`std::generator`: Synchronous Coroutine Generator for Ranges

Document #:	P2168R1
Date:	2021-01-18
Project:	Programming Language C++
Audience:	LEWG

Reply-to:	Lewis Baker <lewissbaker@gmail.com>
	Corentin Jabot <corentin.jabot@gmail.com>

Abstract

We propose a standard library type `std::generator` which implements a coroutine generator compatible with ranges.

Revisions

R1

- Add benchmarks results and discussion about performance
- Introduce `elements_of` to avoid ambiguities when a generator is convertible to the reference type of the parent generator.
- Add allocator support
- Symmetric transfer works with generators of different value / allocator types
- Remove `iterator::operator->`
- Put `generator` in a new `<generator>` header.
- Add an other example to motivate the `Value` template parameter

Example

```
std::generator<int> fib (int max) {
    co_yield 0;
    auto a = 0, b = 1;

    for(auto n : std::views::iota(0, max))  {
        auto next = a + b;
        a = b, b = next;
        co_yield next;
    }
```

```
}

int answer_to_the_universe() {
    auto coro = fib(7) ;
    return std::accumulate(coro | std::views::drop(5), 0);
}
```

Motivation

C++ 20 had very minimalist library support for coroutines. Synchronous generators are an important use case for coroutines, one that cannot be supported without the machinery presented in this paper. Writing an efficient and correctly behaving recursive generator is non-trivial, the standard should provide one.

Design

While the proposed `std::generator` interface is fairly straight-forward, a few decisions are worth pointing out.

`input_view`

`std::generator` is a non-copyable `view` which models `input_range` and spawn move-only iterators. This is because the coroutine frame is a unique resource (even if the coroutine *handle* is copyable). Unfortunately, some generators can satisfy the `view` constraints but fail to model the `view` O(1) destruction requirement:

```
template <typename T>
std::generator<T> all (vector<T> vec) {
    for(auto & e : vec)  {
        co_yield e;
    }
}
```

Header

Multiple options are available as to where put the `generator` class.

- `<coroutine>`, but `<coroutine>` is a low level header, and `generator` depends on bits of `<type_traits>` and `<iterator>`.
- `<ranges>`
- A new `<generator>`

Separately specifyable Value Type

This proposal supports specifying both the ”yielded” type, which is the iterator ””reference”” type (not required to be a reference), and its corresponding value type. This allow ranges to handle proxy types and wrapped `reference`, like this implementation of `zip`:

```
    template<std::ranges::input_range Rng1,
    std::ranges::input_range Rng2>
    generator<
    std::tuple<std::ranges::range_reference_t<Rng1>,
    std::ranges::range_reference_t<Rng2>,
    std::tuple<std::ranges::range_value_type_t<Rng1>,
    std::ranges::range_value_type_t<Rng2>>>
    zip(Rng1 r1, Rng2 r2) {
        auto it1 = std::ranges::begin(r1);
        auto it2 = std::ranges::begin(r2);
        auto end1 = std::ranges::end(r1);
        auto end2 = std::ranges::end(r2);
        while (it1 != end1 && it2 != end2) {
            co_yield {*it1, *it2};
            ++it1; ++it2;
        }
    }
```

In this second example, using `string` as value type ensures that calling code can take the necessay steps to make sure iterating over a generator would not invalidate any of the yielded values

```
// Yielding string literals : always fine
std::generator<std::string_view> string_views() {
    co_yield "foo";
    co_yield "bar";
}

std::generator<std::string_view, std::string> strings() {
    co_yield "start";
    std::string s;
    for (auto sv : string_views()) {
        s = sv;
        s.push_back('!');
        co_yield s;
    }
    co_yield "end";
}

// conversion to a vector of strings
// If the value_type was string_view, it would convert to a vector of string_view,
// which would lead to undefined beavior as the string_views may get invalidated upon iteration!
auto v = std::to<vector>(strings()); // (P1206R3 [3])
```

Recursive generator

A "recursive generator" is a coroutine that supports the ability to directly `co_yield` a generator of the same type as a way of emitting the elements of that `generator` as elements of the current `generator`.

