DR so it's fine to do so.

```cpp
int main()
{
    auto a = new int{ 1 };
    auto b = new Base[2];
    std::println("PtrA = {}, val = {}; PtrB = {}",
                 (void*)a, *a, (void*)b);
    delete[] b;
    delete a;
    return 0;
}
```

```
Called overrided operator new, size=4
INSIDE CLASS: Called overrided operator new[], size=2
PtrA = 0x218395c8d10, val = 1; PtrB = 0x218395c8fd0
INSIDE CLASS: Called overrided operator delete[], ptr=0x218395c8fd0
Called overrided operator delete, ptr=0x218395c8d10
```

new/delete

- However, C++ also allows you to `delete basePtr`, which will call virtual dtor.
  - Ideally, it should call `Derived::operator delete` instead of `Base::operator delete`…
  - Let's try it!

```cpp
class Base
{
public:
    static void* operator new(std::size_t byteCnt)
    {
        std::println("Base: Called overrided operator new, size={}", byteCnt);
        if (auto ptr = malloc(byteCnt); ptr)
            return ptr;
        throw std::bad_alloc{};
    }

    void operator delete(void* ptr) noexcept
    {
        std::println("Base: Called overrided operator delete, ptr={}", ptr);
        free(ptr);
    }

    virtual ~Base() { std::println("Base dtor"); }
};
```

```cpp
class Derived : public Base
{
public:
    static void* operator new(std::size_t byteCnt)
    {
        std::println("Derived: Called overrided operator new, size={}", byteCnt);
        if (auto ptr = malloc(byteCnt); ptr)
            return ptr;
        throw std::bad_alloc{};
    }

    void operator delete(void* ptr) noexcept
    {
        std::println("Derived: Called overrided operator delete, ptr={}", ptr);
        free(ptr);
    }

    virtual ~Derived() { std::println("Derived dtor"); }
};
```

new/delete

```
Base* ptr = new Derived;
delete ptr;
```

```
Derived: Called overrided operator new, size=8
Derived dtor
Base dtor
Derived: Called overrided operator delete, ptr=0x2b96ff5c320
```

- The output is like:
  - So `Derived::operator delete` is called, quite like virtual dtor!
  - But `operator delete` is `static`! How?
- Reason: compiler will generate a "deleting destructor"*.
  - That is, it will generate a new virtual function:
    - For a normal object, just use normal dtor;
    - For `delete ptr`, it will call this new function.

```
virtual void DestroyWhenDelete(void* addr)
{
    this->~Derived();
    Derived::operator delete(addr);
}
```

- With virtual dispatch, we can extract more information from the type to improve `malloc`-like version!

*: This is implementation-defined; here we use method of Itanium ABI. See [this blog]() for details.

new/delete

- Before going on, let's do some recap…
- In ICS, we've written a very basic allocation strategy:
  - Allocate memory block that's slightly larger than requested, then store block size and pointer to next block alongside it.
  - However, many metadata will never change after allocation, which will pollute cache line.
- So in modern memory allocators, it's much more complicated…
  - Roughly speaking, a common way is to split memory into bins **indexed by approximate size**.
  - And metadata may record more info:

```
// Thread local data
struct mi_tld_s {
  unsigned long long  heartbeat;
  bool                recurse;
  mi_heap_t*          heap_backing;
  mi_heap_t*          heaps;
  mi_segments_tld_t   segments;
  mi_os_tld_t         os;
  mi_stats_t          stats;
};
```

Adopted from mimalloc.

Sized-delete

- And in C++, we can almost always know object type exactly…
  - So we can know size of object!
- To facilitate optimization, C++ introduces *size-aware delete* (also called *sized-delete*).
  - Global sized delete is provided in C++14, while class-specific one is from C++11.

```cpp
void operator delete  ( void* ptr, std::size_t sz ) noexcept;
void operator delete[]( void* ptr, std::size_t sz ) noexcept;
```

  - For a naïve example:

```cpp
void operator delete(void* ptr, std::size_t size) noexcept
{
    std::println("Derived: Called overrided operator delete, ptr={}, size={}",
        ptr, size);
    free(ptr);
}
```

```cpp
int* ptr = new int;
delete ptr;
```

```
Called overrided operator new, size=4
Derived: Called overrided operator delete, ptr=0x145540d4ce0, size=4
```

