

Using OpenMP from C++

Victor Eijkhout

PEARC 2023



Justification



OpenMP has the opportunity to exploit features of modern C++ that are not present in C. In this course we will explore:

- range-based iteration,
- differences in treatment between vectors and arrays, and various sophisticated reduction schemes.



Basic stuff



Output streams in parallel



The use of *cout* may give jumbled output: lines can break at each <<. Use stringstream to form a single stream to output.

```
// hello.cxx
#pragma omp parallel

int t = omp_get_thread_num();
stringstream proctext;
proctext << "Hello world from " << t;
cerr << proctext str() << '\n';
}</pre>
```



Parallel regions in lambdas



OpenMP parallel regions can be in functions, including lambda expressions.

```
const int s = [] () {
   int s;
   # pragma omp parallel
   # pragma omp master
   s = 2 * omp_get_num_threads();
   return s; }();
```

('Immediately Invoked Function Expression'_,



Dynamic scope for class methods



```
1 // nested.cxx
class withnest {
 public
  void f() {
      stringstream ss
      SS
   << omp_get_num_threads()</pre>
        << '\n':
  cout << ss.str():
  int main() {
     withnest my_object;
  #pragma omp parallel
    my_object.f();
```

```
executing
     → OMP_MAX_ACTIVE_LEVELS=2
     →OMP PROC BIND=true
     →OMP_NUM_THREADS=2
     \hookrightarrow./nested
```

Privatizing class members



Class members can only be privatized from (non-static) class methods.

In this example **f** can not be static:

```
// private.cxx
class foo {
private:
   int x;
public:
   void f() {
   #pragma omp parallel private(x)
   somefunction(x);
};
};
```

You can not privatize just a member:

```
// privateno.cxx
class foo { public: int x; };
int main() {
foo thing;
#pragma omp parallel private(thing.x) // NOPE
```



Vectors are copied, unlike arrays, 1



```
C arrays: private pointer, but shared array:
```

```
1 // alloc.c
2 int *array =
     (int*) malloc(nthreads*sizeof(int)):
   for (int i=0; i<nthreads; i++)</pre>
     arrav[i] = 0;
   #pragma omp parallel firstprivate(array)
     int t = omp_get_thread_num();
   // ptr arith: needs private array
   array += t;
array[0] = t;
   // ... print the array
```

Output

```
Array result: 0:0, 1:1, 2:2, 3:3,
```

Vectors are copied, unlike arrays, 2





Parallel loops



Questions



- 1. Do regular OpenMP loops look different in C++?
- 2. Is there a relation between OpenMP parallel loops and iterators?
- 3. OpenMP parallel loops vs parallel execution policies on algorithms.



Range syntax



Parallel loops in C++ can use range-based syntax as of OpenMP-5.0:

```
// vecdata.cxx
vector<float> values(100);

#pragma omp parallel for
for ( auto& elt : values ) {
    elt = 5.f;
}

float sum{0.f};

#pragma omp parallel for reduction(+:sum)
for ( auto elt : values ) {
    sum += elt;
}
```

Tests not reported here show exactly the same speedup as the C code



General idea





C++ ranges header



The C++20 ranges library is also supported:



C++ ranges speedup



```
---- Run range on 1 threads ----
sum of vector: 50000005000000 in 6.148
sum w/ drop 1: 50000004999999 in 6.017
sum times 2 : 100000010000000 in 6.012
==== Run range on 25 threads ====
sum of vector: 50000005000000 in 0.494
sum w/ drop 1: 50000004999999 in 0.477
sum times 2 : 100000010000000 in 0.489
==== Run range on 51 threads ====
sum of vector: 50000005000000 in 0.257
sum w/ drop 1: 50000004999999 in 0.248
sum times 2 : 100000010000000 in 0.245
==== Run range on 76 threads ====
sum of vector: 50000005000000 in 0.182
sum w/ drop 1: 50000004999999 in 0.184
sum times 2 : 100000010000000 in 0.185
==== Run range on 102 threads ====
sum of vector: 50000005000000 in 0.143
sum w/ drop 1: 50000004999999 in 0.139
sum times 2 : 100000010000000 in 0.134
==== Run range on 128 threads ====
sum of vector: 50000005000000 in 0.122
sum w/ drop 1: 50000004999999 in 0.11
```

Ranges and indices





Custom iterators, 0



Recall that

Short hand:

```
vector<float> v;
for ( auto e : v )
... e ...
```

for:

```
for ( vector<float>::iter
e=v.begin();
e!=v.end(); e++ )
... *e ...
```

If we want

we need a sub-class for the iterator with methods such as **begin**, **end**, * and ++.