Example: A `generator` can `co_yield` other generators of the same type

```
generator<const std::string&> delete_rows(std::string table, std::vector<int> ids) {
    for (int id : ids) {
        co_yield std::format("DELETE FROM {0} WHERE id = {1}", table, id);
    }
}

generator<const std::string&> all_queries() {
    co_yield elements_of(delete_rows("user", {4, 7, 9 10}));
    co_yield elements_of(delete_rows("order", {11, 19}));
}
```

Example: A `generator` can also be used recursively

```
struct Tree {
    Tree* left;
    Tree* right;
    int value;
};

generator<int> visit(Tree& tree) {
    if (tree.left) co_yield elements_of(visit(*tree.left));
    co_yield tree.value;
    if (tree.right) co_yield elements_of(visit(*tree.right));
}
```

In addition to being more concise, the ability to directly yield a nested generator has some performance benefits compared to iterating over the contents of the nested generator and manually yielding each of its elements.

Yielding a nested `generator` allows the consumer of the top-level coroutine to directly resume the current leaf generator when incrementing the iterator, whereas a solution that has each generator manually iterating over elements of the child generator requires O(depth) coroutine resumptions/suspensions per element of the sequence.

Example: Non-recursive form incurs O(depth) resumptions/suspensions per element and is more cumbersome to write

```
generator<int> slow_visit(Tree& tree) {
    if (tree.left) {
        for (int x : elements_of(visit(*tree.left)))
        co_yield x;
    }
    co_yield tree.value;
    if (tree.right) {
        for (int x : elements_of(visit(*tree.right)))
```

```
        co_yield x;
    }
}
```

Exceptions that propagate out of the body of nested `generator` coroutines are rethrown into the parent coroutine from the `co_yield` expression rather than propagating out of the top-level 'iterator::operator++()'. This follows the mental model that 'co_yield someGenerator' is semantically equivalent to manually iterating over the elements and yielding each element.

For example: `nested_ints()` is semantically equivalent to `manual_ints()`

```
generator<int> might_throw() {
    co_yield 0;
    throw some_error{};
}

generator<int> nested_ints() {
    try {
        co_yield elements_of(might_throw());
    } catch (const some_error&) {}
    co_yield 1;
}

// nested_ints() is semantically equivalent to the following:
generator<int> manual_ints() {
    try {
        for (int x : might_throw()) {
            co_yield x;
        }
    } catch (const some_error&) {}
    co_yield 1;
}

void consumer() {
    for (int x : nested_ints()) {
        std::cout << x << " "; // outputs 0 1
    }

    for (int x : manual_ints()) {
        std::cout << x << " "; // also outputs 0 1
    }
}
```

`elements_of`

`elements_of` is a utility function that prevents ambiguity when a nested generator type is convertible to the value type of the present generator

```
generator<int> f()
{
    co_yield 42;
```

```
}

generator<any> g()
{
    co_yield f(); // should we yield 42 or generator<int> ?
}
```

To avoid this issue, we propose that:

- `co_yield <expression>` always yield the value directly.
- `co_yield elements_of(<expression>)` yield the values of the nested generator.

For convenience, we further propose that `co_yield elements_of(x)` be extended to support yielding the values of arbitrary ranges beyond generators, ie

```
generator<int> f()
{
    std::vector<int> v = /*... */;
    co_yield elements_of(v);
}
```

Symmetric transfer

The recursive form can be implemented efficiently with symmetric transfer. Earlier works in [CppCoro] implemented this feature in a distinct `recursive_generator` type.

However, it appears that a single type is reasonably efficient thanks to HALO optimizations and symmetric transfer. The memory cost of that feature is 3 extra pointers per generator. It is difficult to evaluate the runtime cost of our design given the current coroutine support in compilers. However our tests show no noticeable difference between a `generator` and a `recursive_generator` which is called non recursively. It is worth noting that the proposed design makes sure that HALO [5] optimizations are possible.

