User Manual
Version 3.3.0

Table of Contents

Chapter 1

Getting Started

1.1 Requirements

- A C++ compiler compatible with C++-11
- One or more graphics cards compatible with OpenCL, CUDA, or Metal.
- The following operating systems are supported:
  - Linux (all distributions, glibc >= 2.18)
  - Windows
  - macOS >= 10.9
  - iOS >= 8.0
  - Android
- CPU: all common CPUs are supported (current linux builds include x86_64, i686, arm64, arm32, power8)
- optional: cmake >= 3.3

1.2 Installation

Linux

On linux, the script '`install.sh`' can be run as root to install goopax in the default system directories. Alternatively, the folder '`goopax-linux-3.3.0`' can be moved to a custom location.

MacOS

On MacOS, it is recommended to copy the folder '`goopax.framework`' either to '`/Library/Frameworks/`', or to '`$HOME/Library/Frameworks/`'.

Windows

No install scripts are currently provided for windows. Simply move '`goopax-windows-3.3.0`' to a custom location.

1.2.1 License File

The license file 'goopax_license.h' can be placed in the 'share/goopax/licenses' subfolder.

1.3 Building Programs

1.3.1 Building Programs with cmake

Cmake is the recommended way to build goopax programs. The required scripts are shipped with the goopax package.

If goopax is installed in a non-standard location, the path to <goopax>/share should be added to CMAKE_PREFIX_PATH.

If goopax_license.h is stored in a different location than 'share/goopax/licenses', the path to the folder where it is stored should be set in the 'GOOPAX_LICENSE_PATH' environment variable.

To use goopax in a cmake project, use the goopax package:

```
find_package(goopax)
```

And link the resulting target 'goopax::goopax' to your program:

```
target_link_libraries(my_program goopax::goopax)
```

1.3.2 Building Programs without cmake

When using other build systems, it is up to the user to set the paths correctly. The compiler needs to find the goopax header files, as well as the license file 'goopax_license.h' in its search path. The linker needs to link to the goopax library.

1.4 Building the Example Programs

The example programs are a good place to start. Precompiled versions can be found in the 'bin', 'bin32', or 'bin64' folders. The source code is located in folder 'examples'. The examples can be built with cmake in the usual way, e.g.,

```
cd examples
mkdir build
cd build
cmake ..
cmake --build .
```

Chapter 2

Programming with Goopax

2.1 Program Structure

2.1.1 Including Header Files

To use Goopax, include the header file from your source code:

```
#include <goopax>
```

For OpenGL interoperability, include

```
#include <goopax_gl>
```

For OpenCL interoperability, include

```
#include <goopax_cl>
```

For Metal interoperability, include

```
#include <goopax_metal>
```

2.1.2 Namespaces

[sec:namespaces] The basic Goopax functionalities are found in namespace 'goopax'. For simplicity, we will assume that you import this namespace:

```
using namespace goopax;   // Basic GOOPAX types, such as buffer and gpu_float.
```

Optionally, one of the following namespaces can be used, for debug and release mode data types, respectively:

```
using namespace goopax::debug::types;    // Debugging data types Tint, Tfloat, ...
using namespace goopax::release::types;  // Release data types Tint, Tfloat, ...
```

For some functions that are provided by the C++ standard library, an overloaded function may be provided in the namespace "`std`".

2.2 Memory

In the GPU kernels, there are four types of memory:

- Private memory is not shared. Each thread has its own private memory.
- Local memory is shared between the threads in a work group. It is only valid during thread execution. Local memory can be used for thread communication within a work group, as well as sharing small data structures during kernel execution between the threads of a work group.
- Global memory is usually referred to as the main memory of a video card. It can be used to store the data used in calculations, and to share data between all threads, and between GPU and CPU. Global memory is accessible by all threads, and also from the CPU.
- lvm memory can be used across different devices and CPUs, within a single unified address space.

2.2.1 Private Memory

Private memory is allocated as follows:

```
private_mem<float> A(16);
```

This will allocate an array of 16 floats for each thread.

