

RAST: highly distributed DB project.

Produced by NOA team

Contents

1	Coding patterns	2
1.1	Function specifications cascade merging	2
1.2	Reverse template instantiation	4
2	Algorithmic basis	9
2.1	Free-list multi-level allocator	9
2.2	Lock-free queue (without deallocations)	12
2.3	Multi-counters deallocation technique	14
2.4	Multi-counters lock-free queue (final algorithm)	15
2.5	Multi-counters lock-free queue (productivity)	17
2.6	Low-overhead periodic timer	19
2.7	Low-overhead start-finish timer	21
2.8	Low-overhead waiting timer	22
2.9	Low-overhead periodic clock	23
2.10	Exception top proto storage	23
2.11	Shadow-paging thread-wise sharder	26
2.12	Dynamically-sharded message passing pool	30
2.13	Increasing hex identifier	35
2.14	Finite state machine	36
3	Detailed Code Architecture (classes and methods)	39
3.1	TLS free-list multi-level allocator	39
3.2	Lock-free queue	40
3.3	Message processor	40
3.4	Message passing tree	41
3.5	Low-overhead timers	46
3.6	Exception with backtrace	47
3.7	Exception top proto storage interface	48
3.8	Shadow-paging thread-wise sharder interface	49
3.9	Dynamically-sharded message passing pool interface	50
3.10	Increasing hex identifier interface	51

1. Coding patterns

1.1. Function specifications cascade merging

Let us assume we are to write some template class, which may be customized with any amount of template parameters. Then, of course, we have to use variadic templates. When using variadic templates, we are to define class through its older version with less amount of template arguments, thus forcing compiler to produce a code for several hierarchical classes. What is often used here is so called cascade inheritance. That is, new class which is specified by a larger amount of template parameters is being inherited from an older version.

Imagine that we want target class $X < A, B, C >$ to have a template function, which has concrete specifications for template parameters being equal to A , B or C . For instance, all classes A , B and C have *Output* method and we want class X to have *Output* $< T >$ method which for $[T = A]$ calls *Output* method from class A , for $[T = B]$ calls *Output* method of class B , etc.

First difficulty we would face considering such approach is that template specifications are not allowed in non-global scope (inside class X , for instance). That is due to the fact, that if those specifications depend on template parameters of X class itself, the more complicated version of template processing algorithm would be required to handle this. But, template processing is not that smart enough to guess what author meant here. So, we have to specialize functions another way.

Let us do specifications by means of input arguments. In order to do so let us introduce an empty template class *TypeSpecifier* $< T >$. We may now write three versions of function, which compiler would treat well:

```
class X {
public:
    void Output(const TypeSpecifier<A>&) {
        A().Output();
    }
    void Output(const TypeSpecifier<B>&) {
        B().Output();
    }
    void Output(const TypeSpecifier<C>&) {
        C().Output();
    }
};
```

Such a function would be called as follows:

```
x.Output(TypeSpecifier<A>());
x.Output(TypeSpecifier<B>());
x.Output(TypeSpecifier<C>());
```

What remains is to represent this code using template processing:

```
template <typename ... Args>
class X {
public:
    void Output() {}
};

template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
public:
```

```

using X<Args...>::Output;
void Output(const TypeSpecifier<T>&) {
    T().Output();
}
};

```

By means of “using” construction we add all versions of Output function from previous inheritance level to the next one, afterwards adding new version associated with current new template parameter T. As far as we put “using Output” inside the code, we have to define “Output” symbol inside very base class, which accepts zero template parameters. In order to do that, we firstly define version of class “X”, which accepts any amount of template arguments. The trick is that such implementation would only be considered if only “X” was zero template arguments, otherwise the declaration $X < T, Args... >$ would be chosen by compiler to process.

The only drawback here is that a user of the class has to know about some “TypeSpecifier” class to use “Output” function. But this can be fixed by introducing a final template function:

```

template <typename ... Args>
class X {
public:
    void Output() {}
};
template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
public:
    using X<Args...>::Output;
    void Output(const TypeSpecifier<T>&) {
        T().Output();
    }
    template <typename T2>
    void TemplateOutput() {
        Output(TypeSpecifier<T2>());
    }
};

```

Now a call to “Output” function looks as follows:

```

x.TemplateOutput<A>();
x.TemplateOutput<B>();
x.TemplateOutput<C>();

```

That is what is called “cascade merging of specifications” here.

The example on which technique is explained is rather simple and it would be weird to write such a complicated code, because one would just do something like this:

```

template <typename T>
void Output() {
    T().Output();
}

```

In that, in the example above there was no actual need in inheritance, the example is just used in order to demonstrate the pattern. But when is this pattern really needed then?

A situation where such a complicated technique starts to be useful is when custom function needs to know something about the moment template argument was being added to template variadic list. That is, when in such a “Output” function we have to access a specific inheritance

level. Imagine, for instance, that in class $X < A, B, C >$ we want to have a function *GetIndex*, which would by class name A , B or C return an index 2, 1 or 0. To do it we have to store somewhere a correspondence between classes and indices. The best place for this is an inherited class:

```

template <typename ... Args>
class X {
public:
    X()
    : max_index(0)
    {}

    void GetIndexImpl() {}
protected:
    int max_index;
};

template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
public:
    using X<Args...>::GetIndexImpl;
    using X<Args...>::max_index;

    X()
    : inheritance_level_index(max_index++)
    {}

    int GetIndexImpl(const TypeSpecifier<T>&) {
        return inheritance_level_index;
    }

    template <typename T2>
    int GetIndex() {
        return GetIndexImpl(TypeSpecifier<T2>());
    }
private:
    int inheritance_level_index;
};

```

That we have basically done here is incrementing “inheritance_level_index” in instantiation of each hierarchical class object, thus having increasing indices in the inheritance chain. Afterwards, while user is calling *GetIndex* $< T >$ function, the corresponding “GetIndexImpl” from the corresponding inheritance level would be called, and this function has a direct access to all the data stored in the level of inheritance it was defined. Thus, it can easily return the required index.

1.2. Reverse template instantiation

In the section [1.1. Function specifications cascade merging](#) we have learned how to access data inside specific inherited class in the inheritance chain by given type name. Now assume the inverse task: given class $X < A, B, C >$ we want it to have a function *Output*(const int n) which calls “Output” function of $n - th$ class. That is, *Output*(0) calls $C :: Output$, *Output*(1) calls $B :: Output$, etc.

First and the simplest idea would be to do recursive calls until reaching needed inheritance level, and afterwards call the required “Output” function:

```

template <typename ... Args>
class X {
public:
    X()
    : max_index(0)
    {}

    void Output(const int) {}
protected:
    int max_index;
};
template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
public:
    using X<Args...>::max_index;

    X()
    : inheritance_level_index(max_index++)
    {}

    void Output(const int index) {
        if (inheritance_level_index == index) {
            T().Output();
        } else {
            X<Args...>::Output(index);
        }
    }
}
private:
    int inheritance_level_index;
};

```

However, the drawback of such an approach is obvious: we do a lot of recursive calls, while another strategy would help up to avoid it. To cope with the task without recursive calls, we should have a fast way of accessing former ancestor classes. Let us store all hierarchical instances in vector then. But we cannot store instances of class “X” itself for the reason that it would result in infinite recursion depth if we would call instantiation of “X” class inside itself. Consequently, we have to use another class-helper “Y” the following way:

```

template <typename ... Args>
class X {
protected:
    class Y {
    public:
        virtual void OutputImpl() {}
    };
public:
    X()
    {}

    virtual void OutputImpl() {}
protected:
    std::vector<Y *> instances;
};

```

```

template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
private:
    class Y : public X<>::Y {
    public:
        void OutputImpl() {
            T().Output();
        }
    };
public:
    using X<Args...>::instances;

    X()
    : X<Args...>()
    {
        instances.push_back(new Y());
    }

    void Output(const int index) {
        instances[index]->OutputImpl();
    }
};

```

Here in the constructor we initialize instances for each level of inheritance. Now we may access each level using this vector and use dynamic function resolution to call concrete “OutputImpl” function, which does the job.

Imagine a more complicated situation, where function defined in “A”, “B” or “C” classes may want to have access to $X < A, B, C >$ caller instance. For example, we want “Output” function of class “C” to accept the caller x of type $X < A, B, C >$ and call $x.Output(1)$ in order to trigger $B :: Output$. The problem here is that $Output(1)$, which triggers $C :: Output$, is processed by $X < C >$ class in inheritance chain. That is why, it is quite unclear how do we manage to pass $X < A, B, C >$ instance to $C :: Output$.

One of the most efficient ways to let parent know about a forthcoming child type is to use templates. Let us introduce a template subclass $Y < GlobalX >$, which is going to be defined inside “X”. The aim is to make “GlobalX” template be replaced by $X < A, B, C >$ somehow during compilation process. To accomplish this let us again build up a vector of “Y” instances, but this time a vector will be constructed not during initialization, but rather after all hierarchical instances are ready, so that we could pass a last one as a template parameter. Let the initialization be in separate function called “LazyInit”.

```

template <typename ... Args>
class X {
public:
    template <typename GlobalX>
    class Y{
    public:
        Y(GlobalX& x)
        : x(x)
        {}

        virtual void Output() {}
    protected:
        GlobalX& x;
    };
};

```

```

};

X()
{}

    template <typename GlobalX>
    void LazyInitImpl(GlobalX&, std::vector<X::Y<GlobalX>*>&) {}
};

template <typename T, typename ... Args>
class X<T, Args...>: public X<Args...> {
public:
    template <typename GlobalX>
    class Y : public X<>::template Y<GlobalX> {
    public:
        using X<>::template Y<GlobalX>::Y;
        using X<>::template Y<GlobalX>::x;

        void Output() {
            T().template Output<GlobalX>(x);
        }
    };

    template <typename GlobalX>
    void LazyInitImpl(GlobalX& x, std::vector<X<>::Y<GlobalX>*>&
        y_hierarchical_instances) {
        X<Args...>::template LazyInitImpl<GlobalX>(x,
            y_hierarchical_instances);
        y_hierarchical_instances.push_back(new Y<GlobalX>(x));
    }

    void LazyInit() {
        LazyInitImpl<X>(*this, y_hierarchical_instances);
    }

    X<>::template Y<X>* GetYInstance(const int index) {
        return y_hierarchical_instances[index];
    }
private:
    std::vector<X<>::template Y<X>*> y_hierarchical_instances;
};

```

Usage of this class would be as follows:

```

class A {
public:
    template <typename GlobalX>
    void Output(GlobalX& x) {
        std::cout << "This is A!\n";
    }
};

class B {
public:
    template <typename GlobalX>

```

```

        void Output(GlobalX& x) {
            std::cout << "This is B!\n";
        }
};

class C {
public:
    template <typename GlobalX>
    void Output(GlobalX& x) {
        std::cout << "This is C!\n";
        x.GetInstance(1)->Output();
    }
};

int main() {
    X<A,B,C> x;
    x.LazyInit();
    x.GetInstance(0)->Output();
    return 0;
}

```

Note that now classes “A”, “B” and “C” are able to use *x* instance. Yet, as far as these classes does not know about *x* type, their “Output” function has to be template.

Now, with the usage of “reverse template instantiation” pattern we are able to access methods of classes by index and to access external controller class object inside each of the class passed as a template parameter.

2. Algorithmic basis

2.1. Free-list multi-level allocator

One of the most important steps in order to make a low-latency system is to consider the allocation problem. The problem is that sometimes different components of the program require very small parts of memory to be allocated. Asking the system to allocate memory directly would result in slower performance. That's why it is better to allocate large memory segment and distribute its parts among several consumers.

But here we face another difficulty, so called fragmentation problem. Imagine that an allocated segment is separated into several parts owned by consumers. Some time after, some of the consumers returned their own parts to allocator (calling deallocate method). Afterwards it appears that the returned memory in control of allocator is represented as distinct small fragments. Those fragments cannot be given to a consumer, who asks for a larger memory segment, than the size of the separate fragment. If there is a lot of such unusable fragments, we would have a lot of memory allocated, which will never be used. In order to avoid this problem, we have to force each small fragment to be used by a consumer who asks for a small amount of memory.

The main idea is to use a memory levels.



For each block size we calculate an integer part of $\log(size)$ and put all blocks of memory in the corresponding list.

By means of the list, we can easily find the block of needed size and remove (detach) it from the list, giving ownership to the consumer, who asked to allocate memory. Likewise, when a consumer asks to deallocate the block, we may put (attach) it again in the list.