While we think a single `generator` type is sufficient and offers a better API, there are three options:

- A single `generator` type supporting recursive calls (this proposal).
- A separate type `recursive_generator` that can yield values from either `recursive_generator` or a `generator`. That may offer very negligible performance benefits, same memory usage.
- A separate recursive_generator type which can only yield values from other `recursive_generator`.

  That third option would make the following ill-formed:

  ```
  generator<int> f();
  recursive_generator<int> g() {
      co_yield f(); // incompatible types
  }
  ```

Instead you would need to write:

```
recursive_generator<int> g() {
    for (int x : f()) co_yield x;
}
```

Such a limitation can make it difficult to decide at the time of writing a generator coroutine whether or not you should return a `generator` or `recursive_generator` as you may not know at the time whether or not this particular generator will be used within `recursive_generator` or not.

If you choose the `generator` return-type and then later someone wants to yield its elements from a `recursive_generator` then you either need to manually yield its elements one-by-one or use a helper function that adapts the `generator` into a `recursive_generator`. Both of these options can add runtime cost compared to the case where the generator was originally written to return a `recursive_generator`, as it requires two coroutine resumptions per element instead of a single coroutine resumption.

Because of these limitations, we are not recommending this approach.

Symmetric transfer is possible for different generator types as long as the `reference` type is the same, aka, different value type or allocator type does not preclude symmetric transfer.

How to store the yielded value in the promise type?

The yielded expression is guaranteed to be alive until the coroutine resumes, it is, therefore, sufficient to store its address. This makes `generator` with a large yielded type efficient. However, it might pessimize yielding values smaller than a pointer because of the added indirection. (It is unclear what the cost of this indirection is, as none of these accesses should result in cache misses).

More annoyingly, this prevents conversions in yielding expressions:

```
generator<string_view> f() {
    co_yield std::string(); // error: cannot convert std::string to std::string_view \&
}
```

Storing a copy would allow less indirection and the ability to yield any values convertible to the yielded type, at the cost of more storage. To avoid that storage cost, a `generator<const T&>` can be used.

Allocator support

In line with the design exploration done in section 2 of P1681R0 [4], `std::generator` can support both stateless and stateful allocators, and strive to minimize the interface verbosity for stateless allocators, by templating both the generator itself and the `promise_type`'s `new` operator on the allocator type. Details for this interface are found in P1681R0 [4]. Allocators

passed as parameter to the coroutine function do not need to be default constructible if we mandate their `value_type` is `std::byte` (such that they do not need to be rebound).

`coroutine_parameter_preview_t` such as discussed in section 3 of P1681R0 [4] has not been explored in this paper.

```
std::generator<int, int, std::allocator<std::byte>> stateless_example() {
    co_yield 42;
}

template <typename Allocator>
std::generator<int, int, Allocator>
allocator_example(std::allocator_arg_t, Allocator&& alloc) {
    co_yield 42;
}

my_allocator<std::byte> alloc;
input_range auto rng = allocator_example<my_allocator<std::byte>>(std::allocator_arg, alloc);
```

Supporting allocators requires storing, in all cases, a function pointer adjacent to the coroutine frame (to track a deallocation function), along with the allocator itself in the case of stateful allocators.

The proposed interface requires that, if an allocator is provided, it is the second argument to the coroutine function, immediately preceded by an instance of `std::allocator_arg_t`. This approach is necessary to distinguish the allocator desired to allocate the coroutine frame from allocators whose purpose is to be used in the body of the coroutine function. The required argument order might be a limitation if any other argument is required to be the first, however, we cannot think of any scenario where that would be the case.

We think it is important that all standard and user coroutines types can accommodate similar interfaces for allocator support. In fact, the implementation for that allocator support can be shared amongst `generator`, `lazy` and other standard types.

Can we postpone adding support for allocator later?

A case can be made that allocator support could be added to `std::generator` later. However, because the proposed design has the allocator as a template parameter, adding allocator after `std::generator` ships would represent an ABI break. We recommend that we add allocator support as proposed in this paper now and make sure that the design remains consistent as work on `std::lazy` is made in this cycle. However, it would be possible to extend support for different mechanisms (such as presented in section 3 of P1681R0 [4] later.