2.2.2 Local Memory

Local memory is only available during the execution of a kernel. Each work group has its own local memory. Local memory is useful for communication within a work group. Between work groups no sharing is possible.

Local memory is declared as `local_mem`<*type*>, which has the constructor

```
local_mem(size_t size)
```

For example:

```
local_mem<double> mem(256);
```

will allocate 256 doubles in local memory for each group.

2.2.3 Global Memory

Global memory must be declared on the CPU side as a `buffer` and on the GPU side as a `resource`.

Buffer

On the CPU side, the memory is declared as `buffer`<*type*>, where the element type is specified as a template parameter. The constructor takes the argument

```
buffer(goopax_device device, size_t size)
```

For example:

```
buffer<float> buf(default_device(), 10);
```

will allocate a global buffer on the video card of type `float` and size `global_size()`.

Resource

The resource<type> is declared from within a GPU kernel and has to match a corresponding buffer. Resources can either be declared as parameters, e.g.,

```
struct my_kernel :
  public kernel<my_kernel>
{
  void program(resource<int>& A)
  {
    <GPU code...>
  }
}
```

or it can be declared within the kernel body and linked to a specific memory buffer:

```
struct my_kernel :
  public kernel<my_kernel>
{
  buffer<float> Bb(1000);
  // <...>
  void program(...)
  {
    resource<float> Br(Bb);
  }
}
```

When the former constructor is used, the corresponding memory buffer has to be specified as an argument when calling the kernel. With the latter constructor, the resource can already be linked to a specific buffer, so that the buffer does not have to be supplied when executing the kernel.

2.2.4 Svm Memory

Svm memory is declared in a similar way as global memory, e.g.,

```
buffer<float> buf(default_device(), 10);
```

Instead of declaring a resource in the kernel, svm memory is accessed via a pointer. The pointer can be provided to the kernel via a normal parameter, e.g.,

```
void program(gpu_type<double*> ptr)
```

Not all devices support svm memory. Whether svm memory is supported can be checked with the function device::support_svm().

2.2.5 Memory Access

All memory types can be accessed within a kernel by the usual [] operator or by iterators.

2.2.6 Data Transfer from Host to GPU and from GPU to Host

Data transfer via DMA to and from the video card can be done with the following member functions of global memory buffer types.

copy_to_host(T* p)

Copies data from the buffer to the address in host memory specified by p.

copy_to_host(T* p, size_t beginpos, size_t endpos)

Copies data from the buffer to the address in host memory specified by p, in the range from position beginpos to endpos-1 within the buffer.

copy_from_host(T* p)

Copies data from the host address p to the buffer.

copy_from_host(T* p, size_t beginpos, size_t endpos)

Copies data from the host address p to the buffer, in the range from position beginpos to endpos-1 within the buffer.

2.2.7 Barriers

To avoid race conditions, it is sometimes necessary to place barriers between memory accesses, especially when local threads are communicating with each other. Race conditions can occur when two or more threads access the same memory address and their access is not properly synchronized.

Resource-Specific Barriers

Barriers can be placed by calling the `barrier()` function of local or global resources, for example:

```
local_mem<float> a(local_size());
a[local_id()] = 2;
a.barrier();
gpu_float b = a[(local_id() + 5) % local_size()];
```

Without the barrier, a race condition would occur. Barriers can be placed on local memory or on global memory. Note: Barriers on global memory only synchronize memory access within a work group. Memory between different groups is only synchronized between kernel calls. However, synchronization of global memory between different work groups can be enforced during kernel execution by using atomic instructions (see section [sec:atomic]).

Global Barriers

Alternatively, you can use the `global_barrier()` function to force a barrier for all resources:

```
resource<float> A;
resource<float> B;

A[global_id()] = 5;
B[2*global_id()] = 7;

global_barrier();   // This will put barriers to resources A and B.

gpu_float x = A[0] + B[129];
```

However, one should usually try to use resource-specific barriers instead of global barriers to allow Goopax to make better optimizations.