Also we want to join small segments of memory returned by consumers to one large segment in case if they go one after another in the global memory space. In order to do this, let us have the following structure of the memory block:



The block will consist of three logical parts:

- FrontControl, which stores the *data_size* which is size of the block, *localnext* and *localprev* pointers, *offset* in external allocated segment, and *total_size* of external segment, and *is_owned* – whether the block is owned by the allocator itself or was it given away to the consumer.

- Data – raw byte space, which would be given an external consumer to control
- BackControl, which stores pointer to the FrontControl (pointer to the beginning of the entire block)

localprev and *localnext* pointers will help us to build freelists, which were described above. BackControl block will help us to find the block, previous to some specific Block, so that we could join two blocks together in case they both have been returned to the allocator. Here is how joining procedure works:



The steps are very simple:

- detach first block from its list
- detach second block from its list
- point **BackControl2** to **FrontControl1**
- change **FrontControl1** information about block size
- do nothing about **BackControl1** and **FrontControl2**, because now we treat them as a garbage part of **MergedData**, and we do not have to care about their contents so far
- attach new merged block to its new list (based on $\log(\text{size})$ where *size* is a sum of sizes of these two blocks) by changing *localprev* and *localnext* pointers of **FrontControl**.

Also, one can easily imagine how to split one large block into two smaller.

Blocks contained in the same external memory segment (which allocator asked from the system) may be joined and split so that we know exactly how many consequent free to use bytespace fragments do we have.

Finally, the allocation procedure is the following:

- calculate integral part $N = \text{int}(\log(\text{size}))$ of the size of the memory consumer asks to allocate
- if the corresponding N -th list of blocks is empty, we allocate several external blocks of the sizes 2^N and build Block structure upon each of them, putting blocks to the N -th list
- detach first block from the list
- split the block into two blocks: first block has exactly needed size, second block consists of what is remained
- attach second block to its corresponding list

- mark the first block as not owned by allocator
- return a pointer to the Data section of the first block to a consumer

The deallocation procedure:

- receive from a consumer a pointer to deallocate data
- calculate the pointer to FrontControl (by subtracting $sizeof(FrontControl)$ from Data pointer)
- mark block as owned by the allocator
- read $offset$ from FrontControl, if $offset > 0$, then there is some previous block in the external memory segment.
- calculate the pointer to BackControl of the previous block (by subtracting $sizeof(BackControl)$ from the pointer to the current block)
- get the pointer to the FrontControl of the previous block via BackControl
- check whether the previous block is owned by the allocator and if it is, merge it with the current block
- if $offset < total_size$ where $total_size$ is the size of external allocated memory segment, we may want to check whether we can merge with the forthcoming block
- calculate the pointer to the FrontControl of the forthcoming block (by adding $data_size + sizeof(FrontControl) + sizeof(BackControl)$ to the pointer to the current block)
- check if the forthcoming block is owned by allocator using its FrontControl and if it is, merge it with the current block.

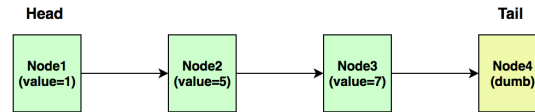
2.2. Lock-free queue (without deallocations)

Queue is a basic container, which is often desired to be used in more or less complicated systems. Although queue implementation is very simple, it is not thread-safe. In order to make it thread-safe, one may decide to use mutexes in Pop and Push operations. The idea is not good enough for several reasons:

- all threads would compete against each other for gaining control over mutex,
- all kernels would have to sync the mutex state via cache ping-pong, which would lead to plenty of time wasted on system calls,
- jobs of threads would be serialized, that is, the queue is 100% not scalable.

The main idea is that if we get down to the details of queue implementation, we may reorganize it in such a way, that we would not need to have mutually exclusive access to guarantee that each operation works fine. That is, we may obtain a lock-free algorithm on the queue.

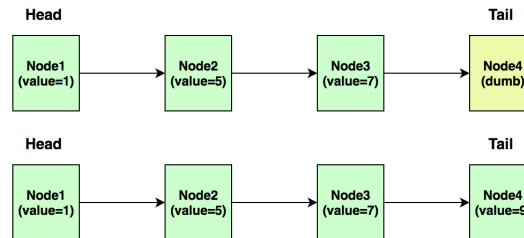
In order to do that, let us represent queue as a linked list, each node of each except for the last one, contains data elements. The last node, which is the tail of the queue would not contain any data at all. Let us call it a dumb node:



Head, Tail and Next pointers inside data structure would be atomic pointers, which would be changed using CAS operations. The main idea of Push is:

- put new data to the dumb node
- create another dumb node and mark it as a tail

But these are two distinct actions, which are not guarded by mutual exclusion and thus are not atomic. So, by choosing this strategy we should agree that it is OK for the queue to be one of the following two:



That is, we agree on purpose that queue may either have dumb element or not have dumb element. But if it is like that, we have to redesign Push:

- try CAS data on the tail (nullptr \rightarrow our new data)
- if CAS succeeded, the queue actually had a dumb node and now we may proceed to adding a new dumb node
- if CAS failed, the queue does not have a dumb element, so we firstly create a dumb node as a tail, and retry the algorithm from the start

Changing tail pointer is also a CAS operation, but in case if it fails, we realize that another thread have already moved tail forward and thus we do not have to do anything, so we will not process tail CAS operation failure on purpose without harming algorithm invariants at all.

The detailed version of Push consists of the following steps:

- allocate new data
- allocate new dumb node
- L1: load tail
- L2: try CAS tail.data (nullptr → allocated new data)
- L3: if L2 succeeded, try CAS tail.next (nullptr → new dumb node)
- if L3 failed, somebody already moved tail.next to new node, so deallocate new dumb node, try CAS tail (old_tail → value of tail.next), ignore the results and finish Push
- if L3 succeeded, try CAS tail (old_tail → new dumb node), do not care about results and finish Push
- L4: if L2 failed, we do not have dumb node in the end of the queue, try CAS tail.next (nullptr → new dumb node)
- if L4 succeeded, we shall allocate new dumb node, try CAS tail (old_tail → value of tail.next), ignore the result of this operation and repeat algorithm from L1 point
- if L4 failed, try CAS tail (old_tail → value of tail.next), ignore the result of this operation and repeat algorithm from L1 point using the same already-allocated new dumb node

Note that the only action among these which may throw is an allocation, but all allocations happen-before successful L2 step. That is, in case if L2 succeeded, the rest of the code in Push is exception-free. Success of L2 operation means that new data is put to the queue, failure means that the data was not put to the queue yet. This implies we are having strong exception safety on Push operation: if it throws, it would surely mean that data was not put to the queue. This is the best available exception guarantee (after "noexcept") for a queue user. Of course, guaranteeing noexcept is impossible, because no one can imagine Push operation without memory allocations, which always may throw.

Now let us consider Pop operation. It would check if head is equal to tail. If it is, the queue consists of dumb node only, thus there is no data. In this case Pop would return empty_ptr. In the other case, we will switch the head to the next element, giving away first node data.

The detailed version of Pop consists of the following steps:

- read head, tail
- L1: check head == tail
- if L1 is true, return empty_ptr
- L2: if L1 is false, do CAS head (current → next)
- if L2 failed, retry from the start
- if L2 succeeded, store and exchange data from old head, return it to the caller

Notice that Pop is exception-free.

The implementation described here does not do deallocations, which poses a memory leak problem. The problem is handled in the forthcoming section [2.4. Multi-counters lock-free queue \(final algorithm\)](#).

2.3. Multi-counters deallocation technique

Almost any algorithm underlying some lock-free data structure requires careful deallocation technique. The [algorithm of lock-free queue](#) we have described above is not an exclusion. The implementation considered does not deallocate memory at all, thus suffering from a memory leak.

The simple but not correct approach is to deallocate node, which is being popped. The problem here is that we have to provide strict guarantee that no other thread is still working with data, which current thread is going to return to the memory heap.

More complicated, but still not correct idea is to count references to the node. We increase reference count if we are going to use a node, and decrease when it is no longer needed. And when counter hits zero, we are going to deallocate the data. Unfortunately, this attempt to write thread-safe deallocation code would also fail, because the counter being equal zero at some point of time does not guarantee that it would be zero the next moment of time. This implies, that checking whether the counter is zero is a useless operation: after checking conditional has passed, and we moved forward to deallocating data, some other thread could ask for one more copy of the object and increase reference count by 1. But, as far as in the current thread we have already decided to deallocate memory, we no longer care about reference counter value. The result is a race condition in the code, which again leads to the fact that some thread uses memory another thread is going to deallocate.

The most complicated, and this time, absolutely correct option, is to always bear in mind reference counter, but start to attempt comparing it with zero and deallocating memory only after the moment when we can guarantee that no other reference to the object is going to be created in the future. That is, we wait until the user-written code somehow manages to say “I will no longer create new references on this specific data”. After this moment, we start to wait until user-written code stops to use existing (created before this moment) references. And after that, we may freely deallocate the memory, without any races.

In a more general case, there are several users, which may produce copies of the references to the object. Then reference counting mechanism should receive a promise not to create copies in the future from each one of them. In order to do that, we would use *external_counters* variable. At the initialization point, it stores the number of sources, which can produce copies of references to the object. Once a source tells that it would no longer produce copies, the *external_counters* decreases. Than it equals zero, we start trying to deallocate the data.

In order to store reference count, we introduce *internal_counter*, which is stored in the data itself. Apart from that, each source, which can create reference copies, would contain *external_counter* – a number of reference copies it created on specific data.

Let us call source, which has not yet promised not to create copies of references, an *active* source. In our three-counter approach the following invariant would hold: for each data element sum of *internal_counter* and all *external_counter* values of all active sources is equal to reference count to this data.

To sum up, data element contains *internal_counter* and *external_counters* inside itself, while each reference to the data element consists of the pointer as well as *external_counter*.

Now we have to define several procedures.

The access protocol:

- source, which want to copy reference, increases its *external_counter*
- ... some operations over data element ...
- send data element a command to decrease its *internal_counter*

Source nocopy promise:

- send a data element command to atomically decrease *external_counters* and to increment *internal_counter* by the value of *external_counter* of the current source

The cleaning condition:

- if *internal_counter* and *external_counters* of data element are both equal zero, deallocate the data

This condition is checked both on the finish of the access protocol and after receiving nocopy promise and if true, data is deallocated.

It is very simple to prove that all the operations described keep invariant that sum of *internal_counter* and all *external_counter* values of all active sources is equal to reference count. Because of that, cleaning condition is proven to be true. Indeed, if *external_counters* equals zero, all sources promised not to create copies in the future. Therefore, by that moment there are no active sources, that is, *internal_counter* (summed up with nothing) now stores the exact amount of references left. And, of course, if it also equals zero, then all access protocols are complete. That surely implies we can deallocate data with no problems.

There are also a few rather important details about the implementation:

- *internal_counter* and *external_counters* are stored in the data element itself and should mutate atomically and simultaneously. This can be lock-free only on processors, which support DWCAS.
- *external_counter* and pointer to the data element should also be packed to the one structure and mutate using DWCAS
- all operations described are exception-free, which means that the method can be used in any data structure without harming exception guarantees of the data structure itself.
- in order to simplify matters in the access protocol, we would incapsulate copying in the constructor of special CopyGuard class, and incapsulate decreasing *internal_counter* in its destructor.

2.4. Multi-counters lock-free queue (final algorithm)

Here we are to combine the ideas in 2.2. Lock-free queue (without deallocations) and 2.3. Multi-counters deallocation technique.

Now the node of the queue is going to contain its *internal_counter* and *external_counters*. In the queue we also need several atomic variables, which point to the nodes: head, tail and next. Let the head, tail and next be objects, which contain *external_counter* and a pointer to the node.

Also, head, tail and next would be associated with three sources, which may produce reference copies. It is required for the correct usage of access protocol, because we would have to access head, tail and next pointers in our algorithm, and the only way to access data element is to duplicate its reference.

While where to use an access protocol and a cleaning condition is a rather simple question, we should consider when do we send nocopy promises.

For tail, we should send nocopy promise when tail is moved further than current data element. This implies that the current data element is never going to be a tail in the future, thus we may guarantee that reference on it cannot be copied from tail.

For next and head, we will send nocopy promise only when the data element is popped out of the queue. This would surely guarantee that no other head and next pointers will ever point to current data element in the future, and the corresponding references will never be copied.