Implementation and experience

`generator` has been provided as part of cppcoro and folly. However, cppcoro offers a separate `recursive_generator` type, which is different than the proposed design.

Folly uses a single `generator` type which can be recursive but doesn't implement symmetric transfer. Despite that, Folly users found the use of `Folly::Generator` to be a lot more efficient than the eager algorithm they replaced with it.

`ranges-v3` also implements a `generator` type, which is never recursive and predates the work on move-only views and iterators [1], [2] which forces this implementation to ref-count the coroutine handler.

Our implementation [Implementation] consists of a single type that takes advantage of symmetric transfer to implement recursion.

Performance & benchmarks

Because implementations are still being perfected, and because performance is extremely dependant on whether HALO optimization (see P0981R1 [**?**]) occurs, it is difficult at this time to make definitive statements about the performance of the proposed design.

At the time of the writing of this paper, Clang is able to inline non-nested coroutines whether the implementation supports nested coroutines or not, while GCC never performs HALO optimization.

When the coroutine is not inlined, support for recursion does not noticeably impact performance. And, when the coroutine yields another generator, the performance of the recursive version is noticeably faster than yielding each element of the range. This is especially noticeable with deep recursion.

	Clang	Clang ST[1]	GCC	GCC ST[1]	MSVC	MSVC ST[1]
Single value	(1) 0.235	(2) 2.36	12.4	13.4	61.9	63.7
Single value, noinline (3)	13.5	13.7	14.1	15.2	63.8	64.4
Deep nesting	43670266.0	(4) 427955.0	58801348	338736	224052033	4760914

[1] `Symmetric transfer.`

The values are expressed in nanoseconds. However, please note that the comparison of the same result across compiler is not meaningful, notably because the MSVC results were obtained on different hardware. That being said we observe:

- Only Clang can perform constant folding of values yielded by simple coroutine (1)
- When the `generator` supports symmetric transfer, clang is not able to fully inline the generator construction, but HALO is still performed (2).
- When HALO is not performed, the relative performance of both approach is similar (3).
- Supporting recursion is greatly beneficial to nested/recursive algorithms (4).

The code for these benchmarks as well as more detailled results can be found on Github.

Wording

The following wording is meant to illustrate the proposed API.

🔲 Header <ranges> synopsis [ranges.syn]

```
template<typename T>
struct elements_of;

template<std::ranges::input_range R>
struct elements_of<R> {
    R&& __range; // exposition only

    explicit constexpr elements_of(R&& r) noexcept : range((R&&)r) {}
    constexpr elements_of(elements_of&&) noexcept = default;

    constexpr elements_of(const elements_of&) = delete;
    constexpr elements_of& operator=(const elements_of&) = delete;
    constexpr elements_of& operator=(elements_of&&) = delete;

    constexpr R && get() && noexcept {
        return std::forward<R>(__range);
    }
};
template<std::ranges::input_range R>
elements_of(R&&) -> return elements_of<R>;
```

🔲 Header <generator> synopsis [generator.syn]

```
#include <coroutine>
#include <ranges>

namespace std {

    template<typename Ref, typename Value  = std::remove_cvref_t<Ref>>
    class generator;

    template <typename Ref, typename Value>
    inline constexpr bool ranges::enable_view<generator<Ref, Value>> = true;
}
```

🔲 Generator View [coroutine.generator]

🔲 Overview [coroutine.generator.overview]

`generator` produces an `input_view` over a synchronous coroutine function yielding values.