2.3 User-Defined Types in Memory Access

In addition to using intrinsic types, memory access can also be done with user-defined classes. This can simplify code development and provide a more structured and safe way of accessing the data structures.

To use user-defines types, minor modifications need to be done to the class. This can be done in an intrusive way (easier and the method of choice if the data type is defined in your own program), or by external type definitions (necessary when using external data types for memory access).

The following restrictions apply:

- The class must fully contain all the data of the data structure. Members that use external memory, such as `std::vector` are not allowed. However, one can use members of type `std::array`, or other classes that don't allocate memory.
- Virtual functions or virtual base classes cannot be used.
- Data members must use special data types. This can either be a type specified as a template parameter, or one of the following types:
  - `float_type`, `double_type`
  - `int_type`, `int64_type`, `int32_type`, `int16_type`, `int8_type`
  - `uint_type`, `uint64_type`, `uint32_type`, `uint16_type`, `uint8_type`

  Pointers and intrinsic types are not allowed.

2.3.1 Intrusive Type Preparations

The following modifications need to be done:

- The class must be defined as a class template, where at least one template parameter is a type that is characteristic for the context in which it is used. This is necessary to distinguish between CPU context, GPU context, and for internal purposes of memory access. If you don't need a template parameter, a dummy parameter can be used that defaults to, e.g., int.
- Call the macro "`goopax_struct_defs(<T>)`" within the type definition, where `T` is the template parameter as described above. This will provide type definitions `int_type`, `float_type`, etc.
- Declare the template type `goopax_struct_changetype` with one template parameter (say, `X`) within the class as shown in the examples below. This provides a type change mechanism for your data type. `goopax_struct_changetype` should be defined as your class, where all the template parameters are changed by means of the `goopax::goopax_struct_changetype` struct.

Example: one template parameter

```
template <typename T> struct complex
{
  // Definitions needed by Goopax for memory access
  goopax_struct_defs(T)
```

```
  template <typename X> using goopax_struct_changetype =
    complex<typename goopax_struct_changetype<T, X>::type>;

  // Now the original content of the class.
  T real;
  T imag;
};

...
buffer<complex<float>> my_buffer(1000);
```

Example: two template parameters

```
template <typename A, typename B> struct pair
{
  goopax_struct_defs(A)
  template <typename X> using goopax_struct_changetype =
    pair<typename goopax_struct_changetype<A, X>::type,
         typename goopax_struct_changetype<B, X>::type>;

  A first;
  B second;
};
```

Example: dummy template parameter

```
template <typename T=int> struct myclass
{
  goopax_struct_defs(T)
  template <typename X> using goopax_struct_changetype =
    myclass<typename goopax_struct_changetype<T, X>::type>;

  float_type some_data;
  int_type some_other_data;
  std::array<double_type, 3> vector_data;
};
...
buffer<myclass<>> my_buffer(1000);
```

2.3.2 Using external classes for memory access.

If the above modifications cannot be done, for example because the data type is defined in an external library, it can still be used for memory access by the following modifications:

- define the template struct "goopax_struct_type" in namespace "goopax" that returns the characteristic data type.
- define the template struct "goopax_struct_changetype" in namespace "goopax" that provides a type change mechanism for your data type.

Example

```
// Complex data type as defined in the STL library.
namespace std
{
  template <typename T> struct complex
  {
    T real;
    T imag;
  };
}
```

```
// Extensions necessary for memory access.
namespace goopax
{
  template <class T> struct goopax_struct_type<std::complex<T>>
  {
    using type = T;
  };

  template <class T, class X> struct goopax_struct_changetype<std::complex<T>, X>
  {
    using type = std::complex<typename goopax_struct_changetype<T, X>::type>;
  };
}
```

2.3.3 Memory Access

Buffers and resources of the new type can be declared in the same way as buffers with intrinsic types, e.g.:

```
...
resource<complex<float>> my_global_resource;
local_mem<complex<float>> my_local_resource(local_size());
...

int main()
{
  ...
  buffer<complex<float>> my_buffer(default_device(), size);
}
```

