SYCL II

Soner Steiner

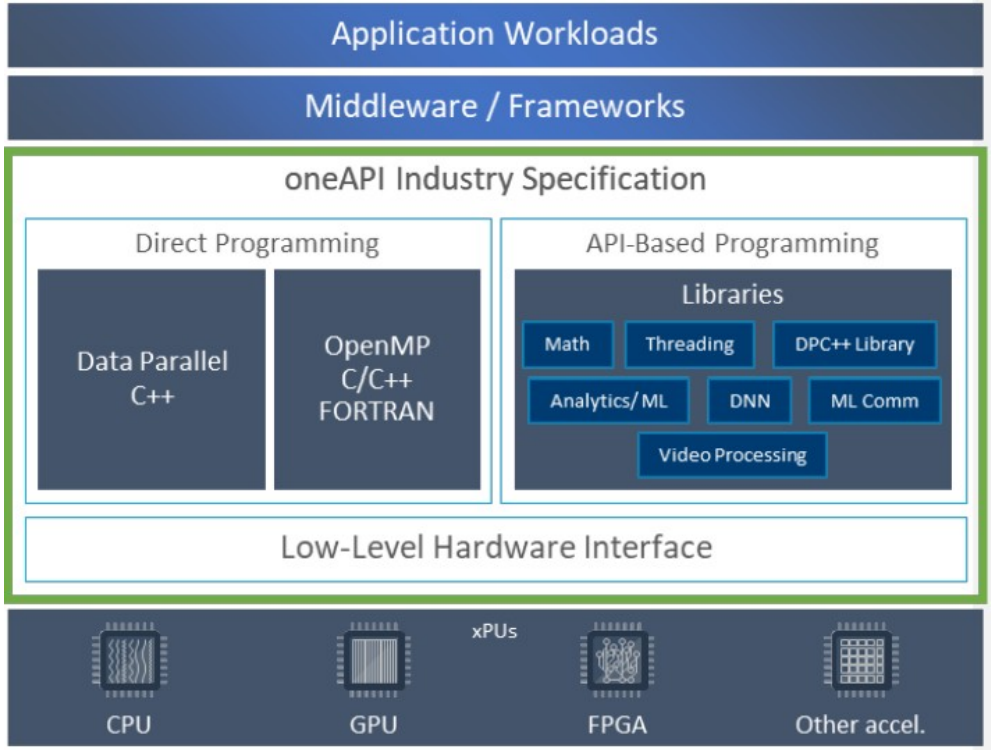
Intel certified oneAPI Instructor
sonersteiner at gmail dot com

Overview

- Manage the data transfer
  Buffers and Unified Shared Memory
- Basic parallel kernels
- ND-Range kernels
- Sub-groups
- Reductions

Programmers' perspective: Three things to consider

1. Offload the code to device
2. Manage the transfer of Data
3. Implement Parallelism



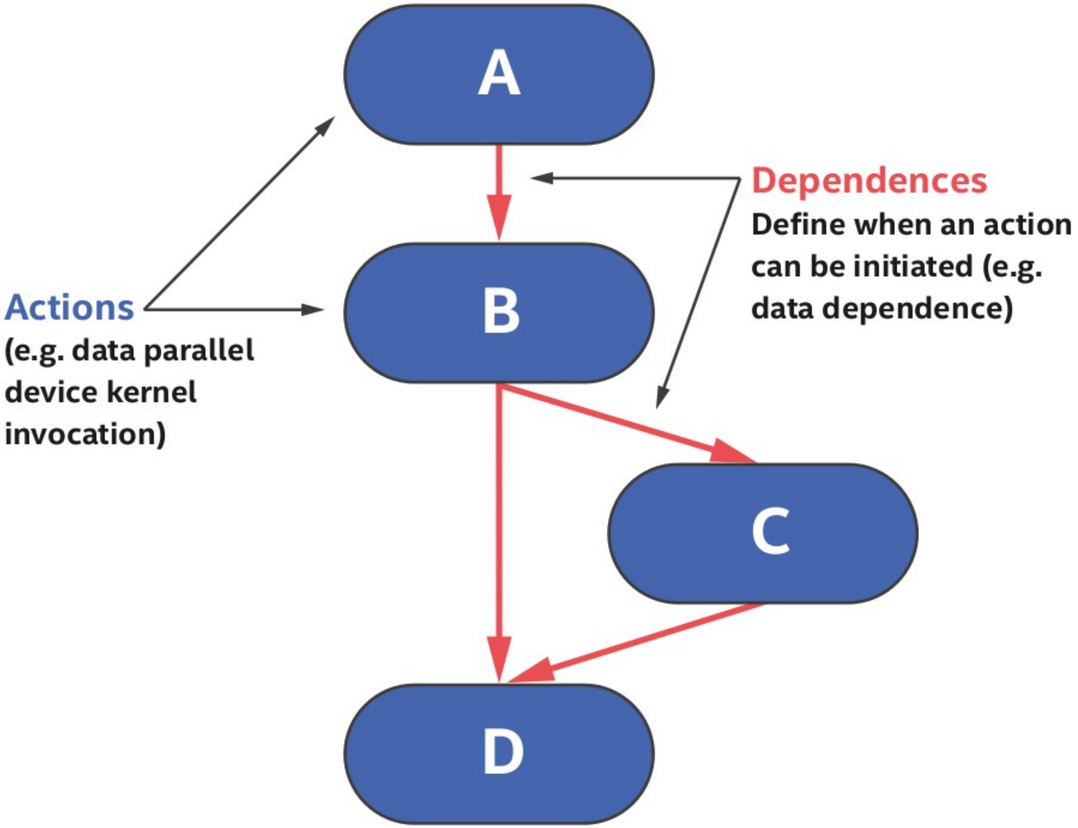
Memory Models

- Buffer Memory Model – abstract view of memory that can be local to the host or a device, and is accessible via accessors.

- Unified Shared Memory (USM)- pointer-based approach for memory model that os familiar for C++ programmers.

- Images: a special type of buffer that has an extra functionality specific to image processing

Task Graphs
(Directed Acyclic Graph)



- Dependency resolution and node execution are controlled by the runtime
- Dependencies determine the order that kernels are executed in
- Dependencies can be explicit or implicit

J. Reinders et al., Data Parallel C++, download link

Explicit Dependencies Using Events

```
constexpr int N = 101;
int main()
{
   queue q;
   int *data = malloc_shared<int>(N, q);

   auto e = q.parallel_for(N, [=] (id<1> i) { data[i] = i ;} );
   q.submit( [&] (handler &h)
     {
       h.depends_on(e);
       h.single_task([=] ()
           {
               for(int i = 1; i < N; ++i)
                   data[0] += data[i];
           } );
     } );
   q.wait();

   std::cout << "printing sum after computation \n" ;
   std::cout << data[0] << " ";
   std::cout << "\n" ;
}
```

- Create event to initialize the data in kernel1
- Kernel2 sums up the elements
- 5050

Buffer Memory Model

Buffers encapsulate data shared between host and device

Accessors provide access to data stored in buffers and create data dependencies in the graph.

Unified Shared Memory (USM) provides an alternative pointer-based mechanism for managing memory

```
queue q;
std::vector<int> v(N, 3);
{
    buffer buf(v);
    q.submit( [&] (handler& h)
     {
        accessor a(buf, h, write_only);
        h.parallel_for(N, [=] (auto i) { a[i] = i; } );
     } );

}

for (int i = 0; i < N; i++) std::cout << v[i] << " ";
```

Buffer Creation – two approaches

- Construct a new buffer using sycl::range to specify the size, data will not be initialized!

```
buffer( const sycl::range<dimensions> &bufferRange,
        const sycl::property_list &proplist={} );
```

- Create buffer from existing data, data will be copied!

```
Buffer( T, hostData,
        const sycl::range<dimensions> &bufferRange,
        const sycl::property_list &proplist={} );
```

Examples of Buffer Creation

```
buffer b1{v};
buffer b2{v.begin(), v.end()};

// create a buffer of ints from std:array
std::array<int, 42> data;
buffer b3{data};

 // create a buffer of 5 doubles and initialize it from
 // a host pointer
double dd[5] = {1.1, 2.2, 3.14, 4.4, 5.5};
buffer b4{dd, range{5} };

std::cout << "printing v before computation \n" ;
for (int i = 0; i < N; i++) std::cout << v[i] << " ";
std::cout << "\n" ;
```

Buffer for vectors

Buffer for std::array

Buffer from a host pointer

Accessors

- Only means of accessing data in Buffers!

- They create the dependencies for the runtime.