Sized-delete

- Note 1: some practical example:
  - Like in jemalloc:

```
void
operator delete(void *ptr, std::size_t size) noexcept {
    sizedDeleteImpl(ptr, size);
}
```

```
JEMALLOC_ALWAYS_INLINE
void
sizedDeleteImpl(void* ptr, std::size_t size) noexcept {
    if (unlikely(ptr == nullptr)) {
        return;
    }
    LOG("core.operator_delete.entry", "ptr: %p, size: %zu", ptr, size);

    je_sdallocx_noflags(ptr, size);

    LOG("core.operator_delete.exit", "");
}
```

- Note 2: compilers are free to choose sized-delete or normal delete.
  - So, programmer should **always provide both of them**.
  - For gcc/msvc/clang (clang needs `-fsized-deallocation` flag):
    - For global override, it will prefer sized version when it exists.
    - For class-specific override, it will prefer normal delete when it exists (since you can easily know its size by `sizeof`).

Aligned new/delete

- But these overloads don’t specify alignment…
  - Before C++17, over-aligned types may be not correctly handled (normally compiler warning in e.g. `-Wall`).

```cpp
// ptr is possibly not aligned to 1024!
A* ptr = new A;
delete ptr;
```

```cpp
struct alignas(1024) A
{
    int a;
};
```

- Since C++17, alignment-aware new/delete are introduced.
  - Here `std::align_val_t` is scoped enumeration as tag.
  - For type whose alignment requirement exceeds macro `__STDCPP_DEFAULT_NEW_ALIGNMENT__`, alignment-aware overloads are preferred.
  - Of course, you can override them too.

```cpp
void* operator new  ( std::size_t count, std::align_val_t al );
void* operator new[]( std::size_t count, std::align_val_t al );
```

```cpp
void operator delete  ( void* ptr, std::align_val_t al ) noexcept;
void operator delete[]( void* ptr, std::align_val_t al ) noexcept;
void operator delete  ( void* ptr, std::size_t sz ) noexcept;
void operator delete[]( void* ptr, std::size_t sz ) noexcept;
void operator delete  ( void* ptr, std::size_t sz,
                        std::align_val_t al ) noexcept;
void operator delete[]( void* ptr, std::size_t sz,
                        std::align_val_t al ) noexcept;
```

- For class-specific ones:

```cpp
void* T::operator new  ( std::size_t count, std::align_val_t al );
void* T::operator new[]( std::size_t count, std::align_val_t al );
void T::operator delete  ( void* ptr, std::align_val_t al );
void T::operator delete[]( void* ptr, std::align_val_t al );
void T::operator delete  ( void* ptr, std::size_t sz );
void T::operator delete[]( void* ptr, std::size_t sz );
void T::operator delete  ( void* ptr, std::size_t sz, std::align_val_t al );
void T::operator delete[]( void* ptr, std::size_t sz, std::align_val_t al );
```

C11/C++17 provides `aligned_alloc` similarly; but MS-STL doesn't provide `aligned_alloc` since Windows doesn't provide ability to allocate aligned memory and thus must over-allocate and align manually. Therefore, it cannot be freed correctly by `free`; instead, `_aligned_alloc` and `_aligned_free` should be used.

new/delete

- Note 1: all new-overloads has nothrow variants:

```
void* operator new  ( std::size_t count, const std::nothrow_t& tag );
void* operator new[]( std::size_t count, const std::nothrow_t& tag );
void* operator new  ( std::size_t count, std::align_val_t al,
                      const std::nothrow_t& tag ) noexcept;
void* operator new[]( std::size_t count, std::align_val_t al,
                      const std::nothrow_t& tag ) noexcept;
```

- Note 2: essentially, new-expression `new(args…) Type{…}` will call `operator new(size(, align), args…)`.
  - The arguments before `args…` are usually determined by compilers, while the latter are specified by users.

new/delete

*: Placement is abused in the context of new/delete; here it just means to provide additional parameters in `new(…)`, which includes nothrow variant and placement-new variant.

- That's why you can use:
  - `new(std::nothrow) Type`, for there exists nothrow variant.
  - `new(ptr) Type`, for there exists placement-new variant.
    - But they can't be overridden by users.