The final implementation of Push operation:

- allocate new data
- allocate new dumb node
- L1: copy tail using CopyGuard (forcing the increase of *external_counter*)
- L2: try CAS tail.data (nullptr → allocated new data)

- L3: if L2 succeeded, try CAS tail.next (nullptr → new dumb node)
- if L3 failed, somebody already moved tail.next to new node, so deallocate new dumb node, try CAS tail (old_tail → value of tail.next). If succeeded, send nocopy promise to the node. In any case finish Push.
- if L3 succeeded, try CAS tail (old_tail → new dumb node). If succeeded, send nocopy promise to the node. In any case finish Push.
- L4: if L2 failed, we do not have dumb node in the end of the queue, try CAS tail.next (nullptr → new dumb node)
- if L4 succeeded, we shall allocate new dumb node, try CAS tail (old_tail → value of tail.next). If succeeded, send nocopy promise to the node. In any case repeat algorithm from L1 point (CopyGuard destructor is called implicitly in the end of the cycle iteration, forcing the decrease of *internal_counter*)
- if L4 failed, try CAS tail (old_tail → value of tail.next). If succeeded, send nocopy promise to the node. In any case repeat algorithm from L1 point using the same already-allocated new dumb node (CopyGuard destructor will be called implicitly)

The final implementation of Pop operation:

- copy head using CopyGuard
- L1: check head == tail
- copy head.next using CopyGuard
- if L1 is true, return empty-unique_ptr
- L2: if L1 is false, do CAS head (current → next)
- if L2 failed, retry from the start (CopyGuard destructors for head and next are called implicitly)
- if L2 succeeded, store and exchange data from old head
- trying until success to do CAS next (current_value → nullptr_special). It cannot hang for an infinite time, because, since the head have already been moved, no new calls to Pop will deal with this “next” pointer of popped head. This means, that we have to wait only finite amount of calls in another threads, which accidentally read same head to be popped, to finish. Threads, which have accidentally read the same head to be popped, would face L2 operation failure, and immediately after that would go to the start of algorithm, where they would reload another head. So, the threads, which may access the same popped head, would not be blocked, they would just go out of the way of the current thread and let it change “next” pointer. Threads, which do Pop, may also want to access current vertex in case if the queue is empty and tail == head. But they also cannot be blocked, because there is only one cycle in the Pop operation, which rereads data at the start, and the next iterations would not see already-popped head and would not interact with the “next” pointer in question. As all interactions with “next” would fade away after finite number of steps in finite number of threads, we would be able to do CAS successfully. That is why it is lock-free to wait until CAS next (current_value → nullptr_special) actually succeeds.
- after that, send nocopy promise of type “next” about the next element of the popped head (because it will never be someone’s else next element, although may still be inside the queue)

- send nocopy promise of type “head” about the popped head (because it will never be a head again)
- return data from popped head to the caller (CopyGuard destructors are called implicitly)

What is specific about Pop operation is this weird `nullptr_special` we used in CAS next. The matter of the fact is that to guarantee strictly that the node, contained in the “next” pointer, would never be copied, we have to assign “next” pointer to something else. Exactly this assignment, as have already been explained, waits for other threads, which want to operate with the same “next” pointer, to stop making “next” reference copies. Still there is a field for bug in the code. If we assign next to `nullptr`, we may open the ability for Push operation to change this “next” pointer, because Push operation does CAS next (`nullptr` → new dumb node). The Push operation is based on the idea that the only node, which has “next” pointer being equal to `nullptr`, is the tail node. By assigning `nullptr` to “next” pointer of the popped head, we corrupt this invariant and Push may attach new node to the popped node, which is totally unacceptable. Thus, if we want to change “next” pointer value, we cannot use `nullptr` for this purpose. But, we may use another special pointer value, which we are sure that nobody else uses. This is what is called a `nullptr_special` here.

What is remained is to deallocate the rest of nodes of the queue in destructor. This is done by calling Pop a lot of times in the destructor (since Pop is `noexcept`, we may freely do so). Afterwards, the only dumb vertex remained, which can be either deallocated directly, or we may send nocopy promises from head and tail and it would also be deallocated automatically by three-counters logic.

To summarize, we designed Push and Pop operations in such a way, that

- all data allocated will at some point be deallocated
- all accessed data in each thread is protected and would not be deleted while read
- Push can throw, strong exception safety is still guaranteed for the reason that reference counting itself is exception-free and all guarantees from lock-free queue are preserved
- Pop is exception-free

2.5. Multi-counters lock-free queue (productivity)

The algorithm of queue appears to be so complex, that one may wonder: is it really faster than the queue, protected by mutex? Let us do a couple of tests to show the difference between these two implementations.

Imagine we have 4-processor hardware. Imagine there are two threads, one of which wants to send 10 million Pop requests, another wants to send 10 million Push requests.

In case of lock-free queue the result is the following:

```
real    0m6.079s
user    0m11.314s
sys     0m0.077s
real RPS = 3.33 million
```

In case of mutex-protected queue the result is the following:

```
real    0m5.500s
user    0m6.980s
sys     0m3.146s
real RPS = 3.63 million
```

One may think that it is not so impressive, because lock-free implementation was 0.5 seconds slower (real time). But a closer look will reveal, that

- mutex-protected queue takes a lot of system time. Of course, it does, because a lot of time is required to synchronize very-fast-changing values of mutex bool flag. Because the threads compete for the ownership of the mutex, cache ping-pong is very likely to happen, forcing queue to be slower
- user time of lock-free queue is even bigger, 11.3 seconds, which is bad comparing with 6.9 seconds of mutex-protected version. But, from a scalability point of view, the lock-free implementation is far better. Using two threads, we gain almost ideal ratio of 11.3 seconds of user time and 6 seconds of real time. As for mutex-protected queue, we get almost no scalability, because each operation is protected with a mutex. Of course, this leads to a threads serialization, and we gain 6.9 seconds of user time resulting in 5.5 seconds of real time, which is a bad result for two-threaded program.

So, lock-free queue is slower, because it contains a lot more logic in its implementation. But, it does not spend so much time on system calls and is ideally scalable.

Let us watch what is going to be if there are two threads, which try to Pop 10 million elements, and also two threads, which try to Push 10 million elements (4 threads in total).

In case of lock-free queue:

```
real    0m10.082s
user    0m37.448s
sys     0m0.949s
real RPS = 4.0 million
```

In case of mutex-protected queue:

```
real    0m17.316s
user    0m23.260s
sys     0m25.856s
real RPS = 2.3 million
```

So here we see the drawbacks of mutex-protected queue in action:

- the system time increased dramatically, because now 4 threads are competing for single mutex, and its value needs to be synchronized to 4 processors instead of 2
- the scalability is still poor: we get 17.3 seconds of real time having 23.2 seconds of user time.
- mutex-protected queue degrades in RPS after we increased threads count: from 3.6 million queries per second to 2.3 million. That is quite a bad degradation.

And again we clearly see the advantages of lock-free queue:

- it wastes less than a second of system time, that is, almost 26 times less than mutex-protected queue. This means, it would not go slower if there would be a lot of other processes doing system calls
- it is still almost perfectly scalable: we get 10 seconds of real time having 37.4 seconds of user time, which is a good result for a 4-threaded application
- the user time of lock-free implementation still exceeds the user-time of mutex-protected queue (the fact that lock-free queue needs more time to execute its logic cannot be changed), but, due to scalability, real time shows that now, with 4 threads, it is almost 1.5 times faster than a mutex-protected implementation

- lock-free queue does not suffer from RPS degradation while we scale concurrency. It even became faster: from 3.33 million requests per second it increased up to 4 million requests per second. That is a very good result.

On the account of these simple measurements, we decided that it is better to use a lock-free algorithm.

2.6. Low-overhead periodic timer

In order to make the system more efficient, sometimes it is required to measure operations in order to distribute them ideally between different threads. However, a lot of calls to system timer would result in slower performance.

The purpose of the current section is to establish easy-to-use time measuring algorithm, which will somehow not slow down the performance being called a lot of times in a short period of time. To simplify matters, let us construct a timer, which has *GetPassedTime* function, which being called the first time, would return unix timestamp (time since year 1970), and being called further would return the time passed since its last call. Let us also have a *Reset* method, which resets timer settings to default ones. This timer will not be thread-safe, it will be designed to be used in one thread only.

Basically, the idea is simple: if we cannot access system clock on each call, we should access them more rarely, thus sometimes we will provide approximate results on *GetPassedTime* call. On the contrary, sometimes on *GetPassedTime* call we will call system clock and catch up or slow down depending on our previous not exact responses. That is, our not-so-exact timer needs to be wound up when we do access system clock.

In order to do that let us introduce what information does such a timer store:

- static constant *min_system_clock_call_period* which is a minimal desired time, after which we want to call *WindUp*
- static constant *max_wind_up_steps* which tells how many calls of *GetPassedTime* at maximum should be until next wind-up
- static constant *min_measured_time* is a minimal value we would return from *GetPassedTime* function
- *wind_up_counter* is a counter, which decreases on each *GetPassedTime* call and being equal zero forces winding up process
- *wind_up_balance* is a difference between real time and sum of all times returned from *GetPassedTime* during program execution
- *last_wind_up_counter* stores the value of counter, which was set just after last wind-up
- *last_system_time* is system time measured while last wind up
- *approx_time_step* is an approximate value, which is a prognosis of time elapsed since previous *GetPassedTime* call

The initial values of parameters would be the following:

- *wind_up_counter* = 1
- *wind_up_balance* = 0
- *last_wind_up_counter* = 1
- *last_system_time* = 0
- *approx_time_step* will be initialized on first *GetPassedTime* call, as it will be seen further

The *Reset* procedure would reset parameters to their initial values.

The *WindUp* procedure would do the following:

- store current system time in *system_time* variable
- calculate $time_passed = system_time - last_system_time$
- calculate $new_wind_up_counter = \max(1, \min(min_system_clock_call_period/time_passed * last_wind_up_counter, max_wind_up_steps))$. That is, if *time_passed* is less than *min_system_clock_call_period*, we want to increase amount of calls we want to perform until next win-up by the ratio of these values. But, we do not want it to be more than *max_wind_up_steps* and less than 1.
- update $approx_time_step = time_passed/last_wind_up_counter + wind_up_balance/new_wind_up_counter$
- update $last_wind_up_counter = new_wind_up_counter$ to store new last counter
- update $wind_up_counter = last_wind_up_counter$ to do next wind up after this amount of steps
- update $last_system_time = system_time$ to save result for future wind-up
- update $wind_up_balance+ = time_passed$ in order to maintain *wind_up_balance* invariant

Now let us consider the following algorithm for *GetPassedTime* operation:

- decrease *wind_up_counter*
- if *wind_up_counter* equals zero, call *WindUp* method
- calculate $result_time = \max(min_measured_time, approx_time_step)$
- update $wind_up_balance- = result_time$ to maintain *wind_up_balance* invariant
- return *result_time*

During the first *GetPassedTime*, *WindUp* would be called immediately and will initialize *approx_time_step*, which has already been stated. And the initialization of *approx_time_step* appears before its first usage.

Why is this supposed to be correct? First of all, imagine a situation when we call *GetPassedTime* so rarely, that *time_passed* is large enough to be more than *min_system_clock_call_period* and also more than *min_measured_time*, and *last_wind_up_counter* equals 1 and *wind_up_balance* equals zero. Let us trace how parameters will be updated in *WindUp* method:

- store current system time in *system_time* variable
- calculate $time_passed = system_time - last_system_time$
- update $new_wind_up_counter = \max(1, \min(min_system_clock_call_period/time_passed * last_wind_up_counter, max_wind_up_steps)) = \max(1, \min(0, max_wind_up_steps)) = \max(1, 0) = 1$.
- update $approx_time_step = time_passed/last_wind_up_counter + wind_up_balance/new_wind_up_counter = time_passed/1 + 0/1 = time_passed$

- update $last_wind_up_counter = new_wind_up_counter = 1$
- update $wind_up_counter = last_wind_up_counter = 1$ so *WindUp* would be called on next *GetPassedTime* call
- update $last_system_time = system_time$ to save result for future wind-up
- update $wind_up_balance+ = time_passed$ so $wind_up_balance$ would equal $time_passed$ now

This way, $last_wind_up_counter$ continues to be equal 1 and next time the same calculations would also hold. Now let us consider what will happen in *GetPassedTime* procedure:

- decrease $wind_up_counter$
- $wind_up_counter$ equals zero because it was being equal 1 before, so we definitely call *WindUp* method
- $wind_up_balance$ equals $time_passed$, as well as $approx_time_step$ does
- calculate $result_time = \max(min_measured_time, approx_time_step) = \max(min_measured_time, time_passed) = time_passed$, because we assumed that $time_passed \geq min_measured_time$
- update $wind_up_balance- = result_time$, which makes $wind_up_balance$ being equal zero again.
- return $result_time$, which is equal to $time_passed$

Thus, all values we assumed at the start ($last_wind_up_counter = 1$, $wind_up_balance = 0$) would still remain the same after *GetPassedTime* call. That is why, the same will repeat on the next *GetPassedTime* call and it will always return correctly calculated precise $time_passed$ result on each call. That is, in case of large enough $time_passed$, our timer is equivalent to simple timer logic.