Accessor Modes

Access Mode	**Description**
read_only	Read only Access
write_only	Write-only accessor Previous Contents not discarded
read_write	Read and Write access

Code Walkthrough

```
#include <CL/sycl.hpp>
using namespace sycl;

int main() {
std::vector<float> A(1024, 1.0f), B(1024, 2.0f), C(1024);
   {
      buffer bufA {A}, bufB {B}, bufC {C};
      queue q;
      q.submit([&](handler &h) {
          auto A = bufA.get_access(h, read_only);
          auto B = bufB.get_access(h, read_only);
          auto C = bufC.get_access(h, write_only);
          h.parallel_for(1024, [=](auto i){
              C[i] = A[i] + B[i];
          });
      });
   }
```

Host Accessor
(up to now our accessors have been in the command group)

- The Host Accessor is an accessor which uses host buffer access target.
- Host accessors make data available for access on the host.
- They synchronize with the host by defining a new dependence between the currently accessing graph and the host.
- Creating host accessor is a blocking call.

Some Dependency Patterns

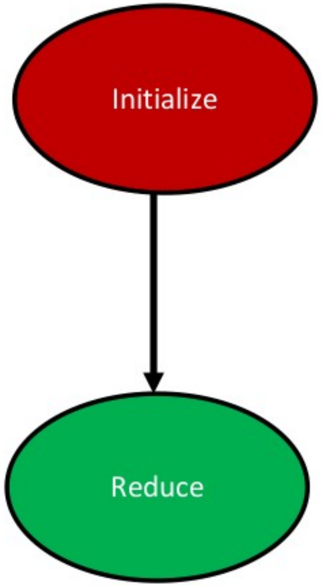
Linear Dependence Using In-order queue

Create In-order queue

Initialize the data in Kernel 1

Kernel 2 sums up the elements

```
constexpr int N=42;

int main()
{
    queue Q{property::queue::in_order()};

    int *data = malloc_shared<int>(N,Q);

    Q.parallel_for(N, [=](id<1> i) { data[i] = 1; });

    Q.single_task([=]()
      {
        for(int i=1; i < N; ++i)
            data[0] += data[i];
      });
    Q.wait();

    assert(data[0] == N);
    for(int i = 0; i < N; ++i)
        std::cout << data[i] << " ";
    std::cout  << "\n";
    return 0;
}
```

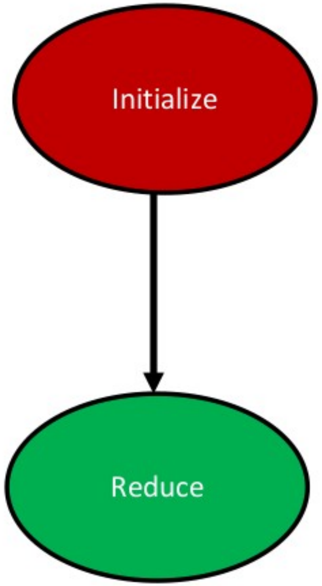


Linear Dependence Using Buffers and Accessors

Use Buffers and Accessors to Initialize the data in Kernel1

Kernel 2 sums up the elements

```
constexpr int N=101;
int main()
{
   queue q;
   buffer <int> data{ range{N} };

   q.submit( [&] (handler &h)
     {
       accessor a{data, h};
       h.parallel_for(N, [=] (id<1> i) { a[i] = i; } );
     } );

   q.submit( [&] (handler &h)
     {
      accessor a{data, h};
      h.single_task([=] ()
       {
           for(int i = 1; i < N; ++i)
               a[0] += a[i];
       } );
     } );
   host_accessor h_a{data};
   std::cout << h_a[0] << "\n";

   return 0;
}
```



5050

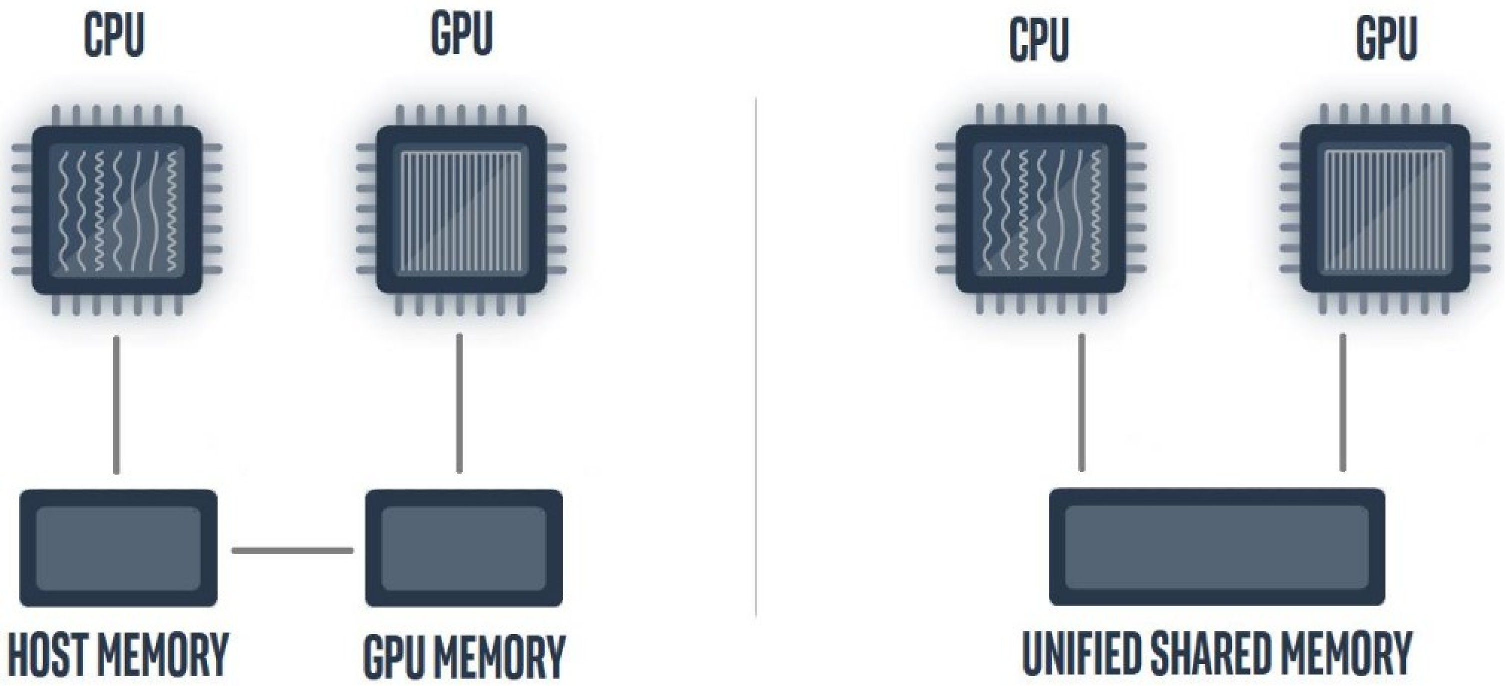
Unified shared memory (USM)

USM provides a pointer-based alternative in SYCL

- Simplifies porting to an accelerator
- Gives programmers the desired level of control
- Complementary to buffers

Developer View of USM

Developers can reference the same memory object in host and device code with USM



Unified shared memory (USM)

USM provides both explicit and implicit models for managing memory.

Allocation Type	Description	Accessible on HOST	Accessible on DEVICE
device	Allocations in device memory (**explicit**)	NO	YES
host	Allocations in host memory (**implicit**)	YES	YES
shared	Allocations can migrate between host and device memory (**implicit**)	YES	YES

Automatic data accessibility and explicit data movement supported.

USM – Explicit Data Movement

```
queue q;
int hostArray[N];
int *deviceArray = (int*) malloc_device(N * sizeof(int), q);

for(int i = 0; i < N; ++i) hostArray[i] = i;

// copy hostArray tp deviceArray
q.memcpy(deviceArray, &hostArray[0], N*sizeof(int));
q.wait();

q.submit( [&] (handler &h)
  {
    h.parallel_for(N, [=] (auto ID)
        {
            deviceArray[ID] = ID*ID ;
        } );
  } );
  q.wait();

//copy deviceArray back to hostArray
q.memcpy(&hostArray[0], deviceArray, N*sizeof(int));
q.wait();
free(deviceArray, q);
```

malloc_device

mem_copy

mem_copy

USM – Implicit Data Movement

```
queue q;
int *hostArray   = (int*) malloc_host(N * sizeof(int), q);;
int *sharedArray = (int*) malloc_shared(N * sizeof(int), q);

for(int i = 0; i < N; ++i) hostArray[i] = i;

q.submit( [&] (handler &h)
  {
    h.parallel_for(N, [=] (auto ID)
        {
            sharedArray[ID] = hostArray[ID] * hostArray[ID];
        } );
  } );
  q.wait();

for (int i = 0; i < N; i++) hostArray[i] = sharedArray[i] ;
free(hostArray, q);
free(sharedArray, q);
```

malloc_host

malloc_shared

Unified Shared Memory – When to use it?