The new type can be accessed in the usual C++ way. Some examples are shown here:

```
// assigning one item
my_global_resource[global_id()].real = 2.3;

// copying a complete element
my_local_resource[local_id()] = my_global_resource[12];

// modifying one item
my_global_resource[global_id()].imag += 25;
```

2.3.4 Pre-Defined Extensions for STL classes

Goopax already provides memory access to the following STL data types:

- std::complex
- std::pair
- std::tuple
- std::array

Memory resources can store data of the above types or of any combination:

```
  buffer<pair<complex<float>, int> > A(1000);
  A.fill(make_pair({1.0, 2.0}, 3.0));
```

2.4 Images

Special data access is provided for images and image arrays. From the host code, images can be allocated with objects of type "image_buffer". From the device code, images can be accessed by using objects of type "image_resource". For more information, see the goopax reference.

2.5 GPU Kernels

Kernels are the functions that run on the video card and are typically used for the computationally demanding calculations.

2.5.1 Writing Kernel Classes

GPU kernel classes are derived from the class template 'kernel', which takes the name of the user-defined kernel class as template argument. They should contain a function program(...) with the code which will be executed on the video card. A kernel might look like this:

```
class my_kernel :
  public kernel<my_kernel>
{
public:
  void program(<...resources...>)
  {
    <GPU code...>
  }
};
```

2.5.2 Kernel Instantiation

Before a kernel is first executed, it has to be instantiated, i.e., an object of the kernel class has to be created:

```
  my_kernel MyKernel;
```

The kernel can be declared as a static object. The kernel object is not bound to any specific device and can be executed on all available devices.

2.5.3 Calling a Kernel

The kernel is executed by calling the '()' operator and passing the required buffers as arguments for all the unspecified resources. For example, if the kernel declares two resources

```
class my_kernel :
  public kernel<my_kernel>
{
public:
  void program(resource<float>& A, const resource<int>& B)
  {
    ...
  }
};
```

then the kernel must be called with two buffers as arguments of the same type and in the same order, i.e.

```
buffer<float> A(default_device(), 100);
buffer<int> B(default_device(), global_size());
MyKernel(A, B);
```

The device on which the kernel is executed is deduced from the buffers that are passed as arguments.

Asynchronous kernel calls

All kernel calls are implicitly asynchronous, meaning that the call will immediately return, and the CPU can perform other calculations while the GPU executes the kernel. Only when the results are accessed, the CPU will wait for the kernel to finish.

2.5.4 Simple Example Program

In the following program, each thread calculates the square root of its thread ID.

```
#include <goopax>         // Include Goopax
using namespace goopax;


struct simple :             // Define kernel program 'simple'
  public kernel<simple>
{
  void program(resource<float>& A)      // This function contains the GPU source code.
  {
    gpu_float x = sqrt((gpu_float)global_id()); // Calculate square root of the thread ID
    A[global_id()] = x;                 // and write it to the resource.
  }
} Simple;


int main()
{
  goopax_device device = default_device();      // Let goopax choose a device.
  buffer<float> A(device, device.default_global_size());   // Allocate buffer of size global_size()

  Simple(A);                                    // Call the kernel.
  std::cout << "A=" << A << std::endl;          // Display the result.
}
```

2.6 Passing Values as Kernel Arguments

2.6.1 Passing Constants to the Kernel

Constant values can be passed as kernel arguments. The parameter must be specified as an argument of the `program` function as a GPU type, and the corresponding value must be passed when calling the kernel.

Example:

```
template <class T> struct fillprog :
  public kernel<fillprog<T>>
{
  using gpu_T = typename gettype<T>::gpu;
  void program(resource<T>& res, gpu_T val)  // Specifying parameters
  {
    gpu_for_global(gpu_uint k=0, res.size())
    {
      res[k] = val;
```

```
    }
  }
};

int main()
{
    fillprog<float> fill;
    buffer<float> A(default_device(), 100);
    fill(A, 3.14159);      // Passing the values
    cout << "A=" << A << endl;
}
```

2.6.2 Gathering Return Values

Return values can be gathered from all threads, by passing references to gather objects to the "program" function, or by returning them as return value. The values of all threads are combined and returned appropriately. Predefined gather classes are:

class	description
gather_add	calculate the sum of all values
gather_min	calculate the minimum of all values
gather_max	calculate the maximum of all values

Instead of using these pre-defined classes, user-defined classes can be used. For more information, see the example program "gather" or the header file "src/gather.h".

goopax_future

Gather values are wrapped into objects of type "goopax_future". Their behavior is similar to "std::future" of the standard C++ library. The actual values can be extracted by the "goopax_future::get()" function.