Another important point is that in any case we are trying to make $wind_up_balance$ being equal zero. Indeed, we set $approx_time_step = time_passed/last_wind_up_counter + wind_up_balance/new_wind_up_counter$. Expression $time_passed/last_wind_up_counter$ evaluates to average time between *GetPassedTime* previous calls. Expression $wind_up_balance/new_wind_up_counter$ results in the fact, that during next $new_wind_up_counter$ calls of *GetPassedTime* we will, apart from main logic, decrease previous error $wind_up_balance$ to zero while updating $wind_up_balance- = result_time$. Simultaneously, a new error, which the prognosis $time_passed/last_wind_up_counter$ poses, will also be accumulated while doing $wind_up_balance- = result_time$, and it will be handled during the next *WindUp* call, while we will count new $approx_time_step$.

$wind_up_balance$ parameter is rather important. It shows our overall accumulated calculation errors and if it is a large negative number, it implies that the sum of returned times from *GetPassedTime* dramatically exceeded real time pace, and $approx_time_step$ would be negative in such a case, forcing $result_time$ we give out to be equal $min_measured_time$. In case if $wind_up_balance$ is a rather big positive number, we will return larger values from *GetPassedTime* because of larger $approx_time_step$, which would help us to catch up with the time pace.

Notice that *GetPassedTime* is designed to be called exactly from one location in the code, because we use the fact that approximately same time passes between *GetPassedTime* calls while calculating $time_passed/last_wind_up_counter$ as an expected virtual time step.

2.7. Low-overhead start-finish timer

Let us construct a timer, which would help to measure some code execution time. It would have *Start* and *Finish* methods, which are to be called before and after desired code block. Also it would have *GetCount* method, which will return count of measurements, and *GetDurationSum* method, which would return sum of all measured times. Moreover, there will be a *Reset* method, which resets counters.

In order to avoid slow performance due to a lot of system timer calls, we will use [2.6. Low-overhead periodic timer](#).

Our timer object would store the following information:

- *start_timer* which is a periodic timer associated with start event
- *finish_timer* which is a periodic timer associated with finish event
- *measurements_counter* which stores count of measurements
- *duration_sum* which stores sum of all durations measured so far

Now let us consider, what methods do.

Algorithm for *Start* method:

- decrease *duration_sum* by *start_timer.GetPassedTime()* (method *GetPassedTime* is defined in the [2.6. Low-overhead periodic timer](#) section)

Algorithm for *Finish* method

- increase *duration_sum* by *finish_timer.GetPassedTime()*
- increase *measurements_counter*

GetCount method returns *measurements_counter*

GetDurationSum method returns *max(PeriodicTimer :: min_measured_time, duration_sum)*.

Reset method reinitializes all parameters *measurements_counter* and *duration_sum* and calls *Reset* method of timers *start_timer* and *finish_timer*.

The implementation here is very simple due to the fact that we have already passed through all the difficulties while building [2.6. Low-overhead periodic timer](#) algorithm.

As far as [2.6. Low-overhead periodic timer](#) has a unique caller requirement (must be called from one specific location in the code), our *Start* and *Finish* methods also have the same requirement.

2.8. Low-overhead waiting timer

Here we are aiming to create a timer, which helps to schedule periodic actions. In constructor it accepts a time period value. It has *CheckTime* function, which returns *true* if given time passed, and *false*, if time have not passed yet. Also there would be a *Reset* function, which resets waiting process from the start.

As far as there can be a lot of *CheckTime* calls in a short period of time, the good idea would be to use [2.6. Low-overhead periodic timer](#).

Our *WaitingTimer* would contain the following variables:

- *time_period* which is set by user in *WaitingTimer* constructor
- *time_elapsed* which is time passed after last *Reset* call
- *periodic_timer*

Reset method has the following logic:

- call *periodic_timer.GetPassedTime()* to fix point since which we are going to trace time

- update *time_elapsed* = 0

Reset should be called in the constructor of the class *WaitingTimer*.
Function *CheckTime* would have the following logic:

- update *time_elapsed* += *periodic_timer.GetPassedTime()*
- return evaluated condition *time_elapsed* > *time_period*

Now our method *CheckTime* acts if it was planned. However, it bears a unique caller requirement as far as [2.6. Low-overhead periodic timer](#) does.

2.9. Low-overhead periodic clock

PeriodicClock would be a class, which is called from one specific location in the code and helps to get approximate current time without doing system call each time.

It is based on [2.6. Low-overhead periodic timer](#)

It would contain the following data:

- *current_time*
- *periodic_timer* of class *PeriodicTimer*

It would have the only method *GetTime*, which does the following:

- increments *current_time* by value of *periodic_timer.GetPassedTime()*
- returns *current_time*

2.10. Exception top proto storage

In order to make it simple for users to administrate the system, let us somehow notify users about exception thrown. Here we face a problem if some exception is thrown a lot of times, it becomes hard to distinguish it from other exceptions. Hence, it is better to store the same exception once with its appearance count. To avoid a lot of memory spent on storing exceptions, let us store only limited top of ones, since it is highly unlikely that any user would ever need to examine over thousand distinct exceptions without fixing the most frequent ones at first.

Let us introduce class *ExceptionTopKeeper*, which would help to update top of exceptions with their frequencies, save it on disk and restore it from disk. Our *ExceptionTopKeeper* would not be thread-safe and has to be instantiated inside TLS. That is, each thread would have its own exceptions top and a specific place to store it.

ExceptionTopKeeper contains the following data:

- variable *path*, which is a path of a file where exceptions are stored
- hashmap *exception_frequency* with key being a string of exception and value being equal exception frequency
- hashmap *exception_last_time* with exception as key and its last appearance time as a value
- set *frequency_exception* of pairs (*frequency*, *exception*) needed in order to maintain sorted order of exceptions by their frequencies
- set *last_time_exception* of pairs (*last_time*, *exception*) needed to maintain sorted order of exceptions by their last appearance time
- *dump_timer* of type *WaitingTimer* (for details see [2.8. Low-overhead waiting timer](#)), which is a timer telling when to *Dump* exceptions on disk

- *store_clock* of a class *PeriodicClock* (for details see [2.9. Low-overhead periodic clock](#)), which helps to get time of storing exception
- static constant *top_size*, which is a maximal amount of exceptions with counters stored
- static constant *keep_time*, which stores a period of time, after which exception disappears
- static boolean constant *debug_mode*, which being *true* implies we call terminate on any exception thrown, and being *false* implies we do not terminate, but rather store exception reason

It would have the following methods:

- *Dump* method, which dumps exception data to file
- *Restore* method, which restores exception data from file
- *Trim* deletes exceptions stored if amount of different exceptions is more than *top_size*
- *RemoveOld(current_time)* deletes exceptions older than $current_time - keep_time$
- *SetPath(new_path)* which changes *path* value to the one user wants
- *WithCatchingException(callback)*, which calls *callback* either with or without catching exceptions, which is defined by *debug_mode*. In case if exception is caught, it is stored in the exception top

Dump has the following algorithm:

- if *dump_timer.CheckTime()* returns *false*, do nothing
- start try block
- convert data of variables *exception_frequency*, *exception_last_time*, *last_time_exception* and *frequency_exception* to proto class
- dump proto to a separate file
- move separate file to *path*
- call *dump_timer.Reset()*
- if any exception caught, ignore them

Replacing file by another moved file is safe even when some other process or thread is reading it.

Dump method absorbs its own exceptions and thus does not throw.

Restore has the following algorithm:

- read proto from *path* file
- fill inner variables *exception_frequency*, *exception_last_time*, *last_time_exception* and *frequency_exception*

Restore may throw, no guarantees provided, caller should terminate in case of *Restore* failure.

Trim has the following algorithm:

- while size of *frequency_exception* set is larger than *top_size*, delete the least pair
- also delete it from *exception_frequency* and *exception_last_time*, *last_time_exception*

Trim does not throw as far as it calls only erase operations.

RemoveOld(current_time) has the following also:

- if *last_time_exception* is empty, do nothing
- start an infinite cycle A
- take first element (*last_time*, *exception*) of *last_time_exception*
- if *last_time* is more than *current_time* – *keep_time*, return from function
- remove *exception* from all containers *exception_frequency*, *frequency_exception*, *exception_last_time* and *last_time_exception*

RemoveOld does not throw, because it calls only erasing functions from set and map.

SetPath has the following algorithm:

- set *path* to its new value
- create a file, which has path *path*
- call *Restore* method

SetPath would be marked as noexcept and terminate on creating file failure or *Restore* failure.

WithCatchingException(callback) has the following algorithm:

- if *debug_mode* is *true*, call *callback* and return
- assert *path* is not empty string (that is, *SetPath* has been called before)
- try to catch exceptions while calling *callback*
- store in *exception* variable a string containing the exception text (if string construction threw, return doing nothing)
- initialize *prev_count* with zero
- if *exception* is found inside *exception_frequency* hashmap, assign *prev_count* to its frequency
- initialize *prev_time* with zero
- if *exception* is found inside *exception_last_time* hashmap, assign *prev_time* to its last appearance time
- store *current_time* = *store_clock.GetTime()*
- start try block A
- add pair (*prev_count* + 1, *exception*) to *frequency_exception*
- add pair (*current_time*, *exception*) to *last_time_exception*
- replace value of *exception_frequency*[*exception*] with *prev_count* + 1
- replace value of *exception_last_time*[*exception*] with *current_time*
- if an exception is caught in try block A, delete (*prev_count* + 1, *exception*) pair from *frequency_exception*, delete (*current_time*, *exception*) pair from *last_time_exception* and return from function doing nothing. This would help to preserve invariants on inner exception.

- if *prev_count* is more than zero, remove pair (*prev_count*, *exception*) from *frequency_exception* set
- if *prev_time* is more than zero, remove pair (*prev_time*, *exception*) from *last_time_exception* set
- call *RemoveOld(current_time)*
- call *Trim*
- call *Dump*

AddException cannot throw, it absorbs all its own exceptions and in case if any exception handled, *AddException* may have no effect. Still, that is not a problem for the reason that the functionality of *AddException* is not so important, thus we do not need to provide strict guarantees. Even if we wanted to provide some guarantees, we would fail for the reason that the function is forced to be *noexcept*, because it is handling some already happened. Therefore, if other exception arises, we could have gone into infinite recursion handling it the same way.

2.11. Shadow-paging thread-wise sharder

Sometimes it is needed to distribute some job between threads. Here we assume that we have a one big job, which can be represented as a composition of several small operations indexed by integer numbers, which we will call “shards”. Here we are aiming to going further and to introduce resharding process, which helps to adjust to circumstances in runtime.

A good idea would be to separate the logic part (which does job shards and knows how to reshard to make job faster) and the infrastructural part (which is about a thread-safe logic of how to organize resharding).

The logic part is going to be processed by any user-provided class *Controller*, and an infrastructural part is going to be handled by *Sharder < Controller >* class.

Controller class would have the following methods:

- *ProcessShard(can_be_updated, shard_num)* which processes actions of shard *shard_num* (*can_be_updated* specifies whether *Controller* can update information of shard, which is used for resharding)
- *PreProcess(thread_num, can_be_updated)* which is called from *Sharder* before processing shards
- *OnSwitch(thread_num, new_shards)* which is called by *Sharder* when a thread started to work with new shards after resharding. A new shards vector is passed to it.
- *GetInitialSharding(thread_count)*, which returns vector of vector of numbers of shards
- *GetResharding(old_shard_configuration, thread_count)*, where *old_shard_configuration* is a current vector of vector of shards, and the function should return new vector of vector of shards
- *GetMaxThreadCount* returns maximum amount of threads needed

Implementation of *Controller* class is user-specific, since that we will not discuss any possible implementation of *Controller* in this section.