Non-allocating placement allocation functions

```
void* operator new  ( std::size_t count, void* ptr );
void* operator new[]( std::size_t count, void* ptr );
```

- More generally, you can provide customized arguments for user-defined placement* allocation new:
  - Class-specific ones also exist, omitted here.

```
void* operator new  ( std::size_t count, /* args... */ );
void* operator new[]( std::size_t count, /* args... */ );
void* operator new  ( std::size_t count,
                      std::align_val_t al, /* args... */ );
void* operator new[]( std::size_t count,
                      std::align_val_t al, /* args... */ );
```

```
void operator delete  ( void* ptr, args... );
void operator delete[]( void* ptr, args... );
```

- Plus placement deallocation delete:
  - Each user-defined `new` must has a matching user-defined `delete`; when **constructor throws**, `new` memory will be freed by corresponding `delete`.
  - Otherwise memory leak! For example (omit sized-delete):

```
struct S
{
    S() = default;
    S(int) { throw std::runtime_error{ "Hi" }; }

    void* operator new(std::size_t byteCnt, const std::string& msg)
    {
        std::println("New {}: {}", msg, byteCnt);
        return ::operator new(byteCnt);
    }

    // Non-placement deallocation function:
    void operator delete(void* ptr)
    {
        std::println("Delete {}", ptr);
        ::operator delete(ptr);
    }

    void operator delete(void* ptr, const std::string& msg)
    {
        std::println("Delete {}: {}", msg, ptr);
        ::operator delete(ptr);
    }
};
```

```
int main()
{
    S* p = new ("123") S;
    delete p;

    try {
        p = new("442") S{1};
    } catch(const std::exception& ex) {
        std::println("Exception: {}", ex.what());
    }
}
```

```
New 123: 1
Delete 0x5a75def022c0
New 442: 1
Delete 442: 0x5a75def022c0
Exception: Hi
```

Prevent memory leak.

Only failed new-expression will call corresponding placement delete!

new/delete

- And similarly, for nothrow new, you need to customize placement delete…

```
void operator delete  ( void* ptr, const std::nothrow_t& tag ) noexcept;
void operator delete[]( void* ptr, const std::nothrow_t& tag ) noexcept;
void operator delete  ( void* ptr, std::align_val_t al,
                        const std::nothrow_t& tag ) noexcept;
void operator delete[]( void* ptr, std::align_val_t al,
                        const std::nothrow_t& tag ) noexcept;
```

- Finally, if a placement allocation corresponds to a non-placement deallocation, then compile error.

```
struct S
{
    // Placement allocation function:
    static void* operator new(std::size_t, std::size_t);

    // Non-placement deallocation function:
    static void operator delete(void*, std::size_t);
};

S* p = new (0) S; // error: non-placement deallocation function matches
                  //        placement allocation function
```

This is sized delete.

new/delete

- Note 3: for the default thrown `operator new`, it will call new handler when allocation fails.
  - As if:

```
void* operator new(std::size_t byteCnt)
{
    auto ptr = operator new(byteCnt, std::nothrow);
    while (ptr == nullptr)
    {
        auto handler = std::get_new_handler();
        if (handler == nullptr)
            throw std::bad_alloc{};
        handler();
        ptr = operator new(byteCnt, std::nothrow);
    }
    return ptr;
}
```

The default new handler is just `nullptr`, so it will `throw` `std::bad_alloc` directly.

new/delete

- You can customize it by `std::set_new_handler(…)` in `<new>` (thread-safe), and the handler is expected to:
  1. Make more memory available (so after calling handler, allocation retry may succeed);
  2. Terminate the program (e.g. by `std::terminate`);
  3. Throw exception derived from `std::bad_alloc`, or `std::set_new_handler(nullptr)`.
  - Return value: previous handler.
  - For example:

```cpp
void handler()
{
    std::cout << "Memory allocation failed, terminating\n";
    std::set_new_handler(nullptr);
}

int main()
{
    std::set_new_handler(handler);
```

```cpp
    try
    {
        while (true)
        {
            new int[1000'000'000ul]();
        }
    }
    catch (const std::bad_alloc& e)
    {
        std::cout << e.what() << '\n';
    }
}
```

new/delete

- Note 4: C++20 introduces class-specific destroying-delete.

```
void T::operator delete( T* ptr, std::destroying_delete_t );
```

```
void T::operator delete( T* ptr, std::destroying_delete_t,
                         std::align_val_t al );
```

```
void T::operator delete( T* ptr, std::destroying_delete_t, std::size_t sz );
```

```
void T::operator delete( T* ptr, std::destroying_delete_t,
                         std::size_t sz, std::align_val_t al );
```