SYCL* Buffers are powerful and elegant

- Use if the abstraction applies cleanly in your application, and/or buffers aren’t disruptive to your development

USM provides a familiar pointer-based C++ interface

- Useful when porting C++ code to SYCL, by minimizing changes
- Use shared allocations when porting code, to get functional quickly
- Note that shared allocation is not intended to provide peak performance out of box
- Use explicit USM allocations when controlled data movement is needed

USM – Data Dependency in Queues

No accessors in USM

Dependences must be specified explicitly using events

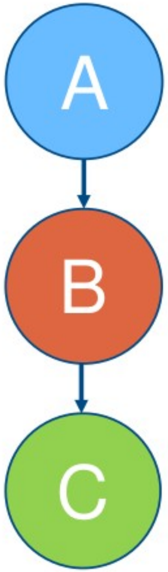
- queue.wait()
- wait on event objects
- use the depens_on method inside a command group

USM – Data Dependency in Queues

```
queue q;
int *data = (int*) malloc_shared(N * sizeof(int), q);
for(int i = 0; i < N; ++i) data[i] = i;

q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskA>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 1;
        } );
  } );
q.wait();
q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskB>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 2;
        } );
  } );
q.wait();
q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskC>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 3;
        } );
  } );
q.wait();

for (int i = 0; i < N; i++) std::cout << data[i] << " ";
free(data, q);
```

Explicit wait() used to ensure Data dependency is maintained

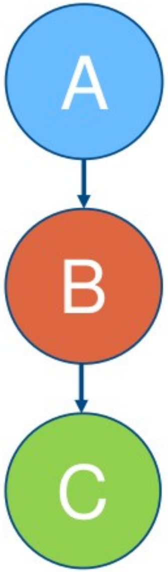
wait() will block execution on host



USM – Data Dependency in Queues

```
queue q;
int* data = malloc_shared<int>(N, q);

for(int i = 0; i < N; ++i) data[i] = i;

auto e1 = q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskA>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 1;
        } );
  } );
  auto e2 = q.submit( [&] (handler &h)
  {
    h.depends_on(e1);
    h.parallel_for<class taskB>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 2;
        } );
  } );
// non-blocking: execution of host code is possible
q.submit( [&] (handler &h)
  {
    h.depends_on(e2);
    h.parallel_for<class taskC>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 3;
        } );
  } );
  q.wait();
std::cout << "printing data after computation \n" ;
for (int i = 0; i < N; i++) std::cout << data[i] << " ";
free(data, q);
```

use depends_on method to let command group handler know that specified event should be complete before specified task can execute.

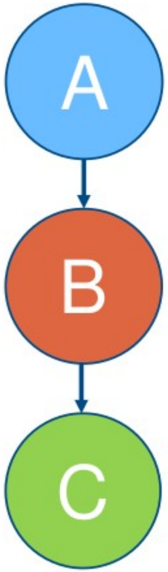


USM – Data Dependency in Queues

```
queue q{property::queue::in_order()};
int* data = malloc_shared<int>(N, q);

for(int i = 0; i < N; ++i) data[i] = i;
q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskA>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 1;
        } );
  } );
q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskB>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 2;
        } );
  } );
q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskC>(range<1> (N), [=] (id<1> i)
        {
            data[i] += 3;
        } );
  } );
q.wait();
free(data, q);
```

use in_queue property for the queue

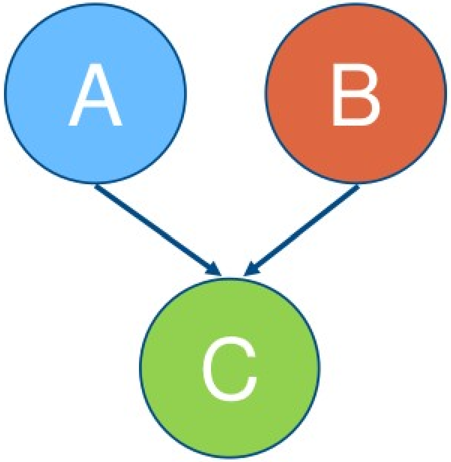
Execution will not overlap even
If the queues have no data dependency



USM – Data Dependency in Queues

```
queue q;
int *data1 = (int*) malloc_shared(N * sizeof(int), q);
int *data2 = (int*) malloc_shared(N * sizeof(int), q);

for(int i = 0; i < N; ++i){ data1[i] = 10; data2[i] = 20;}

auto e1 = q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskA>(range<1> (N), [=] (id<1> i)
        {
            data1[i] += 1;
        } );
  } );
auto e2 = q.submit( [&] (handler &h)
  {
    h.parallel_for<class taskB>(range<1> (N), [=] (id<1> i)
        {
            data2[i] += 2;
        } );
  } );
q.submit( [&] (handler &h)
  {
    h.depends_on({e1, e2});
    h.parallel_for<class taskC>(range<1> (N), [=] (id<1> i)
        {
            data1[i] += data2[i];
        } );
  } );
  q.wait()
for (int i = 0; i < N; i++) std::cout << data1[i] << " ";
free(data1, q); free(data2, q);
```

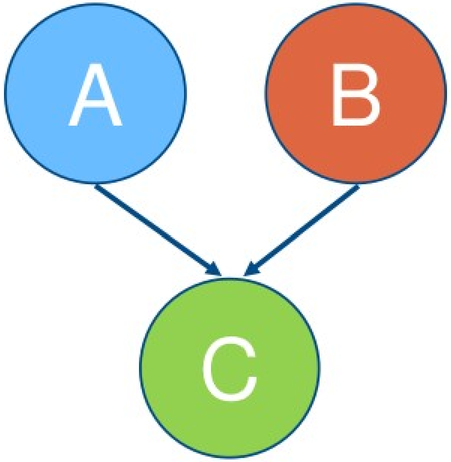
use depends_on() method to let command group handler know that specified events should be complete before specified tasks can execute.



USM – Data Dependency in Queues

```
queue q;
int* data1 = malloc_shared<int>(N, q);
int* data2 = malloc_shared<int>(N, q);

for(int i = 0; i < N; ++i){ data1[i] = 10; data2[i] = 20;}

auto e1 = q.parallel_for<class taskA>(range<1> (N), [=] (id<1> i)
        {
            data1[i] += 1;
        } );
auto e2 = q.parallel_for<class taskB>(range<1> (N), [=] (id<1> i)
        {
            data2[i] += 2;
        } );
q.parallel_for<class taskC>(range<1> (N), {e1, e2}, [=] (id<1> i)
        {
            data1[i] += data2[i];
        } );
  q.wait();

free(data1, q); free(data2, q);
```

A more simplified way of specifying dependency as parameter of parallel_for

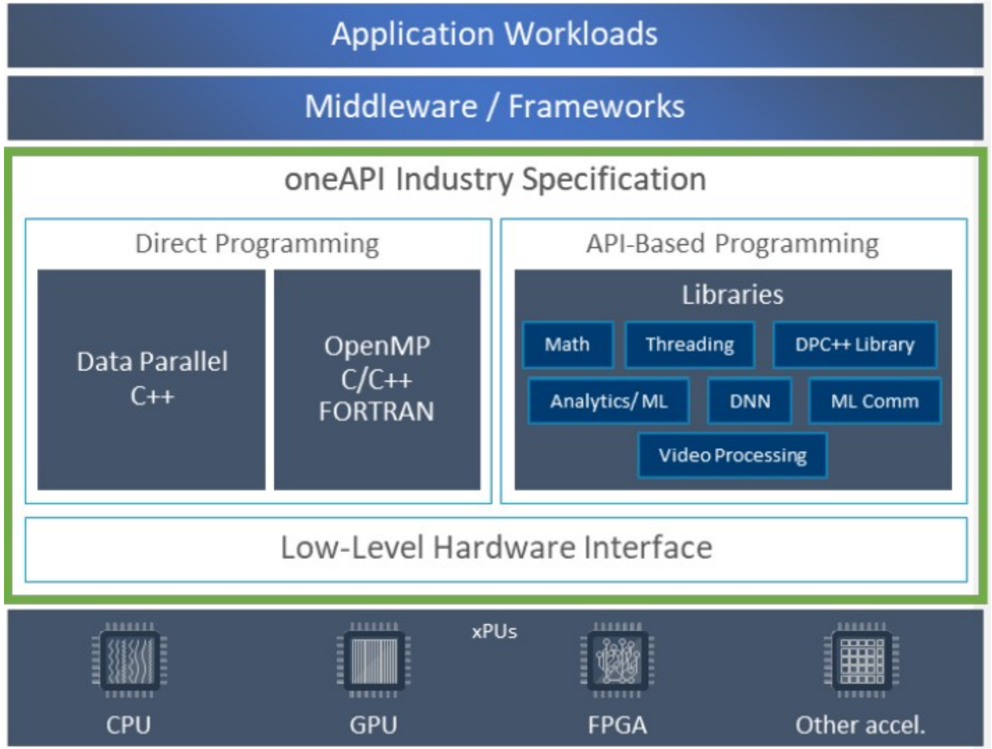


Unified Shared Memory

- Summary
  - What is Unified Shared Memory (USM)?
  - Implicit and Explicit data movement between host and device
  - Handling data dependency in multiple kernel tasks using wait event, depends_on method and in_order queue property