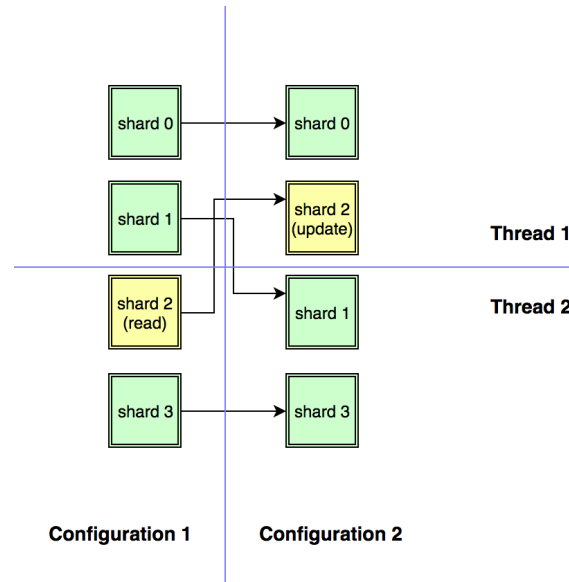
Sharder < Controller > has the following responsibilities:

- guaranteeing read-write lock on shards data. That is, if some shard data is being modified, nobody should read it at this point
- assigning a group of shards to each thread

- doing *Reshard* process periodically by forming new configuration of sharding by means of shadow paging
- call *ProcessShard*, *PreProcess* and *OnSwitch* functions to process shards actions

To avoid several race condition problems, it would be much easier to allow *Reshard* call only from main thread. Still there are some difficulties remained. We are going to do *Reshard* by looking at current state of shards and deciding how to redistribute them. However, while *Reshard* operation is reading shards info, it can be updated by any of the other threads. This poses a risk of data race condition. In order to avoid it, let us introduce *NoUpdatePromise* method. Such a method would help to establish *Reshard* basic requirement: resharding is possible if and only if all threads stopped to update shards data and called *NoUpdatePromise*. After that, we may safely read current shards data and decide how to reshard them. Afterwards, all threads should start to use new sharding configuration. And, if there would be a need in another resharding, all threads must again promise not to update current state of shards.

More than that, there is another race condition. After resharding, threads are going to switch to another sharding configuration, and during switching it is highly likely that thread, which finishes to use old sharding configuration, and a thread, which has already started to use new sharding configuration, may read-write the same shard data:



In the picture above we may see two configurations. Old configuration shows that shards 0 and 1 belong to thread 1 and shards 2 and 3 belong to thread 2. Next resharded configuration show that shards 0 and 2 belong to thread 1, shards 1 and 3 belong to thread 2. While thread 2 is still using old configuration 1, thread 1 may have already switched to configuration 2. In this very moment thread 1 may start to write updates to shard 2 data, while thread 2 is still reading shard 2 data, which is a data race.

In order to avoid this, we will use mutexes. Mutexes will be locked and released only during *Reshard* process, which would take place rarely, that is why we do not need to bother about mutexes being slow. More precisely, each shard would be associated with a single mutex. When switching to new sharding configuration, thread releases all the mutexes associated with shards of old configuration, afterwards acquiring all mutexes associated with shards of new configuration.

In the example above, such improvement would result in the fact that thread 1 cannot switch to the new configuration, because it waits mutex associated with shard 2, which is held by thread 2 processing old configuration. Only after thread 2 stops to work with an old configuration, thread 1 may freely start to work with new one.

Mutexes associated with shards help us to solve data races different threads may create. However, another possible problem arises, which is a deadlock. Fortunately, we do not suffer from this problem. Indeed, the only moment when a thread A waits is a point when it tries to lock mutex associated with some specific shard during configuration changing, that is, some other thread B locked that mutex. If thread B has already switched to new sharding configuration, it cannot lock the same shard as thread A, thus such condition is impossible. If thread B still uses old configuration, it is going to switch to a new one and release its locks, stopping to block thread A. That is, we will never have a case of a deadlock.

After coping with algorithmic problems, let us turn to the implementation issues.

Sharder < *Controller* > contains the following variables:

- *controller* of type *Controller*&
- *first_conf* is a vector of vector of shards
- *second_conf* is a vector of vector of shards
- *can_be_updated* (TLS) boolean variable, which specifies if shard data can be modified by *controller* object
- *is_first_local* (TLS) boolean variable, which indicates which configuration a thread is using: *first_conf* or *second_conf*
- *thread_num* (TLS), which is a current thread number
- *reshard_waiting_timer* (TLS) is a *WaitingTimer* instance needed to measure when next resharding is required (see [2.8. Low-overhead waiting timer](#))
- *threads_count*, which would be initialized on construction
- atomic pair (*is_first_main*, *noupdate_counter*). *is_first_main* indicates whether current active configuration is stored in *first_conf* or in *second_conf*, while *noupdate_counter* shows how many threads promised not to update shards data.
- *threads* is a vector of threads
- *shard_locks* is a vector of mutexes

Data of *Sharder* is initialized the following way:

- *controller* is initialized by reference, passed to constructor
- *threads_count* is initialized as minimum of two values: maximum hardware concurrency and *controller.GetMaxThreadCount()*
- (*is_first_main*, *noupdate_counter*) = (*true*, 0)
- *first_conf* is initialized by call to *controller.GetInitialSharding(threads_count)*
- *second_conf* is initialized by default
- *threads* is initialized as an empty vector
- *shard_locks* is initialized as a vector of size being equal to shards count
- for each *thread_num* call *controller.OnSwitch(thread_num, first_conf[thread_num])*

Sharder would have the following methods:

- *GetConf* accepts boolean and returns *first_conf* if *true* and *second_conf* otherwise
- *Reshard*

- *NoUpdatePromise*
- *GetShards*
- *SwitchConfiguration*
- *ThreadAction*
- *Run*

Reshard method has the following algorithm:

- if it is called not from main thread, do nothing
- read *(is_first_main, nouupdate_counter)* pair
- if *nouupdate_counter* is less than *threads_count*, do nothing. Otherwise, there is a need for resharding
- call *controller.GetResharding(GetConf(is_first_main), thread_count)* catching exceptions
- if no exceptions caught, assign result to *GetConf(!is_first_main)*, return doing nothing otherwise
- try CAS *(is_first_main, nouupdate_counter) ((cur_first_main, threads_count) → (!cur_first_main, 0))* and do nothing on CAS failure
- call *reshard_waiting_timer.Reset()* on CAS success

NoUpdatePromise method

- sets *can_be_updated* to *false*.
- tries to increment *nouupdate_counter* using CAS and re-reading pair *(is_first_main, nouupdate_counter)* in the cycle until CAS succeeds or until after some of reads we find that *is_first_main* is not the same as *is_first_local*.

GetShards would have the following logic:

- check if *is_first_local* is the same as *is_first_main*. If so, we do not need to switch to a new configuration. Otherwise, we need to call *SwitchConfiguration* method
- return *GetConf(is_first_local)[thread_num]*

SwitchConfiguration procedure does the following:

- call *reshard_waiting_timer.Reset()*
- for all shards in *GetConf(is_first_local)[thread_num]* call *unlock* for the corresponding mutex from *shard_locks* vector
- for all shards in *GetConf(is_first_main)[thread_num]* call *lock* for the corresponding mutex from *shard_locks* vector
- update *is_first_local = is_first_main*
- set *can_be_updated* to *true*.
- call *controller.OnSwitch(thread_num, GetConf(is_first_main)[thread_num])*

ThreadAction is a function which would be assigned to threads. It accepts *thread_num* and does the following:

- set TLS variable *thread_num* to the value it accepted from caller
- call *exception_top_keeper.SetPath(...thread_num...)* passing some filename, which contains *thread_num* (for details refer [3.7. Exception top proto storage interface](#))
- start an infinite cycle
- call *exception_top_keeper.WithCatchingException* passing it lambda function with the forthcoming logic:
- call *GetShards* and store value to *shards* variable
- call *controller.PreProcess(thread_num, can_be_updated)*
- for each shard in *shards* call *controller.ProcessShard(can_be_updated, shard_num)*
- if *reshard_waiting_timer.CheckTime()* returns *true*, time for resharding has come. We have to call *NoUpdatePromise* and *Reshard*.

Since *ThreadAction* calls only functions *SetPath* and *WithCatchingException*, which do not throw, it also does not throw.

Finally, *Run* method does the following:

- initializes *threads* vector with *thread_count* – 1 threads associated with *ThreadAction*
- calls *ThreadAction* in main thread also.

Let us also outline the algorithmic and exception requirements to *controller*.

First of all, the algorithmic requirements:

- *ProcessShard(can_be_updated, shard_num)* should manipulate with external not-thread-safe data only if a data is uniquely specified by *shard_num* and there is no intersection between different *shard_num*'s data. Function cannot update such data in case if *can_be_updated* is *false*.
- *PreProcess* cannot manipulate with external not-thread-safe data at all, because it is not protected with mutexes.
- *GetResharding* can safely read all data associated with shards, which was updated during *ProcessShard* calls. It is also safe to assume, that no other thread calls *OnSwitch* until *GetResharding* finishes. That is because before *GetResharding* call all threads promised not to update current shards info, which means that they are already using current shard configuration and not going to switch until CAS, which goes after *GetResharding*.
- while implementing *OnSwitch(thread_num, new_shards)* it is safe to assume, that each shard of *new_shards* is not processed by any thread until *OnSwitch* completion. That is because before calling *OnSwitch*, we locked the shards in the current thread, but have not started to process them yet. It is also safe to assume, that main thread is not processing function *GetResharding* until *OnSwitch* completion, for the reason that otherwise *OnSwitch* is processed at the same time as *GetResharding* is processing, which is impossible as it has been stated above.

The exception requirements:

- *ProcessShard* may throw, should provide at least basic exception safety
- *PreProcess* may throw, should provide at least basic exception safety
- *OnSwitch* should not throw

- *GetInitialSharding* may throw, should provide basic exception safety (however, any reasonable implementation would not throw anyway)
- *GetResharding* may throw, should provide at least basic exception safety
- *GetMaxThreadCount* may throw, should provide at least basic exception safety

2.12. Dynamically-sharded message passing pool

In the section 2.11. [Shadow-paging thread-wise sharder](#) we have discussed how to shard a job. However, what remained uncovered is how to separate some job into shards and how to do re-sharding process. In case of a uniform job, for example, calling function *Func* for each element of some vector, the question of how to shard a job is rather simple. When a job is complicated, it can usually be represented as several distinct logic blocks, some of which use results of some other logical blocks and together they form a full algorithm. That is called a “message passing”. We would have several distinct message processors, which can accept messages, process them, and send messages to another message processors.

In message passing, a shard would be associated with a message processor. All message processors can be called in different threads if we use thread-safe queue for communication between message processor (sending message processor may be processed in one thread, while receiving message processor may be processed in another thread, which is why we have to use thread-safe queue).

Let us use a queue from the section 2.4. [Multi-counters lock-free queue \(final algorithm\)](#).

For thread-safe processing of sharding mechanism, let us use 2.11. [Shadow-paging thread-wise sharder](#). It requires shards being associated with integer numbers, which is why we would need some technical coding pattern to map message processor class names into integer numbers. The full idea of how to cope with it is described in 1.1. [Function specifications cascade merging](#) and 1.2. [Reverse template instantiation](#) and a detailed implementation is discussed in 3.4. [Message passing tree](#).

Here we will assume we already know how to map message processor to its integer index and vice versa.

What remains is to define *Controller* class, which is required in order to use 2.11. [Shadow-paging thread-wise sharder](#). Let this class be called *MessagePassingController*.

It would store the following information:

- *message_passing_tree* is an instance of *MessagePassingTree* class designed in the section 3.4. [Message passing tree](#). *MessagePassingTree* helps to match message passing queues and message processors with their indices.
- *message_wait_cvs* is a vector of condition variables. *message_wait_cvs* contains condition variables by a thread number. These condition variables would help to know when message passing queues are not empty. This vector is initialized in the constructor and never changes.
- *message_send_cvs* is a vector of pointers on same condition variables, indexed by outgoing message passing queues. It would help to notify condition variables on sending messages. The vector has to be rebuilt on each resharding.
- *are_queues_empty* (TLS) boolean variable which specifies whether incoming message queues have been empty on previous step
- static constant *wait_for_messages_time* which specifies how much time a thread waits for message passing queues not to be empty in case all queues was empty at the previous step
- vector *message_processor_timers* of *StartFinishTimer* instances for each message processor

- vector *edge_timers* of *StartFinishTimer* instances for each queue.