Probably also += and -



Custom iterators, 1



OpenMP can parallelize any range-based loop with a random-access iterator

```
Class:

// iterator.cxx

template<typename T>
class NewVector {
protected:
    T *storage;
    int s;

public:
    // iterator stuff
    class iter;
    iter begin();
    iter end();
}
```

Main:

```
NewVector<float> v(s);
#pragma omp parallel for
for ( auto e : v )
cout << e << " ";</pre>
```



Custom iterators, 2



Required iterator methods

```
NewVector<T>::iter& operator++();
T& operator*();
bool operator==( const NewVector::iter & other ) const;
bool operator!=( const NewVector::iter & other ) const;
// needed to OpenMP
int operator-
( const NewVector::iter& other ) const;
NewVector<T>::iter& operator+=( int add );
```

This is a little short of a full random-access iterator; the difference depends on the OpenMP implementation.



Custom iterators, exercise



Write the missing iterator methods. Here's something to get you started.

```
template<typename T>
class NewVector<T>::iter {
  private: T *searcher;
};
template<typename T>
NewVector<T>::iter::iter( T* searcher)
: searcher(searcher) {};
template<typename T>
NewVector<T>::iter NewVector<T>::begin() {
  return NewVector<T>::iter(storage); };
template<typename T>
NewVector<T>::iter NewVector<T>::iter(storage); };
template<typename T>
NewVector<T>::iter NewVector<T>::iter(storage); };
```



Custom iterators, solution



```
template<typename T>
   bool NewVector<T>::iter::operator==
       ( const NewVector<T>::iter &other ) const {
     return searcher other searcher
   template<typename T>
   bool NewVector<T>::iter::operator!=
       ( const NewVector<T>::iter &other ) const {
     return searcher other searcher
   template<typename T>
   NewVector<T>::iter% NewVector<T>::iter::operator++() {
     searcher++: return *this: }
12 template<typename T>
13 NewVector<T>::iter& NewVector<T>::iter::operator+=( int add ) {
     searcher += add: return *this
```



Custom iterators, solution



```
template<typename T>
The NewVector<T>:::iter::operator*() {
    return *searcher; };

// needed for OpenMP
template<typename T>
int NewVector<T>::iter::operator-
(    const NewVector<T>::iterh other ) const {
    return searcher other searcher; };
```



OpenMP vs standard parallelism



Application: prime number marking (load unbalanced)

```
#pragma omp parallel for schedule(guided,8)
for ( int i=0; i<nsize; i++) {
    results[i] = one_if_prime( number(i) );
}

// primepolicy.cxx
transform( std::execution::par,
    numbers.begin(),numbers.end(),
    results.begin(),
    [] (int n ) -> int {
        return one_if_prime(n); }
    );
```

Standard parallelism uses Thread Building Blocks (TBB) as backend



Timing



```
1 ==== Run primepolicy on 1 threads ====
   OMP: found 0 primes; Time: 390 msec (threads= 1)
   TBB: found 0 primes; Time: 392 msec
   ==== Run primepolicy on 25 threads ====
   OMP: found 0 primes; Time: 17 msec (threads=25)
   TBB: found 0 primes; Time: 19 msec
   ==== Run primepolicy on 51 threads ====
   OMP: found 0 primes; Time: 9 msec (threads=51)
   TBB: found 0 primes; Time: 13 msec
10 ==== Run primepolicy on 76 threads ====
   OMP: found 0 primes; Time: 6 msec (threads=76)
   TBB: found 0 primes; Time: 15 msec
   ==== Run primepolicy on 102 threads ====
   OMP: found 0 primes; Time: 5 msec (threads=102)
   TBB: found 0 primes; Time: 71 msec
16 ==== Run primepolicy on 128 threads ====
17 OMP: found 0 primes; Time: 4 msec (threads=128)
   TBB: found 0 primes; Time: 55 msec
```