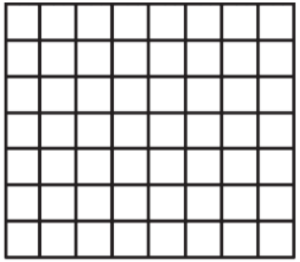
Programmers' perspective: Three things to consider

1. Offload the code to device
2. Manage the transfer of Data
3. Implement Parallelism

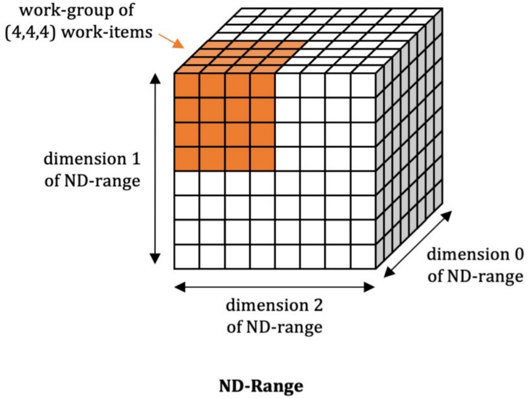


Three forms of Parallel Kernels

- **Basic** Parallel Kernels



- ND-range Parallel Kernels



- Hierarchical Parallel Kernels (‘Experimental alternative syntax’)

Basic Parallel Kernels

- Parallel kernel allows multiple instances of an operation to execute in parallel.
- Useful to offload parallel execution of a basic for-loop in which each iteration is completely independent and in any order.
- Parallel kernels are expressed using the parallel_for function.
- Up to the programmer to handle/confirm that there are no dependencies.

for-loop in CPU application

```
for(int i = 0; i < N; ++i)
{
    c[i] = a[i] + c[i] ;
}
```

Offload to a accelerator using parallel_for

```
h.parallel_for(range<1>(N), [=](id<1> i)
{
    C[i] = A[i] + B[i] ;
});
```

Basic Parallel Kernels

The functionality of basic parallel kernels is exposed via range, id and item classes

- range class is used to describe the iteration space of parallel execution
- id class is used to index an individual instance of a kernel in a parallel execution

```
h.parallel_for(range<1>(N), [=](id<1> idx)
{
    //CODE THAT RUNS ON DEVICE
});
```

Basic Parallel Kernels

The functionality of basic parallel kernels is exposed via range, id and item classes

- range class is used to describe the iteration space of parallel execution
- id class is used to index an individual instance of a kernel in a parallel execution
- item class represents an individual instance of a kernel function, exposes additional functions to query properties of the execution range

```
h.parallel_for(range<1>(N), [=](id<1> idx)
{
    //CODE THAT RUNS ON DEVICE
});
```

```
h.parallel_for(range<1>(N), [=](item<1> item)
{
    auto idx = item.get_id();
    auto R   = item.get_range();

    //CODE THAT RUNS ON DEVICE
});
```

Basic Parallel Kernels

The functionality of basic parallel kernels is exposed via range, id and item classes

- Dimensionality
  <1>, <2> or <3>
  is templated and must be declared at COMPILE time

- Size is dynamic – passed to constructor at runtime

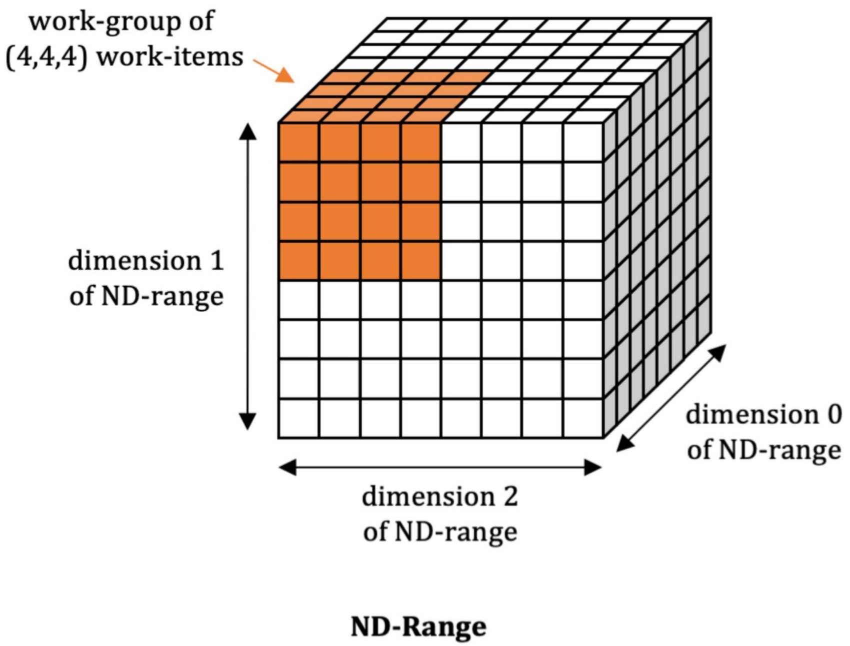
```
h.parallel_for(range<1>(N), [=](id<1> idx)
{
    //CODE THAT RUNS ON DEVICE
});
```

```
h.parallel_for(range<1>(N), [=](item<1> item)
{
    auto idx = item.get_id();
    auto R   = item.get_range();

    //CODE THAT RUNS ON DEVICE
});
```