Now let us turn to the implementation of *Controller* required functions.
GetMaxThreadCount does the following:

- returns total number of message processors, which we can get using *message_passing_tree.GetMessageProcessorsCount()*

PreProcess(thread_num, can_be_updated) does the following:

- if *are_queues_empty* is true, waits for *wait_for_messages_time*, interrupting on *message_wait_cvs[thread_num]* condition variable notification. We will use dumb lock instead of mutex while waiting for condition variable as far as there would be nothing terrible if we miss a wake, because we wait for it a very limited amount of time. Condition variable is needed here merely to interrupt waiting earlier if possible.
- sets *are_queues_empty* to true

ProcessShard(can_be_updated, shard_num) does the following:

- if *can_be_updated* is true, calls *message_processor_timers[shard_num].Start()*
- using *message_passing_tree* calls *Ping* method for message processors with index *shard_num*
- if *can_be_updated* is true, calls *message_processor_timers[shard_num].Finish()*
- gets all incoming queues indices using *message_passing_tree.GetIncomingEdges(shard_num)*
- if *can_be_updated* is true for each index *queue_index* calls *edge_timers[queue_index].Start()*
- calls *message_passing_tree.GetEdgeProxy(queue_index).NotifyAboutMessage()*
- if *can_be_updated* is true calls *edge_timers[queue_index].Finish()*
- if any message passing queue was not empty, sets *are_queues_empty* to false

Additional helper function *SetSenderCVs* would accept vector of vector of shards and do the following:

- calculate max queue index by *message_passing_tree.GetEdgesCount()*
- for all indices *cur_queue_index* get receiver index by calling *message_passing_tree.GetEdgeProxy(cur_queue_index).GetToIndex()*
- find out a number of thread *thread_num*, which is going to process this receiver
- save pointer to condition variable *message_wait_cvs[thread_num]* to the location *message_send_cvs[cur_queue_index]*

GetInitialSharding(thread_count) does the following:

- gets a total number of message processors from *message_passing_tree.GetMessageProcessorsCount()* and distributes them among threads by modulo, returning required vector of vectors
- calls *SetSenderCVs* passing vector of vector of shards

OnSwitch(thread_num, new_shards) does the following:

- for each *new_shard* in *new_shards* calls *message_passing_tree.GetOutgoingEdges(new_shard)* to get outgoing queues indices
- for all outgoing queues sets condition variable of queue (by calling *SetConditionVariable* method) to *message_send_cvs[outgoing_queue_index]*. As far as no *ProcessShard* on any current message processor is processing, we may safely update condition variables of outgoing queues facing no data races.
- calls *message_processor_timers[new_shard].Reset()*
- gets all incoming edges indices by call to *message_passing_tree.GetIncomingEdges(new_shard)*
- calls *edge_timers[incoming_edge_index].Reset()*

In order to define *GetResharding* procedure, let us introduce several local variables it is going to use:

- *chosen_processors* is a vector of message processor indices chosen to be processed by current thread we consider
- *chosen_duration* is a sum of durations of *Ping* and *Receive* calls of chosen message processors
- *durations* is a set of pairs (*duration, message_processor_index*) where *duration* is a sum of durations of *Ping* and *Receive* calls of a corresponding message processor with index *message_processor_index*
- *passing_durations* is a hashmap, which by *message_processor_index* stores sum of durations of *Receive* methods, associated with edges connecting message processor with index *message_processor_index* with message processors from *chosen_processors*

Let us introduce helping procedure *ChoozeProcessor* which accepts all variables defined above as well as a pair (*duration, message_processor_index*) and is going to add a *message_processor_index* to *chosen_processors* and update all the variables listed the following way:

- increment *chosen_duration* by *duration*
- delete pair (*duration, message_processor_index*) from *durations*
- for each *other_message_processor_index* remained in *durations* pairs find index of edges between message processor with index *message_processor_index* and message processor with index *other_message_processor_index* by the call to *message_passing_tree.GetConnectingEdges*
- for each found *edge_index* increment *passing_durations[other_message_processor_index]* by the result of *edge_timers[edge_index].GetDurationSum()* call

GetResharding(shard_configuration, thread_count) does the following:

- calculates sum of all *GetDurationSum* values from times in *message_processor_timers* and in *edge_timers* as well
- divides sum by *thread_count* and saves result to local variable *avg_duration*
- initializes *durations* as set of pairs (*duration, message_processor_index*), where *duration* is sum of the result of *message_processor_timers[message_processor_index].GetDurationSum()* call and the result of *edge_timers[edge_index].GetDurationSum()* call for each *edge_index* of incoming edge

- starts a cycle A for each *thread_num* from 0 to *thread_count* - 1
- initializes *chosen_duration* = 0
- initializes *chosen_processors* as an empty vector
- initializes *passing_durations* is an empty hashmap
- if maximum duration in *durations* is more than *avg_duration*, calls *ChoozeProcessor* for corresponding pair in *durations* set.
- starts a while cycle B until *durations* is not empty
- by means of binary search finds in the set *durations* an upper bound of durations, which can be added to *chosen_duration* so that sum is not more than *avg_duration*. That is, we want to add message processors until sum of their execution durations does not exceed average wanted time.
- if there are no such durations in *durations* set, there is no message processor we can add to the current thread without exceeding average execution time. In this case, algorithm breaks cycle B.
- In the other case, among all suitable message processors with duration less than found upper bound, chooses a *message_processor*, which has the largest passing duration, which is detected by *passing_durations* hash map (absence of key in *passing_durations* map implies zero value here)
- calls *ChoozeProcessor*
- goes to the start of B cycle
- if got here, algorithm has nothing more to add to the current thread
- adds *chosen_processors* to vector of vector of shards we are going to return from this function
- goes to the start of A cycle
- calls *SetSenderCVs* for just built vector of vector of shards
- returns vector of vector of shards

Such algorithm helps to distribute computations evenly and also tries to combine message processors with high communication rate together, which helps to reduce cache ping-pong while two message processors are accessing the same message passing queue.

Let us check that we work with data correctly, without any races. In order to do that, we have to consider every variable we store:

- *message_passing_tree* is a proxy object made to access queues and message processors, it is not modified itself
- *message_wait_cvs* is not modified after initialization
- *message_send_cvs* is accessed in several functions: *SetSenderCVs* which is called inside resharding, also it is accessed in *OnSwitch*. Sharder guarantees *OnSwitch* never intersects with resharding in execution timeline.
- *are_queues_empty* is in TLS, thus no races are possible
- *wait_for_messages_time* is constant

- *message_processor_timers* is accessed in several functions: *ProcessShard* updates *shard_num* index, *OnSwitch* updates indices of switching shards, *GetResharding* reads the whole vector. In order to sort it out, let us consider all pairs of functions. First of all, sharder guarantees that shards passed to *OnSwitch* are not processed by *ProcessShard* at the moment of *OnSwitch* execution, which implies, that there is no race between *OnSwitch* and *ProcessShard*. Secondly, *ProcessShard* does not update information read by *GetResharding* because of *NoUpdatePromise* requirement. Finally, *OnSwitch* and *GetResharding* have not intersecting execution timelines, that is why they do not conflict either.
- *edge_timers* is accessed in several functions: *ProcessShard* updates indices, which are incoming queue indices for message processor *shard_num*, *OnSwitch* also accesses only indices of incoming queues of its shards, *GetResharding* accesses the whole vector. The same discourse as above would show that there are no data races.

As it can easily be concluded, we do not have data races in *MessagePassingController*.

The only question remained is whether it satisfies *Sharder* requirements on *Controller* class:

- *ProcessShard* manipulates with message processor indexed by *shard_num* and its incoming queues. It is obvious that these objects sets do not intersect for different shard numbers.
- *PreProcess* manipulates with thread-safe condition variable and TLS variable *are_queues_empty*, which is ok
- *OnSwitch* does not throw, it calls only noexcept functions
- the rest functions have basic exception safety

All the requirements of sharder are satisfied.

2.13. Increasing hex identifier

This is simply an algorithm for generating string identifiers of the form f12f-abc3-12a4. We want our *HexIdGenerator* to have increment method, which increments the identifier, and *GetHex* method, which returns generated string. Strings should increase lexicographically on each increment.

In order to do that let us introduce *HexToken* class, which contains the following fields:

- *index* is a 16-byte integer number
- *hex_repr* is a hex representation string

HexToken would have the following methods:

- *IsOver* indicates whether *index* reached its maximum, which is 65535.
- *UpdateHexRepr*, which updates the representation of *index* in hex string format and stores the result to *hex_repr* variable
- *GetHex*, which returns reference to string *hex_repr*
- *operator++*, which increments the *index* and calls *UpdateHexRepr* to update hex representation

Methods *IsOver* and *GetHex* would be exception-free. As for the rest, if *UpdateHexRepr* fails inside *operator++*, *operator++* would still be retryable, it would increment *index* again and call *UpdateHexRepr* on next retrying call. That is, in case of failure we may increment *index* by 2, missing *index + 1* representation construction, but this does not break the invariant of increment operation to produce lexicographically larger strings. That is, function provides basic exception safety guarantee.

Now let us consider *HexIdGenerator*, which stores the following information:

- *hex_tokens* vector of *HexToken* objects
- *hex_repr*, which stores full hex representation of the identifier

HexIdGenerator would have the following functions:

- *operator++*, which iterates the identifier
- *GetHex*, which returns reference of string *hex_repr*

operator++ would have the following algorithm:

- iterate through all *hex_token* from vector *hex_tokens*
- if *hex_token.IsOver()* returns *false*, call *++hex_token*, fix *hex_repr* by calling *hex_token.GetHex()* and placing new hex string to the place in whole *hex_repr* where it belongs. Afterwards, return from function
- if we reached to the end of the iteration through all *hex_tokens*, then we have an identifier of the form ffff-ffff-ffff..
- append 0000 to *hex_repr* if its length is not more, than $5 * \text{hex_tokens.size}() - 1$
- add new element to *hex_tokens* vector

GetHex method of *HexIdGenerator* would return reference to *hex_repr* and do nothing else. This method does not throw.

As for *operator++*, it may throw. There are several cases:

- *++hex_token* failed. In this case, the length of *hex_repr* remained unchanged. While it is unchanged, it consists of fixed amount of strings from *HexToken*, which are guaranteed to increase even in case of failures. That is, *hex_repr* also will be guaranteed to increase in this specific case
- appending 0000 to *hex_repr* failed. In this case, function has no effect
- adding new element to *hex_tokens* vector failed. In such a case, we would have a string of the form ffff-ffff-ffff-0000 and not updated vector of *hex_tokens*. On the next retry of *operator++*, it will again find, that all *hex_tokens* are over, but this time we will not append 0000 to the *hex_repr*, because of length check. Afterwards, function would finish appending element of *hex_tokens* and the structure would become consistent again.

That is, *operator++* provides basic exception safety.

2.14. Finite state machine

In order to build a consensus algorithm over several different hosts we have to design a FSM it is going to use to synchronize data it works with. Finite state machine would be represented as an object of some class, which works with so called write-ahead-log, which is a list of updates, which are going to be applied to the final state.

Our finite state machine is going to be used as part of [2.12. Dynamically-sharded message passing pool](#), that is why it has to be designed as a [3.3. Message processor](#)

Assume we already have *TState* class, which represents an object state, and a *TUpdate* class, which represents an update element, which can be applied to state by calling *update.Update(state)*. Also, update can be reverted by calling *update.Revert(state)*.

Let us introduce a class *FSMMP* < *TController*, *TState*, *TUpdate* >, which is inherited from *MessageProcessorBase*. *TController* represents another message processor type, which is controlling current state machine processor by sending messages.

FSMMP will store the following data:

- *write_ahead_log*, which is a list of *std::unique_ptr < TUpdate >* objects
- *cleanup_border*, which is a list iterator pointing at *write_ahead_log* element, up to which the WAL can be cleaned up
- *applied_border*, which is a list iterator pointing to the state, which is going to be applied next
- *state* of type *TState*

For the sake of the communication with *FSMMP* we would have to introduce several helping message classes:

- *FSMCommitMessage* is used for passing commit operation to clean up the WAL
- *FSMUpdatesMessage < TUpdate >* is used for passing updates to the WAL
- *GetFSMUpcomingUpdateMessage* is used to retrieve upcoming updates from the WAL
- *FSMUpcomingRevertMessage* is used to revert upcoming updates from WAL

FSMCommitMessage is a dumb class, it contains nothing.

FSMUpdatesMessage < TUpdate > contains the following information:

- *updates*, which is a list of *TUpdate* instances

GetFSMUpcomingUpdateMessage and *FSMUpcomingRevertMessage* are dumb classes.