  - Which will be preferred over all other overloads.
  - delete-expression will call destroying-delete directly, without calling dtor.
    - That is, it’s duty of the destroying-delete to call dtor.
  - Array doesn’t have this overload.
- Note 5: it should be thread-safe to call `operator new/delete`.

new/delete in coroutine

- Special example: control allocation of coroutine.
  - Coroutine will allocate its state/frame by `new`;
  - C++ allows you to customize `operator new/delete` of `promise_type` to control such allocation!
- It's specially treated so not exactly same as normal class-specific allocation/deallocation.
  - Class-specific ones need lots of overloads to cover every possible case;
  - But `promise_type` only needs to define a few for compiler to choose!
  - For new, it only needs: `void* operator new ( std::size_t count );`
  - For delete, it only needs: `void operator delete ( void* ptr, std::size_t sz ) noexcept;`
    - When this overload doesn't exist, it needs: `void operator delete ( void* ptr ) noexcept;`

new/delete in coroutine

Memory resource will be covered in later sections.

- For example:

```cpp
class CoroTask {
    inline static std::array<std::byte, 200000> memory;
    // covered in t
    inline static std::pmr::monotonic_buffer_resource buffer{
        memory.data(), memory.size(), std::pmr::null_memory_resource()
    };
    inline static std::pmr::synchronized_pool_resource mempool{ &buffer };

public:
    struct promise_type {
        void* operator new(std::size_t size) {
            return mempool.allocate(size);
        }
        void operator delete(void* ptr, std::size_t size) {
            mempool.deallocate(ptr, size);
        }
    };
}
```

new/delete in coroutine

- Note 1: when defining `get_return_object_on_allocation_failure`, you should make `operator new` act as if nothrow instead of defining nothrow variant.
  - For example:

```cpp
void* operator new(std::size_t size) {
    return new(std::nothrow) std::byte[size];
}
```

- Note 2: compilers are allowed to omit your `operator new/delete` when performing HALO.
  - So theoretically, one way to ensure HALO is to only declare `operator new/delete` without definition, so allocating on heap will lead to link error.
- Note 3: `operator new` is allowed to accept parameters of coroutine.
  - A naïve example:
  - And it's preferred if exist.

```cpp
struct promise_type {
    void* operator new(std::size_t sz, int, const std::string&) {
        return mempool.allocate(sz);
    }
};

CoroTask coro1(int a, std::string s); // 这个协程会使用重载的operator new.
CoroTask coro2(int a); // 这个协程不会使用。
```

new/delete in coroutine

- Take `std::generator` as an example:

```
void* operator new( std::size_t size )
    requires std::same_as<Allocator, void> ||
             std::default_initializable<Allocator>;
```

```
template< class Alloc, class... Args >
void* operator new( std::size_t size, std::allocator_arg_t,
                    const Alloc& alloc, const Args&... );
```

For member coroutine.

```
template< class This, class Alloc, class... Args >
void* operator new( std::size_t size, const This&, std::allocator_arg_t,
                    const Alloc& alloc, const Args&... );
```

  - Implementation may then allocate more bytes than `size`, then put allocator on additional space.
  - The `delete` can extract allocator from the frame to do deallocation.

```
void operator delete( void* ptr, std::size_t n ) noexcept;
```

new/delete in coroutine

- Use it by passing additional parameters.

```cpp
template< class Alloc, class... Args >
void* operator new( std::size_t size, std::allocator_arg_t,
                    const Alloc& alloc, const Args&... );
```

```cpp
using Alloc = std::allocator<void>;
template<typename T>
using Generator = std::generator<T, void, Alloc>;

Generator<char> TestImpl(std::allocator_arg_t, Alloc alloc,
                         std::string str)
{
    for (auto ch : str)
        co_yield ch;
}

Generator<char> Test(std::string str)
{
    return TestImpl(std::allocator_arg, Alloc{}, std::move(str));
}
```

new/delete

- Final note: in shared library, global override of `operator new/delete` should be paid special attention.
  - Reason: if each shared library has its own override, it may be unclear which one is used.
    - For example, when A is loaded, its memory is allocated by its `operator new`;
    - And B is loaded, then `operator delete` is replaced;
    - And when A frees its memory, it uses `operator delete` of B, causing unknown results.
    - The behaviors are totally implementation-defined.
  - In static library, this will cause link error for symbol conflict.