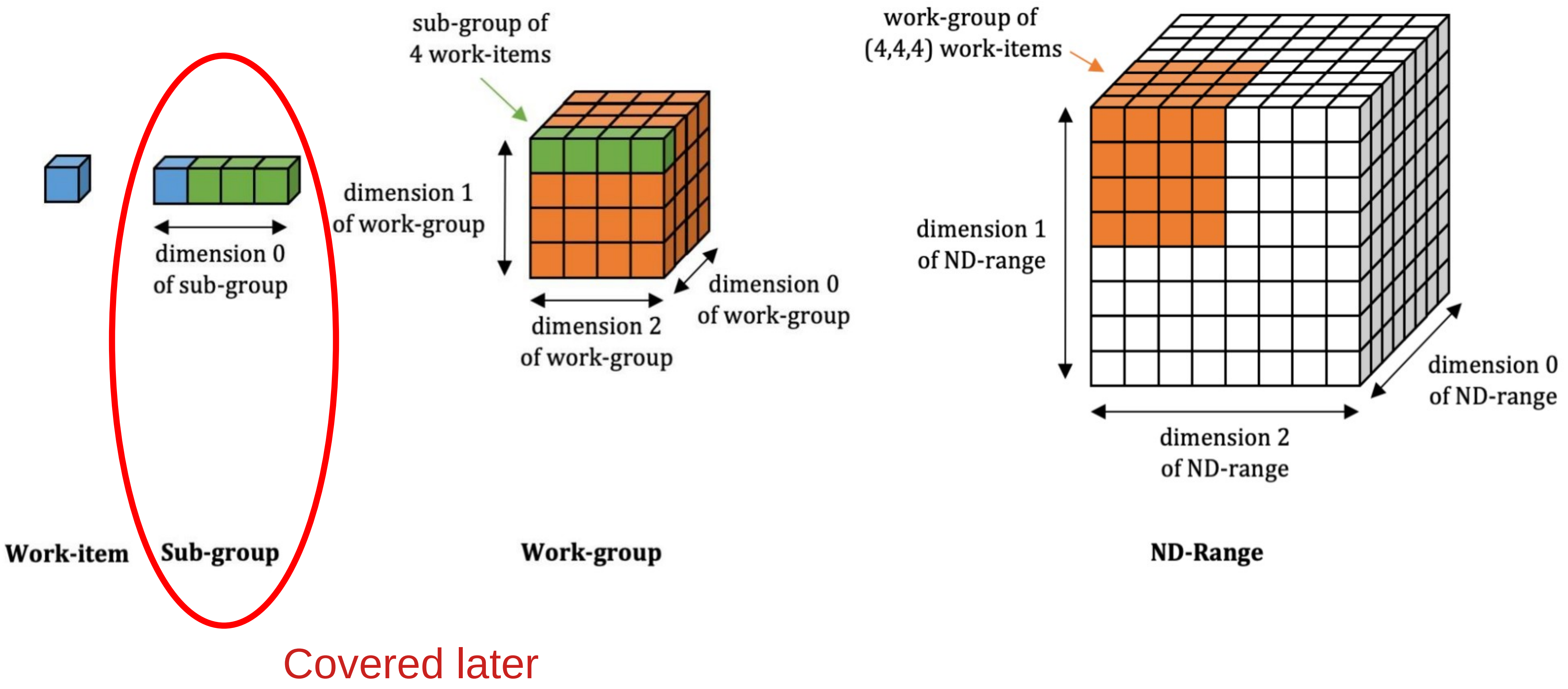
ND-range Kernels

- ND-range kernels enable low level performance tuning by providing access to local memory and mapping executions to compute units on hardware.

- The entire iteration space is divided into smaller groups called work-groups, work-items within a work-group are scheduled on a single compute unit on hardware.
- The grouping of kernel executions into work-groups will allow control of resource usage and load balance work distribution.

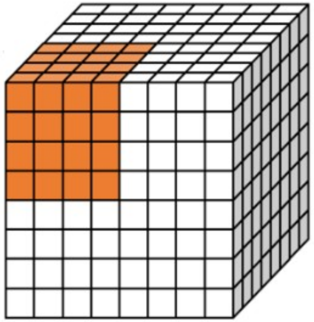


SYCL Thread Hierarchy and Mapping

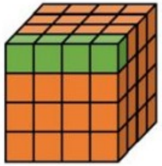


SYLC Thread Hierarchy and Mapping

All work-items in a **work-group** are scheduled on one Compute Unit, which has its own local memory



All work-items in a **sub-group** are mapped to vector hardware



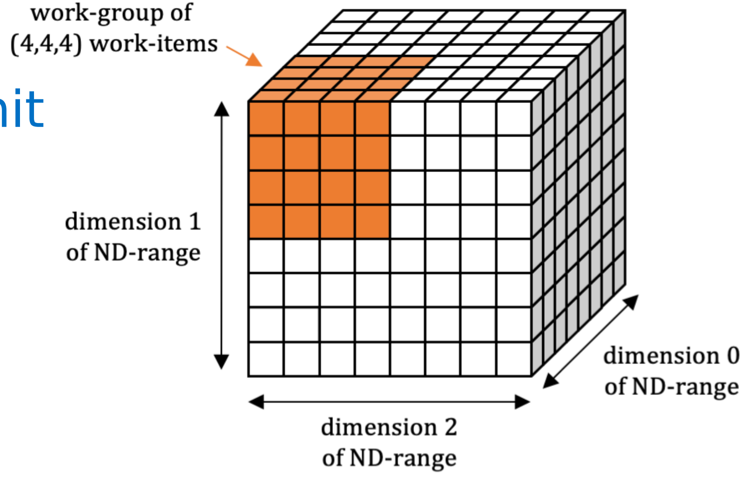
ND-range Kernels

- Basic Parallel Kernels are easy way to parallelize a for-loop but does not allow performance optimization at hardware level.

- ND-range kernel is another way to express parallelism which enable low level performance tuning by providing access to local memory and mapping executions compute units on hardware.

  - The entire iteration space is divided into smaller groups called work-groups, work-items within a work-group are scheduled on a single compute unit on hardware.

  - The grouping of kernel executions into work-groups will allow control of resource usage and load balance work distribution.



ND-Range

ND-range Kernels

The functionality of nd_range kernels is exposed via nd_range and nd_item classes

nd_range class represents a grouped execution range using global execution range and the local execution range of each work-group.

nd_item class represents an individual instance of a kernel function and allows to query for work-group range and index.

Sub-groups

Sub-groups

Understand how Sub-Groups map to GPU hardware

Understand how using Sub-Groups shuffle operations can achieve better performance and avoid repeated global memory access

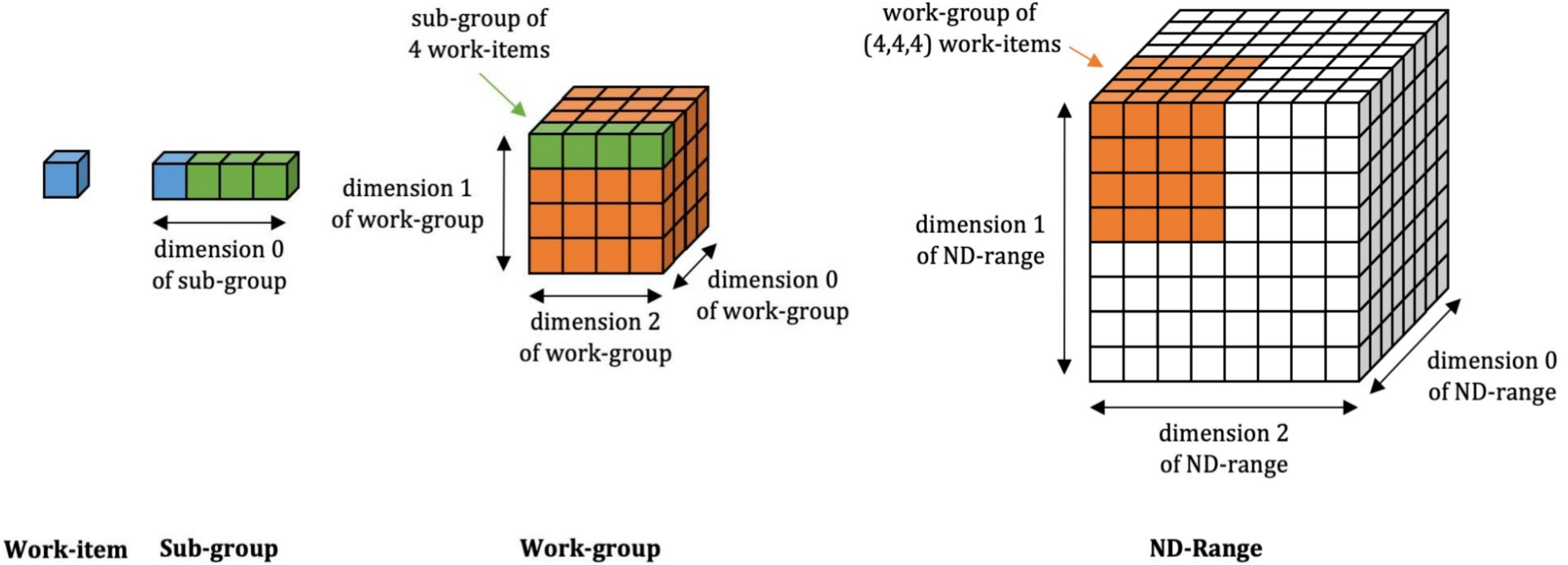
Write a SYCL program using Sub-Group and group algorithms to accomplish computation

Sub-groups

Sub-groups are a subset of the work-items that are executed Simultaneously or with additional scheduling guerantees.

Leveraging sub-groups will help to map execution to low level hardware and may help in achieving higher performance.