Reductions vs standard parallelism



Application: prime number counting (load unbalanced)

```
#pragma omp parallel for schedule(guided,8) reduction(+:prime_count)
for (auto n numbers
prime_count += one_if_prime( n );
// reducepolicy.cxx
prime_count = transform_reduce
(std execution par
    numbers begin(), numbers end(),
std::plus<>{},
[] ( int n ) -> int {
       return one_if_prime(n); }
```



Timing



```
Run reducepolicy on 1 threads
   OMP: found 9592 primes; Time: 390 msec (threads= 1)
   TBB: found 9592 primes; Time: 392 msec
  ==== Run reducepolicy on 25 threads ====
   OMP: found 9592 primes; Time: 17 msec (threads=25)
   TBB: found 9592 primes; Time: 20 msec
  ==== Run reducepolicy on 51 threads ====
   OMP: found 9592 primes; Time: 8 msec (threads=51)
   TBB: found 9592 primes; Time: 13 msec
10 ==== Run reducepolicy on 76 threads ====
   OMP: found 9592 primes; Time:
                              6 msec (threads=76)
   TBB: found 9592 primes; Time: 23 msec
   ==== Run reducepolicy on 102 threads ====
   OMP: found 9592 primes; Time: 5 msec (threads=102)
   TBB: found 9592 primes; Time: 105 msec
   ==== Run reducepolicy on 128 threads ====
   OMP: found 9592 primes; Time: 4 msec (threads=128)
   TBB: found 9592 primes; Time: 54 msec
```

Reductions



Questions



- 1. Are simple reductions the same as in C?
- 2. Can you reduce std::vector like an array?
- 3. Precisely what can you reduce?
- 4. Any interesting examples?
- **5**. Compare reductions to native C++ mechanisms.



Scalar reductions



Same as in C, you can now use range syntax for the loop.

```
// range.cxx
pragma omp parallel for reduction(+:count)
for ( auto e : data )
count += e;
```



Reductions on vectors



Use the data method to extract the array on which to reduce. However, this does not work:

```
vector<float> x;
pragma omp parallel reduction(+:x.data())
```

because the reduction clause wants a variable, not an expression, for the array, so you need an extra bare pointer:

Reduction on class objects



Reduction can be applied to any class for which the reduction operator is defined as *operator*+ or whichever operator the case may be.

A default constructor is required for the internally used init value; see figure 32.



Reduction illustrated



u



User-defined reductions, syntax



```
#pragma omp declare reduction
 identifier typelist combiner
[initializer(initializer-expression)]
```



Reduction over iterators



Support for C++ iterators



Lambda expressions in declared reductions TACC



```
// reductexpr.cxx
#pragma omp declare reduction\
  (minabs int \
omp_out = \
       [] (int x, int y) \rightarrow int { \
          return abs(x) > abs(y) ? abs(y) : abs(x); }
       (omp_in,omp_out) )
  initializer (omp_priv=limit::max())
```

because omp_in/out are the only variables allowed in the explicit



Example category: histograms



Count which elements fall into what bin

```
for ( auto e : some_range )
    histogram[ value(e)]++;
```

Collisions are possible, but unlikely, so critical section is very inefficient



Histogram: intended main program



Declare a reduction on a histogram object

Q: why does the *inc* not have to be atomic?