The class *FSMMP* itself would have the following methods:

- *Cleanup*, which tries to clean the WAL
- *Comeup*, which tries to apply updates from *write_ahead_log*
- *Commit*, which marks queue elements to be deleted
- *AddUpdates(updates)* adds updates to the end of the WAL
- *Revert*, which reverts all upcoming updates from the WAL
- *Receive(ReceivingFrom < TController > (), commit, sender)*, which receives a commit message. *commit* is an instance of *FSMCommitMessage*.
- *Receive(ReceivingFrom < TController > (), add_updates, sender)*, which receives a message to add updates from controller. *add_updates* is an instance of *FSMUpdatesMessage*.
- *Receive(ReceivingFrom < TController > (), get_upcoming_update, sender)*, which requests to receive upcoming updates. *get_upcoming_updates* is an instance of *GetFSMUpcomingUpdateMessage*.
- *Receive(ReceivingFrom < TController > (), revert, sender)*, which requests to revert upcoming updates from the WAL. *revert* is an instance of *FSMUpcomingRevertMessage*.

The *Cleanup* method does the following:

- while *write_ahead_log.begin() != cleanup_border*, does *pop_front*

Cleanup method does not throw.

The *Comeup* method does the following:

- if *applied_border* equals *write_ahead_log.end()*, does nothing
- takes an update *applied_border* is pointing at

- calls *update* \rightarrow *Update(state)*
- increments *applied_border*
- calls *Cleanup*

Comeup method may throw in case of *Update* failure. For this reason, it requires *Update* method to be retryable on failure, that is, to provide basic exception safety. In such a case, *Comeup* method is also retryable and guarantees basic exception safety.

The *Commit* method does the following:

- assigns *cleanup_border* to *write_ahead_log.end()*

Commit method does not throw.

The *AddUpdates(updates)* method does the following:

- if *applied_border* equals *write_ahead_log.end()*, assigns it to *add_updates.begin()*
- if *cleanup_border* equals *write_ahead_log.end()*, assigns it to *add_updates.begin()*
- *write_ahead_log.splice(write_ahead_log.end(), std::move(add_updates.updates))*

AddUpdates does not throw.

The *Revert* method does the following:

- while *applied_border* is not equal to *write_ahead_log.begin()*
- stores to another variable *last_update* iterator, which precedes *applied_border*
- calls *last_update* \rightarrow *Revert(state)*
- assigns *applied_border* to *last_update*

Revert method requires *Revert* method of *last_update* to provide basic exception safety and be retryable. In such a case, it also guarantees basic exception safety and is also retryable.

Receive(ReceivingFrom < TController > (), commit, sender) method does the following:

- calls *Commit()*

This message processing does not throw.

Receive(ReceivingFrom < TController > (), add_updates, sender) method does the following:

- calls *AddUpdates(add_updates.updates)*
- calls *Commit()*

This message processing does not throw.

Receive(ReceivingFrom < TController > (), get_upcoming_update, sender) method does the following:

- creates empty *fsm_updates* of type *FSMUpdatesMessage*
- fills *fsm_updates* \rightarrow *updates* list with states from iterator *cleanup_border* to the end
- calls *sender.Send < TController > (updates)*

This message processing may throw on allocating new *fsm_updates*, and may throw in *sender.Send < TController > (updates)*. In either cases, method has no effect. This implies it guarantees strong exception safety.

Receive(ReceivingFrom < TController > (), revert, sender) method does the following:

- calls *Revert*

This message processing provides basic exception safety and is retryable.

3. Detailed Code Architecture (classes and methods)

3.1. TLS free-list multi-level allocator

The class implements the allocation logic described in the section 2.1. Free-list multi-level allocator.

```
class FreeListMultiLevelAllocator {
    FreeListMultiLevelAllocator()
    // Exceptions:
    //     may throw, strong exception safety

    FreeListMultiLevelAllocator(
        const FreeListMultiLevelAllocator&) = delete;

    FreeListMultiLevelAllocator(
        FreeListMultiLevelAllocator&&) = delete;

    FreeListMultiLevelAllocator& operator=(
        const FreeListMultiLevelAllocator&) = delete;
    // We will use object as s singleton, no copying and assigning

    template <typename T>
    T* Allocate(const size_t size);
    // Return value:
    //     a pointer to allocated block
    // Exceptions:
    //     may throw std::bad_alloc
    //     strong exception safety: no side effects in exception case

    template <typename T>
    void Deallocate(T* pointer, const size_t size) noexcept;
};

thread_local FreeListMultiLevelAllocator global_allocator;
// Memory allocations would be controlled by this TLS singleton

template <typename T>
class FixedFreeListMultiLevelAllocator<T> {
    FixedFreeListMultiLevelAllocator() noexcept;

    FixedFreeListMultiLevelAllocator(
        const FixedFreeListMultiLevelAllocator&) noexcept;

    template <class U>
    FixedFreeListMultiLevelAllocator(
        const FixedFreeListMultiLevelAllocator<U>&) noexcept;
    // These three methods are empty,
    //     they are needed for compatibility with std::allocator

    T* allocate(const size_t n);
    // Calls global_allocator.Allocate<T>(n)
    // Return value:
    //     a pointer to allocated block
    // Exceptions:
```

```

//      may throw std::bad_alloc
//      strong exception safety: no side effects in exception case

void deallocate(T* p, const size_t n) noexcept;
// Calls global_allocator.Deallocate<T>(p, n)
};

```

3.2. Lock-free queue

LockFreeQueue < *TElement* > is an implementation of lock-free queue described in the section 2.4. Multi-counters lock-free queue (final algorithm). It contains the following functions:

```

class LockFreeQueue<TElement> {
    void Push(TElement new_element);
    // Exceptions:
    //      Strong exception safety: in case of failure
    //      the element is not pushed to the queue,
    //      there are no visible side effects

    std::unique_ptr<TElement> Pop() noexcept;
    // Return value:
    //      nullptr unique_ptr if the queue was empty,
    //      unique_ptr pointing on TElement in case of successful pop
};

```

3.3. Message processor

Message processor is a class with *Ping* method, which is called periodically, and *Receive* method, which is called on receiving a message from another message processor. Both methods accept object *sender* of unspecified type and may use it in order to send messages to another message processors.

All message processors have base class *MessageProcessorBase*:

```

class MessageProcessorBase {
public:
    template <typename Sender>
    void Ping(const Sender&) {}
    // Empty method, needed if we do not want to define
    // corresponding Ping method in child class

    virtual ~MessageProcessorBase() noexcept {}
};

```

All message processors look like this:

```

class MessageProcessor : public MessageProcessorBase {
public:
    using MessageProcessorBase::MessageProcessorBase;

    template <typename Sender>
    void Ping(const Sender& sender) {
        // here a call to sender.template
        Send<YetAnotherMessageProcessor>(new AnotherMessageType())
    }
};

```



```

    // may be or may not be present
}
// Ping method may be absent here
// Exceptions: basic exception safety

template <typename Sender>
void Receive(const ReceivingFrom<AnotherMessageProcessor>&,
    const MessageType& message, const Sender& sender) {
    // here a call to sender.template
    Send<YetAnotherMessageProcessor>(new AnotherMessageType())
    // may be or may not be present
}
// Exceptions: basic exception safety
};

```

In here *ReceivingFrom* is a specification helper dumb class:

```

template <typename T>
class ReceivingFrom {};

```

This class is needed to specify several Receive methods in the *MessageProcessor* in order to specialize, from which *AnotherMessageProcessor* are we receiving a message.

MessageType is an arbitrary class, which stores the message itself.

Sender is a template parameter, which is needed for the reason that sender object is of an unspecified type.

3.4. Message passing tree

MessagePassingTree < *Edge* < *MessageProcessor1*, *MessageProcessor2*, *MessageType1* >, *Edge* < *MessageProcessor3*, *MessageProcessor4*, *MessageType2* >, ... > is a message-processor controller. It helps to register message passing communication edges and check are there unregistered send operations inside [message processors](#) at a compile time. Additionally, it stores message passing queues, which help to send messages from one message processor to another.

Moreover, this class provides an opportunity to access distinct message processors by index as well as access message queues by their indices. Furthermore, it can tell which indices of message queues are connected with which message processor indices.

Hence, this is a generic structure which controls message passing process and provides access by indices, which simplifies further programming.

The structure here is rather complicated and based on the [1.1. Function specifications cascade merging](#) and [1.2. Reverse template instantiation](#) patterns.

First of all, there is a zero-template very-base class *Piper* <> definition:

```

template <typename ... Args>
class Piper {
protected:
    template <typename GlobalPiper>
    class EdgeProxy {
    public:
        EdgeProxy(GlobalPiper& piper) noexcept;
        // Remembers piper reference

        virtual bool NotifyAboutMessage() const = 0;
        virtual void
            SetConditionVariable(std::condition_variable_any *) const

```

```

        noexcept = 0;
    virtual const int GetFromIndex() const noexcept = 0;
    virtual const int GetToIndex() const noexcept = 0;

    virtual ~EdgeProxy() noexcept {}
protected:
    GlobalPiper& piper;
};

template <typename GlobalPiper>
class MessageProcessorProxy {
public:
    MessageProcessorProxy(GlobalPiper& piper, const int
        message_processor_index) noexcept;
    // init piper and message_processor_index

    template <typename MP>
    MP& GetMessageProcessor() const noexcept;
    // Returns MessageProcessor instance
protected:
    GlobalPiper& piper;
    int message_processor_index;
}
public:
    Piper() noexcept;
    // initializes max_message_processor_index and max_edge_index
    as zero

    template <typename MP>
    int GetMessageProcessorIndexImpl(const TypeSpecifier<MP>&)
        noexcept;
    // returns -1

    template <typename GlobalPiper>
    void FillEdgeProxysImpl(GlobalPiper&,
        std::vector<std::unique_ptr<Piper::EdgeProxy<GlobalPiper>>>&);
    // Empty function
    // May throw, basic exception safety

    void GetEdgeIndexImpl() noexcept;
    // Empty function

    template <typename GlobalPiper>
    void AddMessageProcessorsImpl(GlobalPiper&,
        std::vector<std::unique_ptr<Piper::MessageProcessorProxy<GlobalPiper>>>&);
    // Empty function
    // May throw, basic exception safety

    virtual ~Piper();
protected:
    int max_message_processor_index;
    int max_edge_index;
    std::vector<std::vector<int>> dest_pipes;
    std::vector<LockFreeQueue<std::unique_ptr<MessageBase>>> queues;

```

```

        std::vector<std::unique_ptr<MessageProcessorBase>>
            message_processors;
};

```

Now the definition of a complete inherited class

```

template <typename From, typename To, typename Message, typename
    ... Args>
class Piper<Edge<From, To, Message>, Args...> : public
    Piper<Args...> {
protected:
    template <typename GlobalPiper, typename From2>
    class SenderProxy {
    // Sender class for MessageProcessors.
    // GlobalPiper is a last Piper<...> child class in the
    // inheritance chain
    public:
        SenderProxy(GlobalPiper& piper) noexcept;
        // Initialize piper reference

        template <typename To2, typename Message2>
        void Send(std::unique_ptr<Message2> message);
        // notifies condition variable from
        // notify_condition_variables vector
        // May throw, strong exception safety
    private:
        GlobalPiper& piper;
    };

    template <typename GlobalPiper, typename MP>
    class MessageProcessorProxy : public
        Piper<>::MessageProcessorProxy<GlobalPiper> {
    public:
        using
            Piper<>::MessageProcessorProxy<GlobalPiper>::MessageProcessorProxy;
        using Piper<>::MessageProcessorProxy<GlobalPiper>::piper;

        virtual void Ping() const;
        // Calls MessageProcessor Ping method, passing a SenderProxy
        // object
        // May throw, basic exception safety
    };

    template <typename GlobalPiper>
    class EdgeProxy : public Piper<>::EdgeProxy<GlobalPiper> {
    public:
        using Piper<>::EdgeProxy<GlobalPiper>::EdgeProxy;
        using Piper<>::EdgeProxy<GlobalPiper>::piper;

        virtual bool NotifyAboutMessage() const;
        // if queue is not empty, calls Receive method of
        // MessageProcessor, passing as SenderProxy object
        // and returns true, returns false otherwise
    };