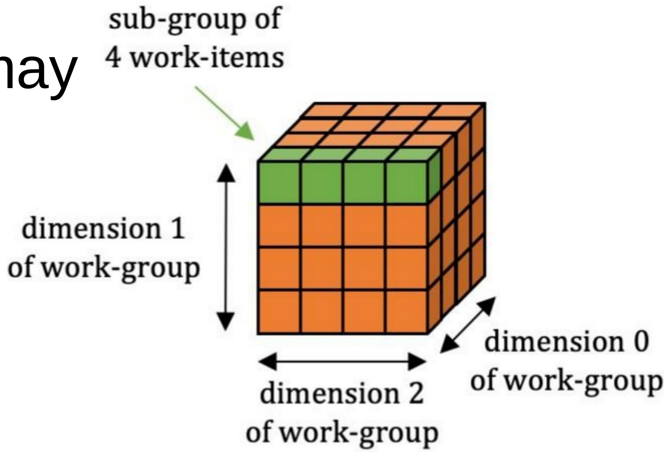
Sub-groups



sub-group of
4 work-items
work-group of
(4,4,4) work-items
dimension 0
of sub-group
dimension 1
of work-group
dimension 0
of work-group
dimension 2
of work-group
dimension 1
of ND-range
dimension 0
of ND-range
dimension 2
of ND-range
Work-item
Sub-group
Work-group
ND-Range

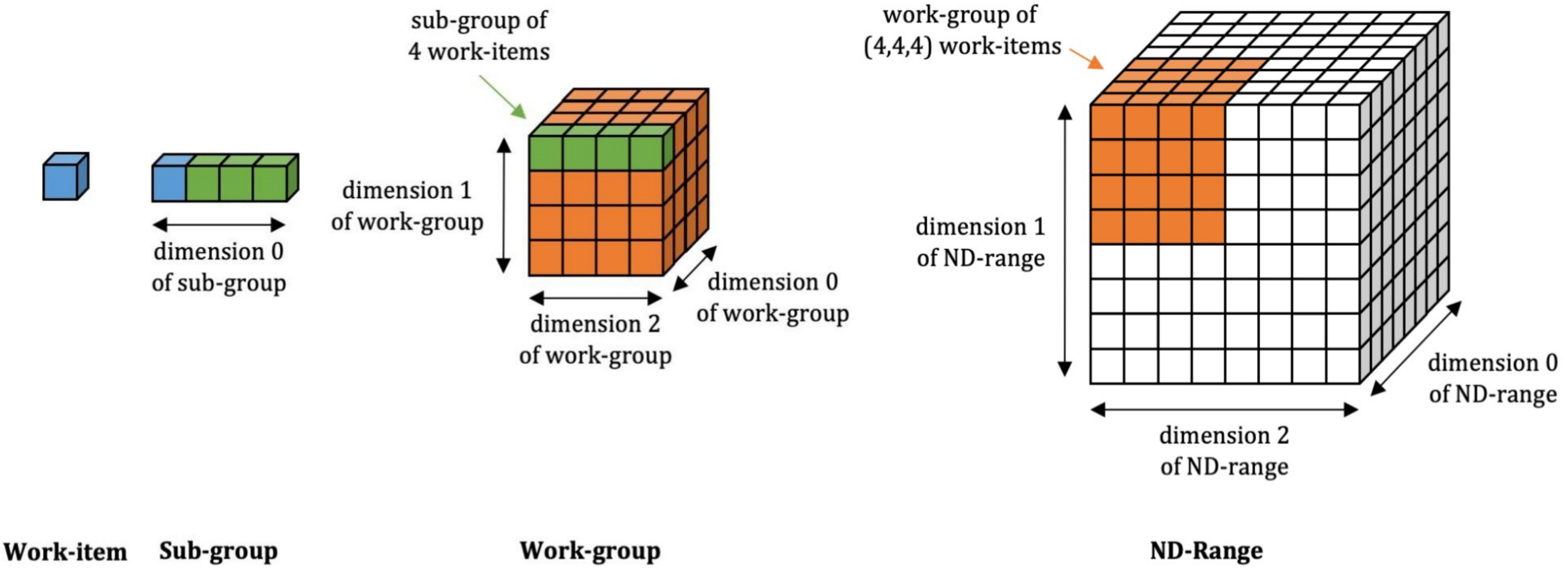
Sub-groups

- A subset of work-items withing a work-group that may map to vector hardware.
- Why use sub-groups?
  - Work-items in a sub_group can communicate directly using shuffle operations
  - Work-items in a sub_group can synchronize using sub_group barriers and guarantee memory consistency using sub_group memory fences
  - Work-items in a sub_group have access to sub_group collectives, providing fast implementations of common parallel patterns.



Sub-groups

- Sub-group = subset of work-items withing a work-group
- Parallel execution with ND-RANGE kernel helps to get access to work-group and sub-group



Sub-groups

```
h.parallel_for(nd_range<1>(N,B), [=](nd_item<1> item)
{
            auto sg = item.get_sub_group();
            // KERNEL CODE
});
```

sub_group class

- The sub-group handle can be obtained from the nd_item using the get_sub_group() .

- Once you have the sub-group handle, you can query for more information about the sub-group, do shuffle operations or use collective functions.

- Explicit kernel attribute

  [[intel::reqd_sub_group_size(N)]]

  to control the sub-group size

Sub-groups

The sub-group handle can be quired to get other information:

- get_local_id() returns the index of the work-item within its sub-group
- get_local_range() returns the size of sub_group
- get_group_id() returns the index of the sub-group
- get_group_range() returns the number of sub-groups within the parent work-group

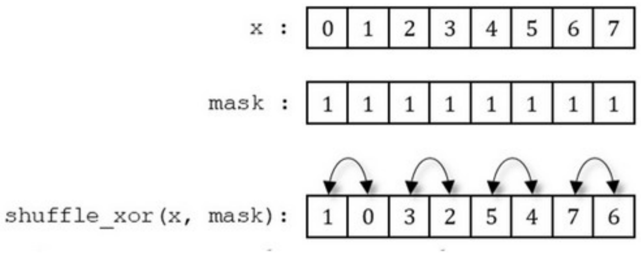
```
h.parallel_for(nd_range<1>(N,B), [=](nd_item<1> item){
        auto sg = item.get_sub_group();

        if(sg.get_local_id() == 0){
            out << "sub_group id: " << sg.get_group_id()[0]
                << " of " << sg.get_group_range()
                << ", size=" << sg.get_local_range()[0]
                                        << endl;
        }
});
```

```
sub_group id: 1 of 4, size=16
sub_group id: 3 of 4, size=16
sub_group id: 2 of 4, size=16
sub_group id: 0 of 4, size=16
```

Sub-group Shuffles

- One of the most useful features of sub-groups is the ability to communicate directly between individual work-items without explicit memory operations.
- Shuffle operations enable us to remove work-group local memory usage from our kernels and/or to avoid unnecessary repeated accesses to global memory.

```
h.parallel_for(nd_range<1>(N,B), [=](nd_item<1> item){
        auto   sg = item.get_sub_group();
        size_t i = item.get_global_id(0);

        /* Shuffles */
        //data[i] = sg.shuffle(data[i], 2);
        //data[i] = sg.shuffle_up(0, data[i], 1);
        //data[i] = sg.shuffle_down(data[i], 0, 1);
        data[i] = sg.shuffle_xor(data[i], 1);
});
```



Sub-group Collectives

- The collective functions provide implementations of closely- related common parallel patterns.
- Providing these implementations as library functions increases developer productivity and gives implementations the ability to generate highly optimized code for individual target devices.

```
h.parallel_for(nd_range<1>(N,B), [=](nd_item<1> item){
        auto sg = item.get_sub_group();

        size_t i = item.get_global_id(0);


        /* Collectives */

        data[i] = reduce(sg, data[i],   plus<>());

        //data[i] = reduce(sg, data[i], std::maximum<>());

        //data[i] = reduce(sg, data[i], std::minimum<>());
});
```

Sub Groups

Sub-Group Group Algorithms

- Group algorithms provide implementations of closely-related common parallel patterns.
- Providing implementations as library functions increases developer productivity and gives implementations the ability to generate highly optimized code for individual target devices.

```
h.parallel_for(nd_range<1>(N,B),[=](nd_item<1> item)
{
      auto sg = item.get_sub_group();
      size_t i = item.get_global_id(0);

      /* Collectives */
      data[i] = reduce(sg, data[i], plus<>());
      //data[i] = reduce(sg, data[i], maximum<>());
      //data[i] = reduce(sg, data[i], minimum<>());
});
```

Specifying the Sub-Group Size

The sub-group size can be configured separately for each kernel.

The set of available sub-group sizes is hardware-specific

```
q.parallel_for(range<1>(N),
               [=](id<1> id) [[intel::reqd_sub_group_size(16)]]
{
  // KERNEL CODE
});
```

The sub-group size can be tuned even for kernels that do not use the `sub_group` class (e.g. to tune for SIMD width and register usage).

Sub Groups

- **Summary**
  - What are Sub-Groups?
  - Why are they useful?
  - Learned about sub-group shuffle operations and using sub-group collectives

Reductions

A reduction produces a single value by combining multiple values in an unspecified order.