Histogram solution: reduction operator



Give the class a += operator to do the combining:

```
// charcount.cxx
2 template<typename key>
   class bincounter    public map<key int> {
   public
   // merge this with other map
     void operator+=( const bincounter<key>& other ) {
       for (auto [k, v] other)
         if ( map<key,int>::contains(k) ) // c++20
          this \rightarrow at(k) += v:
        else
         this->insert( {k, v} ):
   // insert one char in this map
   void inc(char k)
      if ( map<key,int>::contains(k) )
  this->at(k) += 1:
17 else
       this->insert( {k,1} );
```

Histogram in native C++



Use atomics because there is no reduction mechanisms

```
// mapreduceatomic.cxx
   class CharCounter : public array<atomic<int>,26> {
   public
   CharCounter() {
    for ( int ic=0: ic<26: ic++ )
        (*this)[ic] = 0;
   // insert one char in this map
   void inc(char k) {
  if (k==', ') return;
int ik = k-'a';
12 (*this)[ik]++;
```

Histogram in native C++, comparison



OpenMP reduction on array<int,26>:

```
Using atomics on 1 threads: time= 20.19 msec
UpenMP reduction on 1 threads: time= 1.966 msec
Using atomics on 5 threads: time= 315.855 msec
UpenMP reduction on 5 threads: time= 0.52 msec
Using atomics on 10 threads: time= 91.968 msec
UpenMP reduction on 10 threads: time= 0.364 msec
Using atomics on 30 threads: time= 249.171 msec
UpenMP reduction on 30 threads: time= 0.556 msec
Using atomics on 50 threads: time= 164.177 msec
UpenMP reduction on 50 threads: time= 0.904 msec
```



Exercise: mapreduce



Make an OpenMP parallel version of

```
intcounter primecounter;
for ( auto n : numbers )
if ( isprime(n) )
primecounter.add(n);
```

where primecounter contains a map<int,int>

Use skeleton: mapreduce.cxx



Example category: list filtering



The sequential code is as follows

```
vector<int> data(100);
// fil the data
vector<int> filtered;
for ( auto e : data ) {
   if ( f(e) )
   filtered.push_back(e);
}
```



List filtering, solution 1



Let each thread have a local array, and then to concatenate these:

```
#pragma omp parallel

vector<int> local;

pragma omp for

for ( auto e : data )

if ( f(e) ) local.push_back(e);

filtered += local;

}
```

where we have used an append operation on vectors:

```
// filterreduct.cxx
template<typename T>
vector<T>& operator+=(vector<T>& me, const vector<T>& other) {
    me.insert(me.end(),other.begin(),other.end());
    return me;
};
```



List filtering, not quite solution 2



We could use the plus-is operation to declare a reduction

Problem: OpenMP reductions can not be declared non-commutative, so the contributions from the threads may not appear in order.

Codo

```
#pragma omp parallel
reduction(+: filtered)
{
    vector<int> local;
    pragma omp for
    for ( auto e : data )
    if ( f(e) )
        local.push_back(e);
    filtered += local;
```

Output:

```
Mod 5: 80 85 90 95 

\hookrightarrow100 5 10 15 20 

\hookrightarrow25 30 35 40 45 

\hookrightarrow50 55 60 65 70 

\hookrightarrow75
```

List filtering, task-based solution



```
Parallel region, without for:
// filtertask.cxx
vector<int> filtered;
3 int ithread=0:
   #pragma omp parallel
   vector<int> local:
     int threadnum = omp_get_thread_num();
   # pragma omp for
  for ( auto e : data )
if (e%5==0)
           local push_back(e);
   // create task to add local to filtered
```

Output

```
Mod 5: 5 10 15 20 25

\hookrightarrow30 35 40 45 50

\hookrightarrow55 60 65 70 75

\hookrightarrow80 85 90 95 100
```

List filtering, task-based solution'



The task spins until it's its turn:

Output:



Templated reductions



You can reduce with a templated function if you put both the declaration and the reduction in the same templated function:

```
template<typename T>
T generic_reduction( vector<T> tdata ) {
    #pragma omp declare reduction
    (rwzt:T.omp_out=reduce_without_zero<T>(omp_out,omp_in))
    initializer(omp_priv=-1.f)

T tmin = -1;
#pragma omp parallel for reduction(rwzt:tmin)
    for (int id=0; id<tdata.size(); id++)
        tmin = reduce_without_zero<T>(tmin,tdata[id]);
    return tmin;
};
```

which is then called with specific data:

```
auto tmin = generic_reduction<float>(fdata);
```