[*Example:*

```
        generator<int> iota(int start = 0) {
            while(true)
            co_yield start++;
        }

        void f() {
            for(auto i : iota() | views::take(3))
            cout << i << " " ; // prints 0 1 2
        }
```

— end example]

🔲 Class template `generator` [coroutine.generator.class]

```
namespace std {

template <typename Ref, typename Value = std::remove_cvref_t<Y>,
          typename Allocator = std::allocator<std::byte>>
class generator  {
    public:
    class promise_type;
    class iterator;
    class sentinel {};

    private:
    std::coroutine_handle<promise_type> coroutine_ = nullptr; // exposition only

    explicit generator(std::coroutine_handle<promise_type> coroutine) noexcept // exposition only
    : coroutine_(coroutine) {}

    public:
    generator() noexcept;
    generator(const generator &other) = delete;
    generator(generator && other) noexcept
    : coroutine_(exchange(other.coroutine_, nullptr)){}

    ~generator() {
        if (coroutine_) {
            coroutine_.destroy();
        }
    }

    generator &operator=(generator && other) noexcept {
        swap(other);
        return *this;
    }

    iterator begin();
```

```
    sentinel end() noexcept
    { return {};  }

    void swap(generator & other) noexcept {
        std::swap(coroutine_, other.coroutine_);
    }

};
}
```

Mandates:

- `Allocator` meets2 the `Cpp17Allocator` requirements,
- `same_as<Allocator::value_type>, std::byte>` is `true`.

```
iterator begin();
```

Preconditions: `!coroutine_` is `true` or `coroutine_` refers to a coroutine suspended at its initial suspend-point.

Effects: Equivalent to:

```
if(coroutine_)
coroutine_.resume();
return iterator{coroutine_};
```

[*Note:* It is undefined behavior to call `begin` multiple times on the same coroutine. — *end note*]

🔲 Class template `generator::promise_type` [coroutine.generator.promise]

```
template <typename Ref, typename Value, typename Allocator>
class generator<Ref, Value, Allocator>::promise_type {

    friend generator;

    union {
        Ref value_; // exposition only
    };

public:
    using value_type = V;
    using reference  = Y;

    generator<Y, V> get_return_object() noexcept;

    std::suspend_always initial_suspend() const {
        return {};
    }
```

```
    auto final_suspend() const;

    std::suspend_always yield_value(reference && value) n
        noexcept(std::is_nothrow_copy_constructible_v<reference, T>));

    template <typename T>
    requires(!std::is_reference_v<Ref>) && std::is_convertible_v<T, Ref>
    std::suspend_always yield_value(T &&x) noexcept(std::is_nothrow_constructible_v<Ref, T>);

    template <typename TVal, typename TAlloc>
    unspecified yield_value(elements_of<generator<Ref, TVal, TAlloc>>&& g) noexcept; // see below

    template<std::ranges::input_range R>
    requires convertible_to<ranges::range_reference_t<R>, Y>
    unspecified yield_value(elements_of<R&&>&& rng) noexcept; // see below

    void await_transform() = delete;

    void return_void() noexcept {}

    void unhandled_exception();

    static void* operator new(std::size_t size);

    template<typeame Alloc, typename... Args>
    static void* operator new(std::size_t size, std::allocator_arg_t, Alloc&& alloc, Args&...);
    template<typename This, typeame Alloc, typename... Args>
    static void* operator new(std::size_t size, This&, std::allocator_arg_t, Alloc&& alloc, Args&...);

    static void operator delete(void *pointer, size_t size) noexcept;
};

generator<Ref, Value, Allocator> get_return_object() noexcept;
```

Effects: Equivalent to:

```
    return generator<Ref, Value, Allocator>{
        std::coroutine_handle<promise_type>::from_promise(*this)};
```

```
std::suspend_always yield_value(reference && value)
    noexcept(std::is_nothrow_copy_constructible_v<reference, T>));

template <typename T>
requires(!std::is_reference_v<Ref>) && std::is_convertible_v<T, Ref>
std::suspend_always yield_value(T &&x)
    noexcept(std::is_nothrow_constructible_v<Ref, T>);
```

Effects: Assign `x` to `value_`;

If the execution control has been transferred from this promise to another `generator`'s `promise_type` noted `other`, assign `x` to `other.value_`;

[*Note:* Generators can transfer control recursively, `value` returns the value set on promise associated with the child-most generator coroutine. — *end note*]

```
template <typename TVal, typename TAlloc>
auto yield_value(elements_of<generator<Ref, TVal, TAlloc>>&& g) noexcept;
```

Mandates:

- `TAlloc` meets the `Cpp17Allocator` requirements,
- `same_as<TAlloc::value_type>, std::byte>` is `true`.