- Parallelizing reductions can be tricky because of the nature of computation and accelerator hardware.
- SYCL 2020 introduces a simplified approach for reductions in heterogenous programming

Simple Reduction

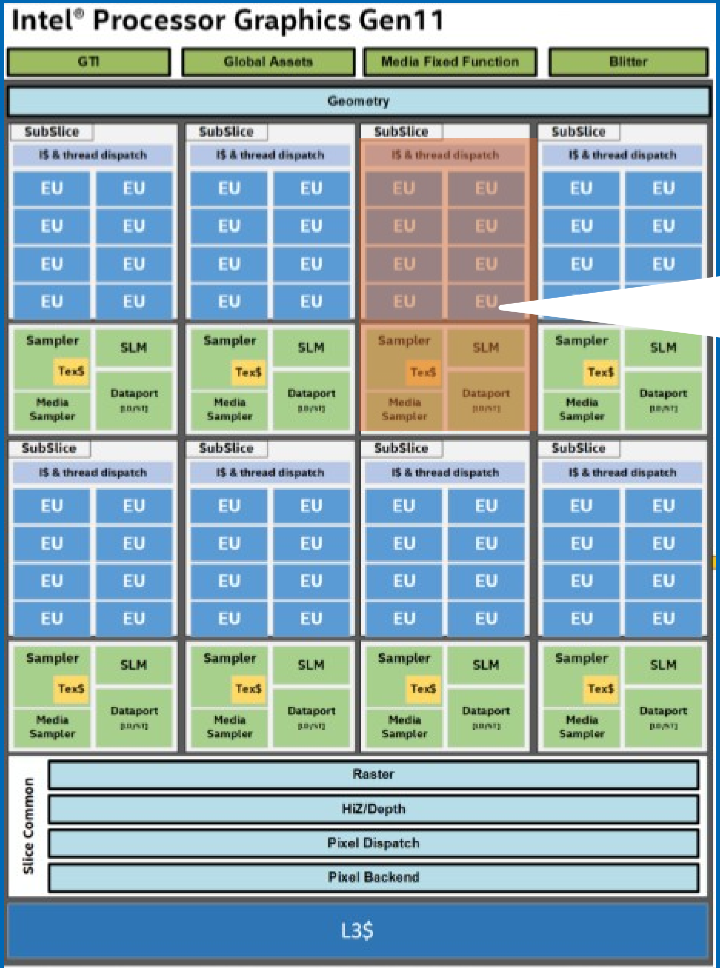
Let's look a simple reduction example: ***Addition of N items***

A simple for-loop in kernel function can accomplish reduction.

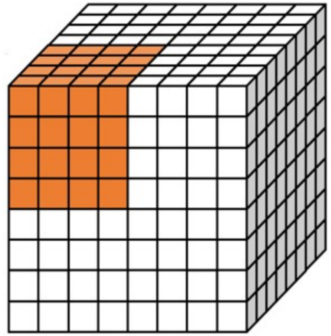
But, for-loop is not efficient and does not take advantage of parallelism in hardware.

```
queue q;
int *data = malloc_shared<int>(N, q);
for (int i = 0; i < N; i++) data[i] = i;

q.single_task([=]()
{
  int sum = 0;
  for(int i = 0; i < N; i++)
  {
      sum += data[i];
  }
  data[0] = sum;
}).wait();

std::cout << "Sum = " << data[0] << std::endl;
```

Parallelizing Reductions



work-group executions are mapped to Compute Units on hardware.



Reduction can be parallelized by first reducing items in each work-group using ND-range kernel, multiple work-groups can execute in parallel depending on number of compute units on hardware.

Work-Group Reduction

ND-Range kernel can be used to compute sum of all items in each work-group

reduce() function will simplify reduction of items in a work-group

A simple for-loop in single_task kernel function can then accomplish final reduction of each work-group sums.

```
q.parallel_for(nd_range<1>(N, B), [=](nd_item<1> item)
{
  auto wg = item.get_group();
  size_t i = item.get_global_id(0);

  //# Adds all elements in work_group using work_group reduce
  int sum_wg = reduce(wg, data[i], plus<>());

  //# write work_group sum to first location for each work_group
  if (item.get_local_id(0) == 0) data[i] = sum_wg;
});
```

```
q.single_task([=]()
{
  int sum = 0;
  for(int i=0;i<N;i+=B){
      sum += data[i];
  }
  data[0] = sum;
});
```

Some parallelism achieved but code is still complex with 2 kernel functions

Simplified Reduction

SYCL 2020 introduces reduction object in parallel_for

reduction object in parallel_for encapsulates the reduction variable, an optional operator identity and the reduction operator.

Removes the need for two step approach using two kernel functions.

```
queue q;
auto data = malloc_shared<int>(N, q);
for (int i = 0; i < N; i++) data[i] = i;

auto sum = malloc_shared<int>(1, q);
sum[0] = 0;

q.parallel_for(nd_range<1>{N, B},
          reduction(sum, plus<>()), [=](nd_item<1> it, auto& sum)
{
     int i = it.get_global_id(0);
     sum += data[i];
}).wait();

std::cout << "Sum = " << sum[0] << std::endl;
```

Multiple Reductions in one kernel

```
myQueue.submit([&](handler& cgh)
{

  // Input values to reductions are standard accessors (or USM pointers)
  auto inputValues = accessor(valuesBuf, cgh);

  // Create temporary objects describing variables with reduction semantics
  auto sumReduction = reduction(sumBuf, cgh, plus<>());
  auto maxReduction = reduction(maxBuf, cgh, maximum<>());

  // parallel_for performs two reduction operations
  cgh.parallel_for(range<1>{1024}, sumReduction, maxReduction,
    [=](id<1> idx, auto& sum, auto& max)
  {
      sum += inputValues[idx];
      max.combine(inputValues[idx]);
  });
});
```

https://www.khronos.org/registry/SYCL/specs/sycl-2020/html/sycl-2020.html#sec:reduction

Useful Links

Open source projects

oneAPI Data Parallel C++ compiler: github.com/intel/llvm

Graphics Compute Runtime: github.com/intel/compute-runtime

Graphics Compiler: github.com/intel/intel-graphics-compiler

SYCL 2020: tinyurl.com/sycl2020-spec

DPC++ Extensions: tinyurl.com/dpcpp-ext

Environment Variables: tinyurl.com/dpcpp-env-vars

DPC++ book: tinyurl.com/dpcpp-book

SYCL Academy github.com/codeplaysoftware/syclacademy/tree/main

Code samples:

github.com/intel/llvm/tree/sycl/sycl/test
github.com/intel/llvm/tree/sycl/sycl/test-e2e
github.com/oneapi-src/oneAPI-samples

Hands-on Exercises

SYCL Lab 2 - Unified Shared Memory

Notices and Disclaimers

- Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.
- The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document, with the sole exception that a) you may publish an unmodified copy and b) code included in this document is licensed subject to the Zero-Clause BSD open source license (0BSD), https://opensource.org/licenses/0BSD. You may create software implementations based on this document and in compliance with the foregoing that are intended to execute on the Intel product(s) referenced in this document. No rights are granted to create modifications or derivatives of this document.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document, with the sole exception that code included in this document is licensed subject to the Zero-Clause BSD open source license (OBSD), http://opensource.org/licenses/0BSD.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.
-