Gathering values as references

When using references, multiple gather values can be used.
Example:

```
struct testprog :
  public kernel<testprog>
{
  void program(gather_add<Tuint>& minID, gather_max<Tuint>& maxID)
  {
    minID = global_id();
    maxID = global_id();
  }
};

testprog P;

goopax_future<Tuint> minID;
goopax_future<Tuint> maxID;
P(minID, maxID);
cout << "minimum id=" << minID.get() << endl
     << "maximum id=" << maxID.get() << endl;
```

Gathering values as return values

Returning gathered values is done as shown in the following example:

```
struct testprog :
  public kernel<testprog>
{
  gather_add<Tfloat> program()
  {
    return global_id();
  }
};

testprog P;

goopax_future<Tfloat> sum = P();
cout << "sum␣of␣thread␣IDs=" << sum.get() << endl;
```

2.7 Data Types

2.7.1 Basic GPU Data Types

Special data types are used to declare local variables that reside on the video card. The programming of GPU kernels relies on these data types. The following basic types are available:

GPU type	Corresponding CPU type
gpu_double	double
gpu_float	float
gpu_int64	int64, long
gpu_uint64	uint64_t, unsigned long
gpu_int	int32_t, int
gpu_uint	uint32_t, unsigned int
gpu_int16	int16_t, short
gpu_uint16	uint16_t, unsigned short
gpu_int8	int8_t, signed char
gpu_uint8	uint8_t, unsigned char
gpu_char	char
gpu_bool	bool
gpu_type<T>	T

Variables of GPU type are used within the function "program()", or as local variables in other functions and classes that are called from the program() function. They may never be used as static variables, and they may never be alive outside the life span of the program() function.

2.7.2 CPU Types and GPU Types

In kernel development, we distinguish between variables of CPU type and of GPU type. This is not to be mistaken with the device type (section [sec:device]). Regardless of the device type, a CPU variable is an ordinary variable of the main program, whereas a GPU variable is a variable (typically a register) that resides on the device. Although kernel development relies on the use of GPU types, using CPU types within a kernel can sometimes simplify programming, and improve performance,

as they are treated as constant expressions from the perspective of the GPU kernel.

Both CPU and GPU types can be used together in kernel programs. Instructions that use CPU types are calculated during kernel compilation. They are the equivalent of what constant expressions are in ordinary programs, and they will not use any GPU resources.

Hence, variables can be sorted into four different types of life cycle:

- **CPU-based compile-time** – Constant expressions that may be evaluated by the C++ compiler. Has no effect on the runtime of your program.
- **CPU-based run-time** – CPU variables evaluated at runtime. Has no effect on GPU kernels.
- **GPU-based compile-time** – GPU variables that can be evaluated by Goopax during kernel creation. May increase compilation time of the GPU kernels, but has no effect on the kernel runtime.
- **GPU-based run-time** These are the runtime variables that are used in kernel.