Effects: This function returns an implementation defined `awaitable` type which takes ownership of the generator `g`.

[*Note:* This ensures that local variables in-scope in `g`'s coroutine are destructed before local variables in-scope in this coroutine being destructed. — *end note*]

Execution is transferred to the coroutine represented by `g.coroutine_` until its completion. After `g.coroutine_` completes, the current coroutine is resumed.

[*Note:* Generators can transfer control recursively. — *end note*]

```
template<std::ranges::input_range R>
requires convertible_to<ranges::range_reference_t<R>, Ref>
std::suspend_always yield_value(yield_value(elements_of<R>&& rng) noexcept;
```

Effects: Calls `co_yield elem` for each element `elem` of the range `rng`.

🔒 Class template `generator::iterator` [coroutine.generator.iterator]

```
template <typename Ref, typename Value, typename Allocator>
class generator::iterator {
    private:
    std::coroutine_handle<promise_type> coroutine_ = nullptr;


    public:
    using iterator_category = std::input_iterator_tag;
    using difference_type = std::ptrdiff_t;
    using value_type = promise_type::value_type;
    using reference = promise_type::reference;

    iterator() noexcept = default;
    iterator(const iterator &) = delete;


    iterator(iterator && other) noexcept
    : coroutine_(exchange(other.coroutine_, nullptr)) {}

    iterator &operator=(iterator &&other) noexcept {
        coroutine_ = exchange(other.coroutine_, nullptr);
```

```
    }

    explicit iterator(std::coroutine_handle<promise_type> coroutine) noexcept
    : coroutine_(coroutine) {}

    bool operator==(sentinel) const noexcept {
        return !coroutine_ || coroutine_.done();
    }

    iterator &operator++();
    void operator++(int);

    reference operator*() const noexcept (noexcept(std::is_nothrow_copy_constructible_v<reference>));

};
```

```
iterator &operator++();
```

Preconditions: `coroutine_ && !coroutine_.done()` is `true`.

Effects: Equivalent to:

```
coroutine_.resume();
return *this;
```

```
void operator++(int);
```

Preconditions: `coroutine_ && !coroutine_.done()` is `true`.

Effects: Equivalent to:

```
(void)operator++();
```

```
reference operator*() const
noexcept (noexcept(std::is_nothrow_copy_constructible_v<reference>));
```

Preconditions: `coroutine_ && !coroutine_.done()` is `true`.

Effects: Equivalent to:

```
return coroutine_.promise().value();
```

Feature test macros

Insert into [version.syn]

```
#define __cpp_lib_generator <DATE OF ADOPTION>
```

References

[1] Casey Carter. P1456R1: Move-only views. https://wg21.link/p1456r1, 11 2019.

[2] Corentin Jabot. P1207R0: Movability of single-pass iterators. https://wg21.link/p1207r0, 8 2018.

[3] Corentin Jabot, Eric Niebler, and Casey Carter. P1206R3: ranges::to: A function to convert any range to a container. https://wg21.link/p1206r3, 11 2020.

[4] Gor Nishanov. P1681R0: Revisiting allocator model for coroutine lazy/task/generator. https://wg21.link/p1681r0, 6 2019.

[5] Richard Smith and Gor Nishanov. P0981R0: Halo: coroutine heap allocation elision optimization: the joint response. https://wg21.link/p0981r0, 3 2018.

[CppCoro] Lewis Baker *CppCoro: A library of C++ coroutine abstractions for the coroutines TS* https://github.com/lewissbaker/cppcoro

[Folly] Facebook *Folly: An open-source C++ library developed and used at Facebook* https://github.com/facebook/folly

[range] Eric Niebler *range-v3 Range library for C++14/17/20* https://github.com/ericniebler/range-v3

[Implementation] Lewis Baker, Corentin Jabot *`std::generator` implementation* https://godbolt.org/z/Tb36xj

[N4861] Richard Smith *Working Draft, Standard for Programming Language C++* https://wg21.link/N4861