```

```

    // May throw, basic exception safety

    virtual void
        SetConditionVariable(std::condition_variable_any*) const
        noexcept;
    // sets corresponding condition variable in vector
    notify_condition_variables

    virtual const int GetFromIndex() const noexcept;
    // returns cur_to_index of current Piper layer

    virtual const int GetToIndex() const noexcept;
    // returns cur_from_index of current Piper layer

};
public:
    using Piper<Args...>::max_message_processor_index;
    using Piper<Args...>::GetMessageProcessorIndexImpl;
    using Piper<Args...>::max_edge_index;
    using Piper<Args...>::dest_pipes;
    using Piper<Args...>::queues;
    using Piper<Args...>::message_processors;
    using Piper<Args...>::GetEdgeIndexImpl;

    Piper() noexcept;
    // Does nothing

    template <typename GlobalPiper, typename MP>
    void AddMessageProcessorIfNotExists(int& target_index,
        GlobalPiper& piper,
        std::vector<std::unique_ptr<Piper<>::MessageProcessorProxy<GlobalPiper>>>&
            message_processor_handlers
        const bool is_to);
    // Calls GetMessageProcessorIndexImpl to find MP's message
    // processor index. If returns -1
    // (from a base Piper class implementation), nothing found.
    // In this case function generates new
    // message processor index (target_index =
    // max_message_processor_index++) and pushes a new instance
    // of MessageProcessor to message_processors and
    // a new instance of MessageProcessorProxy<GlobalPiper, MP> to
    // message_processor_handlers,
    // passing current new generated index of message processor to
    // constructor of MessageProcessorProxy.
    // Also it pushes new empty vector to dest_pipes.
    // In any case it saves old or new generated message processor
    // index to target_index.
    // May throw, basic exception safety

    template <typename GlobalPiper>
    void AddMessageProcessorsImpl(GlobalPiper& piper,
        std::vector<std::unique_ptr<Piper<>::MessageProcessorProxy<GlobalPiper>>>&
            message_processor_handlers);
    // Calls itself for parent class,

```

```

// then calls AddMessageProcessorIfNotExists for 'From' and for
// 'To' Message Processor.
// Generates cur_edge_index (cur_edge_index = max_edge_index++)
// Appends to dest_pipes and queues
// May throw, basic exception safety

template <typename GlobalPiper>
void FillEdgeProxysImpl(GlobalPiper& piper,
    std::vector<std::unique_ptr<Piper<>::EdgeProxy<GlobalPiper>>>&
    edge_handlers);
// Calls itself for parent class
// Fills edge_handlers with EdgeProxy<GlobalPiper> instance
// May throw, basic exception safety

void InitializeNotifyConditionVariables();
// Initializes notify_condition_variables with nullptrs
// May throw, basic exception safety

int GetMessageProciessorIndexImpl(const TypeSpecifier<From>&)
    noexcept;
// Specification which returns index of message processor 'From'

int GetMessageProcessorIndexImpl(const TypeSpecifier<To>&)
    noexcept;
// Specification which returns index of message processor 'To'

int GetEdgeIndexImpl(const TypeSpecifier<Edge<From, To,
    Message>>&) noexcept;
// Specification which returns index of edge
private:
    int cur_to_index;
    int cur_from_index;
    int cur_edge_index;
    std::vector<std::condition_variable_any *>
        notify_condition_variables;
};

```

And finally, an enclosing class *MessagePassingTree*, which does all the lazy initialization and provides access to proxy objects:

```

template <typename ... Args>
class MessagePassingTree : public Piper<Args...> {
private:
    using GlobalPiper = Piper<Args...>;
public:
    using GlobalPiper::dest_pipes;

    MessagePassingTree();
    // Calls AddMessageProcessorsImpl<GlobalPiper> to fill
    // message_processor_handlers
    // calls FillEdgeProxysImpl<GlobalPiper> to fill edge_handlers
    // calls InitializeNotifyConditionVariables to fill condition
    // variables

```

```

//
// May throw, basic exception safety

const Piper<>::EdgeProxy<GlobalPiper>* GetEdgeProxy(const int
    index) noexcept;
// returns edge handler from edge_handlers vector

const Piper<>::MessageProcessorProxy<GlobalPiper>*
    GetMessageProcessorProxy(const int index) noexcept;
// returns message processor handler from
    message_processor_handlers vector;

const size_t GetMessageProcessorsCount() noexcept;
// returns message_processor_handlers size

const size_t GetEdgesCount() noexcept;
// returns edge_handlers size

std::vector<int>& GetIncomingEdges(const int
    message_processor_index) noexcept;
// Gets all incoming edges

std::vector<int>& GetOutgoingEdges(const int
    message_processor_index) noexcept;
// Gets all outgoing edges

std::vector<int>& GetConnectingEdges(const int first_mp_index,
    const int second_mp_index) noexcept;
private:
    std::vector<std::unique_ptr<Piper<>::EdgeProxy<GlobalPiper>>>
        edge_handlers;
    std::vector<std::unique_ptr<Piper<>::MessageProcessorProxy<GlobalPiper>>>
        message_processor_handlers;
};

```

3.5. Low-overhead timers

Let us outline classes and methods for several algorithms described above:

- [2.6. Low-overhead periodic timer](#)
- [2.7. Low-overhead start-finish timer](#)
- [2.8. Low-overhead waiting timer](#)
- [2.9. Low-overhead periodic clock](#)

```

class PeriodicTimer {
private:
    void WindUp() noexcept;
public:
    int64_t GetPassedTime() noexcept;
    void Reset() noexcept;
private:

```

```

    static const int64_t min_system_click_call_period;
    static const int64_t max_wind_up_steps;
    static const int64_t min_measured_time;
    int64_t wind_up_counter;
    int64_t wind_up_balance;
    int64_t last_wind_up_counter;
    int64_t last_system_time;
    int64_t approx_time_step;
};
class StartFinishTimer {
public:
    void Start() noexcept;
    void Finish() noexcept;
    uint64_t GetCount() noexcept;
    int64_t GetDurationSum() noexcept;
    void Reset() noexcept;
private:
    PeriodicTimer start_timer;
    PeriodicTimer finish_timer;
    uint64_t measurements_counter;
    int64_t duration_sum;
};
class WaitingTimer {
public:
    WaitingTimer(const uint64_t time_period) noexcept;
    bool CheckTime() noexcept;
    void Reset() noexcept;
private:
    uint64_t time_period;
    int64_t time_elapsed;
    PeriodicTimer periodic_timer;
};
class PeriodicClock {
public:
    int64_t GetTime() const noexcept;
private:
    PeriodicTimer periodic_timer;
    int64_t current_time;
};

```

3.6. Exception with backtrace

For various reasons, while handling exception, one may want to get access to stacktrace. For instance, having stacktrace makes it possible to pass it to the user, without aborting the program.

Let us introduce a class *ExceptionWithBacktrace* which on construction will accept exception reason as string and will remember current stacktrace from place in the code it was constructed. Moreover, it would register itself in global *exception_top_keeper* instance.

Let the *ExceptionWithBacktrace* class be the following:

```

class ExceptionWithBacktrace : public std::exception {
public:
    ExceptionWithBacktrace(const std::string& reason) noexcept
        // stores reason and current backtrace, and their

```

```

// human-readable concatenation in reason_with_backtrace,
// catches inner exceptions and if caught, may leave
// some of inner string fields empty

const std::string& GetStackStrace() const noexcept;
// returns backtrace;

const char* what() const noexcept;
// return reason_with_backtrace.c_str()

const std::string& GetReasonWithBacktrace() const noexcept;
// returns reason_with_backtrace
private:
    std::string reason;
    std::string backtrace;
    std::string reason_with_backtrace;
};

```

3.7. Exception top proto storage interface

Here we are to outline classes and methods of *ExceptionTopKeeper*, which is defined in the section 2.10. Exception top proto storage.

```

class ExceptionTopKeeper {
public:
    void Dump() const noexcept;

    void Restore();

    void Trim() noexcept;

    void RemoveOld(const int64_t current_time) noexcept;

    void SetPath(const std::string& new_path) noexcept;

    template <typename Function>
    void WithCatchingException(Function func) noexcept;
private:
    std::string path;
    std::unordered_map<std::string, uint64_t> exception_frequency;
    std::unordered_map<std::string, int64_t> exception_last_time;
    std::set<std::pair<uint64_t, std::string>> frequency_exception;
    std::set<std::pair<int64_t, std::string>> last_time_exception;
    WaitingTimer dump_timer;
    PeriodicClock store_clock;
    static const size_t top_size;
    static const int64_t keep_time;
    static const debug_mode;
}

```

Let us also declare a global thread_local instance of the class above

```

thread_local ExceptionTopKeeper exception_top_keeper;

```

3.8. Shadow-paging thread-wise sharder interface

Here we are to outline classes and methods required for *Sharder*, which is described in the section [2.11. Shadow-paging thread-wise sharder](#)

Here is how helper class looks

```
class Controller {
public:
    void ProcessShard(const bool can_be_updated, const int
        shard_num);
    // Exceptions: basic safety

    void PreProcess(const int thread_num, const bool
        can_be_updated);
    // Exceptions: basic safety

    void OnSwitch(const int thread_num, const std::vector<int>&
        new_shards) noexcept;

    std::vector<std::vector<int>> GetInitialSharding(const int
        thread_count);
    // Exceptions: basic safety

    std::vector<std::vector<int>> GetResharding(
        const std::vector<std::vector<int>>& old_shard_configuration,
        const int thread_count);
    // Exceptions: basic safety

    const int GetMaxThreadCount();
    // Exceptions: basic safety
};
```

Now let us define *Sharder* class itself:

```
struct ShadowCounter {
    bool is_first_main;
    unsigned noupdate_counter:7;
};

template <typename Controller>
class Sharder {
public:
    Sharder(Controller& controller);
    // Constructor may throw, basic exception safety

    std::vector<std::vector<int>>& GetConf(bool is_first) noexcept;

    void Reshard() noexcept;

    void NoUpdatePromise() noexcept;

    std::vector<int>& GetShards() noexcept;

    void SwitchConfiguration() noexcept;
```

```

void ThreadAction(const int thread\_num) noexcept;
// Inside we have to catch exceptions, though

void Run() noexcept;
// Thread launching may throw, but if it does, we cannot do
// anything better than to terminate
// That is why we mark function as noexcept
private:
    std::vector<std::vector<int>> first_conf;
    std::vector<std::vector<int>> second_conf;
    int threads_count;
    ShadowCounter shadow_counter;
    std::vector<std::thread> threads;
    std::vector<std::mutex> shard_locks;

    static thread_local bool can_be_updated;
    static thread_local bool is_first_local;
    static thread_local int thread_num;
    static thread_local WaitingTimer reshard_waiting_timer;
};

```

3.9. Dynamically-sharded message passing pool interface

Controller object have the following interface:

```

template <typename ... Args>
class MessagePassingController {
public:
    MessagePassingController();
    // Exceptions: basic safety

    void ProcessShard(const bool can_be_updated, const int
        shard_num);
    // Exceptions: basic safety

    void PreProcess(const int thread_num, const bool
        can_be_updated);
    // Exceptions: basic safety

    void OnSwitch(const int thread_num, const std::vector<int>&
        new_shards) noexcept;

    std::vector<std::vector<int>> GetInitialSharding(const int
        thread_count);
    // Exceptions: basic safety

    std::vector<std::vector<int>> GetResharding(
        const std::vector<std::vector<int>>& old_shard_configuration,
        const int thread_count);
    // Exceptions: basic safety

    const int GetMaxThreadCount();

```

```

    // Exceptions: basic safety
private:
    MessagePassingTree<Args...> message_passing_tree;
    std::vector<std::condition_variable_any> message_wait_cvs;
    std::vector<std::condition_variable_any*> message_send_cvs;
    static thread_local bool are_queues_empty;
    static const uint64_t wait_for_messages_time;
    std::vector<StartFinishTimer> message_processor_timers;
    std::vector<StartFinishTimer> edge_timers;
};

```

Let us also introduce easy-to-use completion wrapper:

```

template <typename ... Args>
class DSMPPool {
public:
    DynamicallyShardedMessagePassing()
    // creates controller by default constructor,
    // creates sharder passing controller reference
    // Exceptions: basic safety

    void Run() noexcept;
    // Proxies sharder.Run() call

private:
    MessagePassingController<Args...> controller;
    Sharder<MessagePassingController<Args...>> sharder;
};

```

3.10. Increasing hex identifier interface

Here we are to outline an interface for [2.13. Increasing hex identifier](#)

```

class HexToken {
private:
    void UpdateHexRepr();
public:
    bool IsOver() const noexcept;

    const std::string& GetHex() const noexcept;

    HexToken& operator++();
    // May throw, basic exception safety is provided
private:
    uint16_t index;
    std::string hex_repr;
};

class HexIdGenerator {
public:
    const std::string& GetHex() const noexcept;

    HexIdGenerator& operator++();

```

```
    // May throw, basic exception safety is provided
private:
    std::vector<HexToken> hex_tokens;
    std::string hex_repr;
};
```