2.7.3 Matching CPU and GPU Types

It is sometimes necessary to get the corresponding CPU type from a GPU type or vice versa, especially when the type in question is a template parameter. This can be done with the `gettype` struct, which takes a type as template parameter, and contains type definitions for the corresponding types. For a given input type `T` (regardless of whether T is a GPU type or a CPU type),

```
gettype<T>::cpu
```

is the CPU type, and

```
gettype<T>::gpu
```

is the GPU type.
Some Examples:

```
template <class T> struct foo
{                                  // If T is, say, float, then
  T x;                             // x is a float,
  typename gettype<T>::gpu y;      // y is a gpu_float,
  typename gettype<T>::cpu z;      // z is a float,

  typedef typename gettype<T>::gpu gpu_T;
  gpu_T x2;                        // x2 is a gpu_float,
  typename gettype<gpu_T>::gpu y2; // y2 is a gpu_float,
  typename gettype<gpu_T>::cpu z2; // and z2 is a float.
};
```

2.7.4 Changing GPU/CPU Types

When writing template functions that should be valid both for CPU types and for GPU types, it is sometimes necessary to specify a type without knowing whether it will be used by the CPU or by the GPU.

The "`gettype`" struct contains a helper struct template called "`change`" that can generate a type regardless of whether it is for CPU or for GPU context. It takes a CPU type as template argument.

```
template <class T> struct foo  // T may be, e.g., float or gpu_float
{
  using D = typename gettype<T>::template change<double>::type;
  // If T is float, then D is double. If T is gpu_float, then D is gpu_double.
}
```

2.7.5 Type Conversion

Implicit Type Conversion

To avoid performance pitfalls, implicit type conversion is stricter than the usual type conversion of C++. Conversion will only be done automatically if the precision of the resulting type is at least as large as of the source type. Also, type conversion is done implicitly from integral types to floating point types, but not in the other direction from floating point types to integral types. No implicit type conversion is done from signed integral type to unsigned integral type.
Some examples of implicit type conversion:

```
gpu_float a = 2.5;        // Conversion from CPU double to GPU float is ok.
gpu_double b = -3.7;
gpu_float c = b;          // Error: No implicit conversion from gpu_double to gpu_float.
gpu_float c = (gpu_float)b;  // With explicit type conversion ok.
gpu_double d = a + b;     // Implicit conversion to type with higher precision is ok.
```

Explicit Type Conversion

If the type conversion is not done implicitly, the type conversion can still be done explicitly in the usual C/C++ way, for example:

```
gpu_int64 a = 12345;
gpu_int b = (gpu_int)a;
gpu_int16 c = gpu_int16(b);
gpu_uint d = static_cast<gpu_uint>(b);
```

Type Conversion from CPU to GPU

Types are implicitly converted from CPU type to GPU type. The rules are slightly relaxed, as compared to GPU/GPU type conversion: CPU types can implicitly be converted to GPU types of lower precision. However, conversion from a CPU floating point type to a GPU integer type still requires explicit conversion.
Some examples:

```
int a = 5;
gpu_int b = a;
gpu_float c = 3.0;  // Conversion from double to gpu_float is ok.
gpu_int d = 3.0;    // Error: no implicit conversion from CPU floating point to GPU int.
```

Conversion in the other direction is not possible. A GPU type cannot be converted to a CPU type.

2.7.6 Reinterpret

Sometimes it is necessary to change the data type of a value without changing the binary content. This can be done by the `reinterpret` function templates. It is defined as:

```
template <class TO, class FROM>
  TO reinterpret(const FROM& from);
```

The source data is provided as a function parameter, and the destination type is provided as template argument. The "reinterpret" function can be used on various data types (GPU types, CPU types, arrays, arrays of arrays, user-defined GPU classes, etc.), and is meant as a safe substitute of the "union" construct of C/C++ with additional compile-time checks.
Some Examples:

```
gpu_int a = 5;

gpu_float b = reinterpret<gpu_float>(a);   // Reinterpreting gpu_int to gpu_float

gpu_double c = reinterpret<gpu_double>(a); // Error: different size!

array<gpu_int, 2> v = {2, 3};
gpu_double d = reinterpret<gpu_double>(v); // Array of 2 gpu_ints to gpu_double.
```

"reinterpret" also works on ordinary data types, e.g.,

```
int i = 2;
float f = reinterpret<float>(v);
```