Back Up
Details about Intel® oneAPI Toolkits

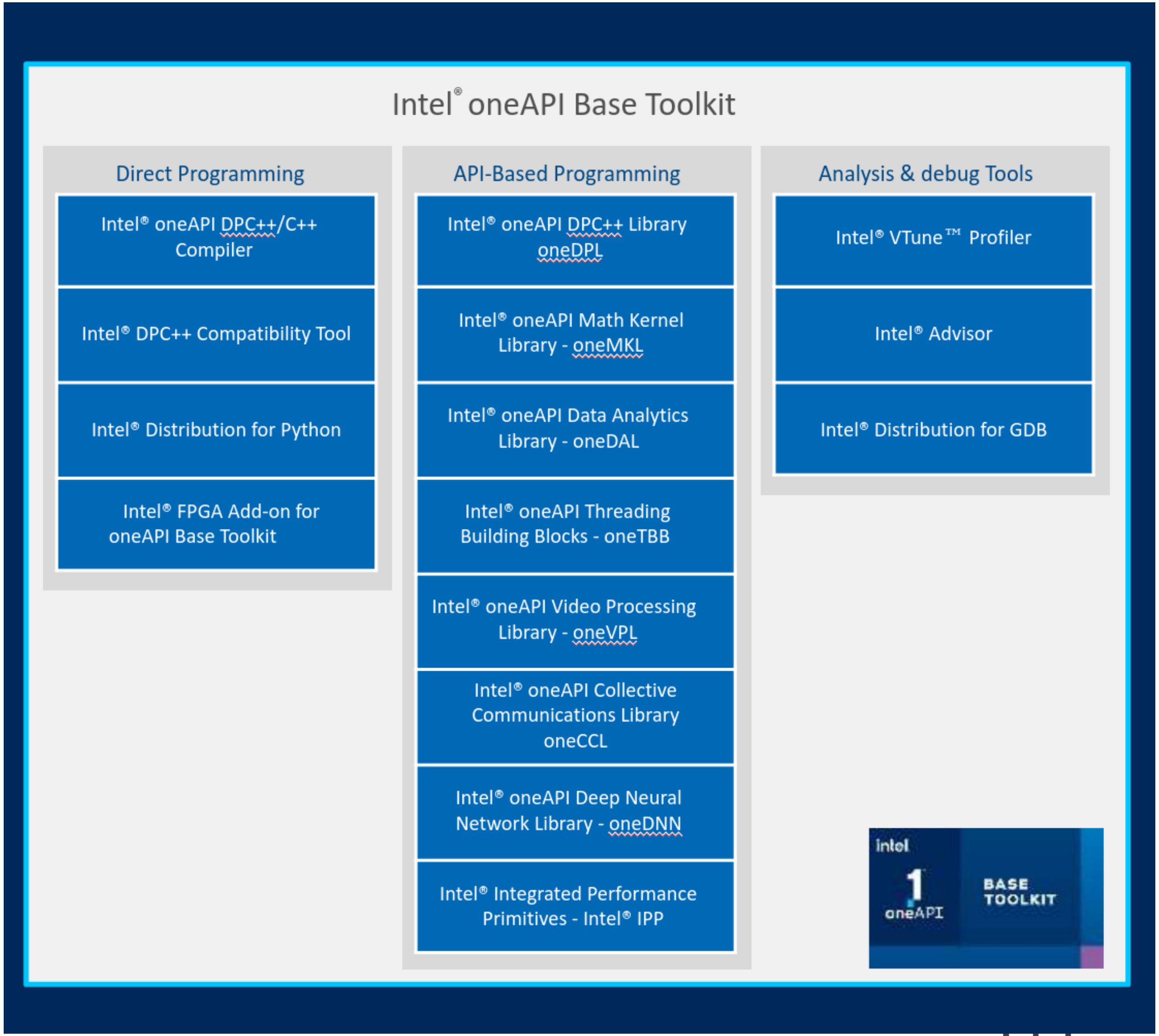
Intel® oneAPI Base Toolkit

Accelerate Data-centric Workloads

A core set of core tools and libraries for developing high-performance applications on Intel® CPUs, GPUs, and FPGAs.

Who Uses It?

- A broad range of developers across industries
- Add-on toolkit users since this is the base for all toolkits



Intel® oneAPI Base Toolkit

Accelerate Data-centric Workloads

Top Features/Benefits

- Data Parallel C++ compiler, library and analysis tools
- SYCLomatic / DPC++ Compatibility tool helps migrate CUDA code to C++ with SYCL
- Python distribution includes accelerated scikit-learn, NumPy, SciPy libraries
- Optimized performance libraries for threading, math, data analytics, deep learning, and video/image/signal processing

Learn More

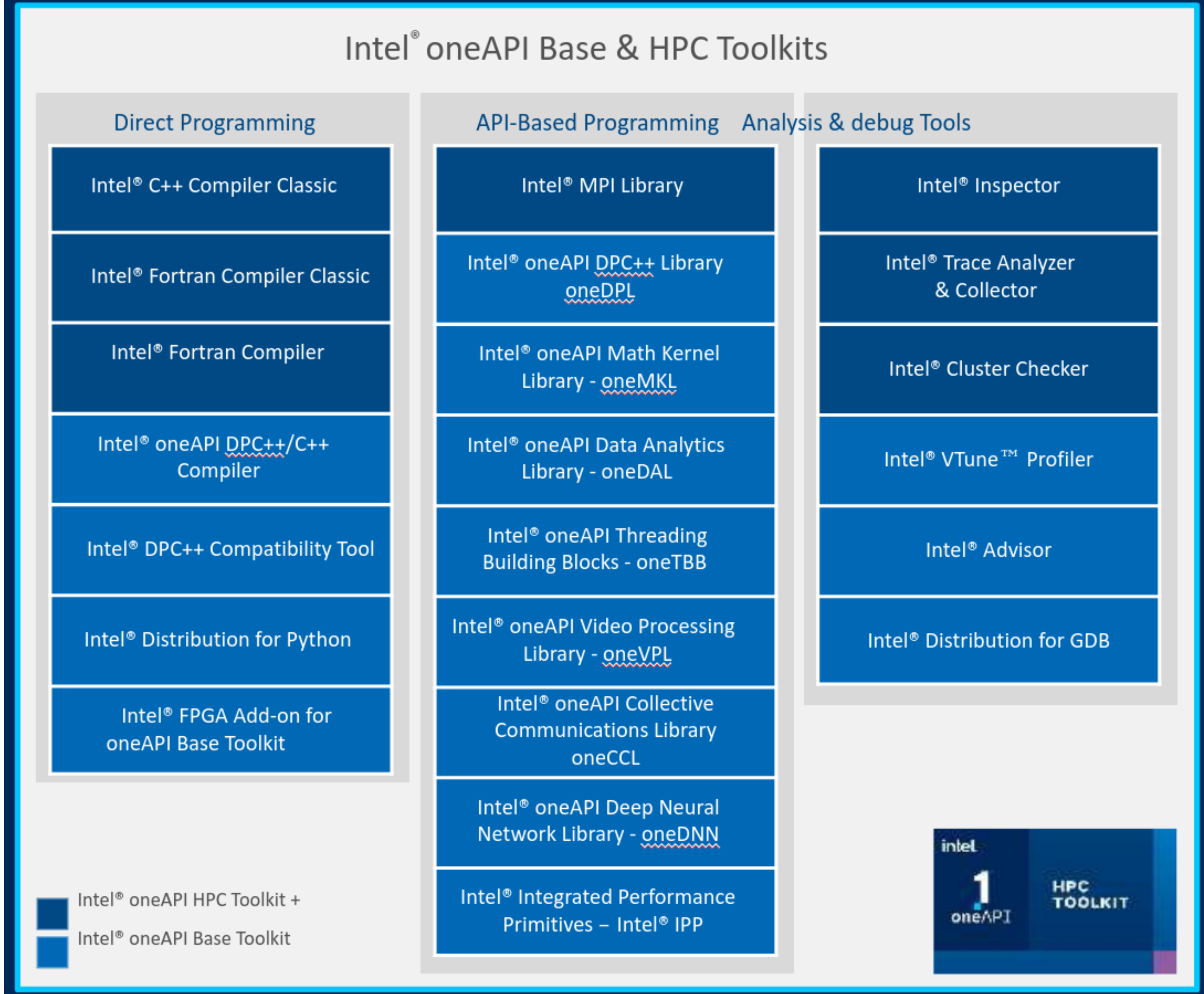
Intel® oneAPI HPC Toolkit

Accelerate Data-centric Workloads

A core set of core tools and libraries for developing high-performance applications on Intel® CPUs, GPUs, and FPGAs.

Who Uses It?

- A broad range of developers across industries
- Add-on toolkit users since this is the base for all toolkits



Intel® oneAPI HPC Toolkit

Accelerate Data-centric Workloads

Top Features/Benefits

- Data Parallel C++ compiler, library and analysis tools
- SYCLomatic / DPC++ Compatibility tool helps migrate CUDA code to C++ with SYCL
- Python distribution includes accelerated scikit-learn, NumPy, SciPy libraries
- Optimized performance libraries for threading, math, data analytics, deep learning, and video/image/signal processing

Summary

- oneAPI cross-architecture, one source programming model provides freedom of XPU choice.
  Apply your skills to the next innovation, not to rewriting software for the next hardware platform.

- Intel® oneAPI Toolkit products take full advantage of accelerated compute by maximizing performance across Intel CPUs, GPUs, and FPGAs.

- Develop confidently with a proven set of cross-architecture libraries and advanced tools that interoperate with existing performance programming models.

Intel® oneAPI HPC Toolkit

Accelerate Data-centric Workloads

Top Features/Benefits

- Data Parallel C++ compiler, library and analysis tools
- SYCLomatic / DPC++ Compatibility tool helps migrate CUDA code to C++ with SYCL
- Python distribution includes accelerated scikit-learn, NumPy, SciPy libraries
- Optimized performance libraries for threading, math, data analytics, deep learning, and video/image/signal processing