More topics



Threadprivate random number generators



The new C++ random header has a threadsafe generator, by virtue of the statement in the standard that no STL object can rely on global state. The usual idiom can not be made threadsafe because of the initialization:

```
static random_device rd;
static mt19937 rng(rd);
```

However, the following works

```
// privaterandom.cxx
static random_device rd;
static mt19937 rng;
#pragma omp threadprivate(rd)
#pragma omp threadprivate(rng)
int main() {
#pragma omp parallel
rng = mt19937(rd());
```

Threadprivate random use



Based on the previous note, you can use the generator safely and independently:



Uninitialized containers



```
double *x = (double*)malloc( N*sizeof(double));
#pragma omp parallel for
 for (int i=0; i<N; i++)
   x[i] = f(i);
   std::vector<double> x(N);
#pragma omp parallel for
for (int i=0; i<N; i++)</pre>
     x[i] = f(i);
```

because of value initialization in the vector container.



Uninitialized containers, 2



Trick to create a vector of uninitialized data

```
// heatalloc.cxx
2 template<typename T>
   struct uninitialized
  uninitialized() {};
5 T val:
  constexpr operator T() const {return val;};
     T operator=( const T&& v ) { val = v; return val; };
   vector<uninitialized<double>> x(N),y(N);
   #pragma omp parallel for
4 for (int i=0; i<N; i++)
y[i] = x[i] = 0.;
\mathbf{x}[0] = 0; \mathbf{x}[N-1] = 1.;
```

(Question: why not use reserve?)



Uninitialized containers, 3



Easy way of dealing with this

```
template<typename T>
class ompvector : public vector<uninitialized<T>> {
  public:
    ompvector( size_t s )
    : vector<uninitialized<T>>::vector<uninitialized<T>>(s) {};
};
```



Atomic scalar updates



Can you atomically update scalars'

- Make an object that has data plus a lock;
- Disable copy and copy-assignment operators;
- Destructor does omp_destroy_lock;
- Overload arithmetic operator.

(Quick self-test: why lock, not critical?)



Atomic updates: class with OMP lock



```
// lockobject.cxx
class atomic_int
private
  omp_lock_t the_lock
  int _value{0};
public
  atomic_int() {
    omp_init_lock(&the_lock);
  atomic_int( const atomic_int& )
      = delete
  atomic_int& operator=( const atomic_int& )
      = delete
~atomic int() {
    omp_destroy_lock(&the_lock);
```



Atomic updates: atomic ops



```
int operator +=( int i ) {
// atomic increment
omp_set_lock(&the_lock);
value += i; int rv = _value;
omp_unset_lock(&the_lock);
return rv;
};
```



Atomic updates: usage





Atomic updates, comparison to native



Timing comparison on simplest case:

Object with built-in lock:

Vative C++ atomics:

Native solution is 10x faster



False sharing prevention



```
#include <new>
#ifdef __cpp_lib_hardware_interference_size
const int spread = std::hardware_destructive_interference_size
          sizeof(datatype);
#else
const int spread = 8;
#endif
vector<datatype> k(nthreads*spread)
#pragma omp parallel for schedule( static, 1 )
for ( datatype i = 0; i < N; i++ ) {</pre>
  k[(i\%nthreads) * spread] += 2;
```

Since $\mathsf{C}{+}{+}17$



Beware vector-of-bool!



```
1 // boolrange.cxx
vector<bool> bits(1000000);
3 for ( auto& b : bits )
b = true:
```

```
// booliter.cxx
vector<bool> bits(3000000);
   #pragma omp parallel for schedule(static,4)
   for ( int i=0; i < bits.size(); i++ )</pre>
     bits[i] = ( i%3==0 );
   // and then count the million 1s
```

```
#threads=1: should be
   #threads=2; should be
   #threads=3; should be
4 #threads=4: should be
```

Different bits[i] are falsely shared.