In C/C++, this would correspond to

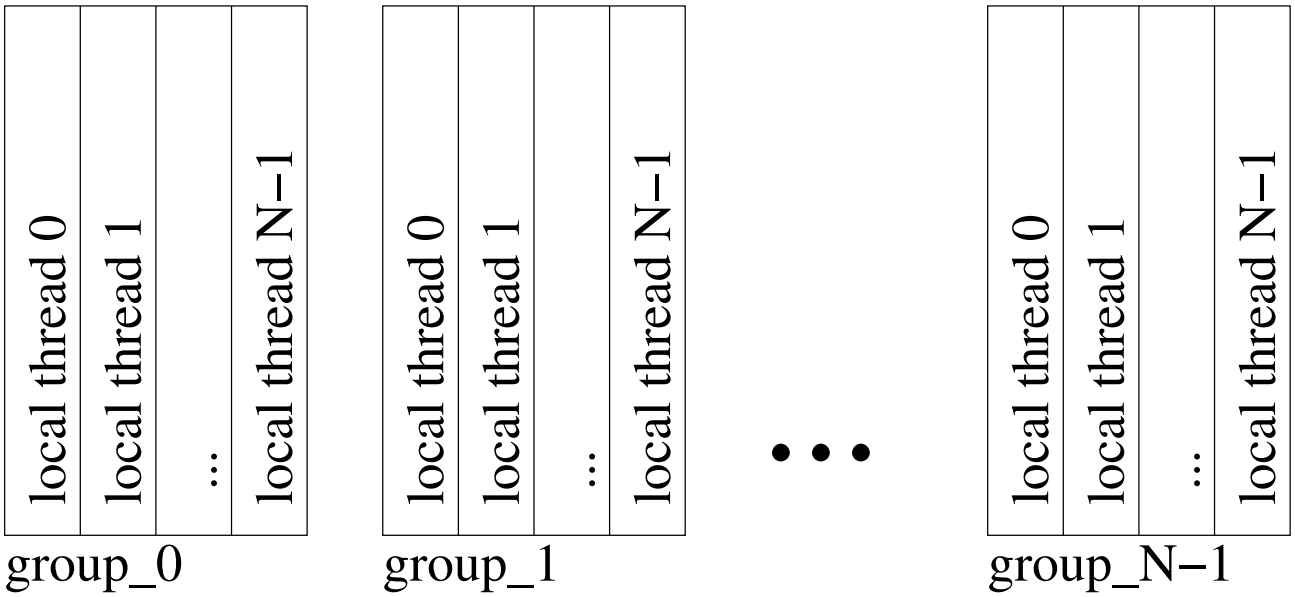
```
union
{
  int i;
  float f;
} u;
u.i = 2;
float d = u.f;
```

2.8 Thread Model

The major difference between CPUs and GPUs is the level of parallelism. While CPUs are designed for serial computation, GPUs work vastly parallel, with thousands of threads working in parallel on a common goal.

2.8.1 Thread Numbers

Kernels are executed by all GPU threads in parallel. Each thread can query its ID by the functions `local_id()`, `group_id()`, and `global_id()` as described below. The threads are organized into several work groups, where each work group in turn consists of several local threads. Depending on the hardware, the threads in a work group may or may not work in unison, executing SIMD instructions that apply to all threads in that group at once. In any case, threads within a work group have better means of communicating with each other than threads that are in different work groups.



The number of local threads per group is given by `local_size()`. The number of groups is given by `num_groups()`. The total number of threads is therefore the product

```
global_size() = num_groups() * local_size().
```

Within a kernel, the local thread id can be queried by `local_id()`, the group id by `group_id()`, and the global id by `global_id()`. The global id is calculated as

```
global_id() == group_id() * local_size() + local_id().
```

For simple programs that can easily be parallelized, it may be sufficient to ignore the detailed thread model and to simply assume that the total number of threads is `global_size()`, and that each individual thread has the ID `global_id()`. For more complex programs, it may be beneficial to take the thread model into account and to separate the threads into work groups.

The group size and the number of groups can vary from one video card to another. The programmer should normally not make assumptions on these values, but always query them from `local_size()` and `num_groups()`. Although it is possible to manually set these values by the member functions "`force_local_size()`", "`force_num_groups()`", etc (see reference guide), one should normally not do this and let Goopax decide which thread sizes to use.

2.8.2 Threads of the same Work Group

Work groups are assumed to work in sync. They cannot branch off into different if-clauses or loops. Whenever only some threads of a work group enter an if clause, the other threads in the group must wait. The same is true for loops. To optimize performance, one should usually make sure that either all threads in a work group enter an if-clause or none, and that the number of loop executions is similar or equal for all threads in the group in order to avoid waiting times.

Threads of the same work group can communicate with each other via local memory or global memory.

2.8.3 Different Thread Groups

Different work groups can branch off into different if-clauses or for-loops. However, they cannot easily communicate with each other within one kernel execution. Atomic operations (section [sec:atomic])

offer some means of communication between different work groups.

Another possibility for communication between threads in different work-groups is to wait until the kernel execution has finished. Between two kernel calls it is guaranteed that all memory resources are synchronized.

2.9 Flow Control

2.9.1 Conditional Function

Simple if-statements can be expressed by a conditional function (the equivalent of the C/C++ conditional operator "a ? b : c"):

```
T cond(gpu_bool condition, T return value if true, T return value if false)
```

2.9.2 If Clauses

For more complex statements that cannot be expressed as a conditional move, gpu_if can be used instead. Example:

```
gpu_if (local_id() == 0)
{
  a = 17;
}
```

2.9.3 Loops

gpu_for

Usage:

```
gpu_for(gpu_type var=begin, end [,step] [,comparison operator])
{
  // ... do something ...
}
```

The default comparison operator is "<", and the default step size is 1. Examples:

```
gpu_for(gpu_uint i=0, 10)  // Counts from 0 to 9.
{
  ...
}

gpu_for(gpu_uint i=0, 10, 2, "<=")  // Counts from 0 to 10 with step 2.
{
  ...
}
```

gpu_for_global

The gpu_for_global loop is parallelized over all threads. It is provided for convenience.

```
gpu_for_global(gpu_uint i=0, N)
{
  ...
}
```

is equivalent to

```
gpu_for(gpu_uint i=global_id(), N, global_size())
{
  ...
}
```

gpu_for_local

The gpu_for_local loop is parallelized over all threads in a work-group.

```
gpu_for_local(gpu_uint i=0, N)
{
  ...
}
```

is equivalent to

```
gpu_for(gpu_uint i=local_id(), N, local_size())
{
  ...
}
```

gpu_break

Breaks out of a loop. It is called as a function of the loop variable. Usage example:

```
gpu_for(gpu_int i=0, 10)
{
  gpu_if(i==5)
  {
    i.gpu_break();
  }
}
```

It is also possible to break out of a two-dimensional loop:

```
gpu_for(gpu_uint a=0, 10)
{
  gpu_for(gpu_uint b=0, 10)
  {
    gpu_if(a + b == 15)
    {
      a.gpu_break();   // Break out of both loops.
    }
  }
}
```

C-Style Loops

Traditional for-loops can also be used. They require that all loop boundaries are CPU values and therefore known to Goopax. The advantage is that the loop variable is also a CPU value, so that it can be used to address elements in vectors or arrays. In the following example, the sum of all elements of a 3D vector is calculated. The loop will be explicitly unrolled and all the calculation is done in registers.

```
array<gpu_float, 3> x;    // A 3D vector

gpu_float sum = 0;
for (int k=0; k<3; ++k)
{
  sum += x[k];
}
```

Warning: Using traditional C-style loops results in explicit loop unrolling. They should only be used in GPU kernels when the number of loop cycles are reasonably low!

2.10 Atomic Operations

Atomic memory operations are guaranteed to be indivisable and thread-safe. For example, if two threads atomically increase the value at the same memory location by 1, then the value will be increased by 2 in total, without the danger of getting into a race condition.

Atomic memory operations are supported on both global and local memory. The referenced memory objects must be 32 bit integers, or 64 bit integers. Atomic operations on 64 bit integers is not supported on all devices.

Example:

```
void program(resource<Tuint>& a)
{
  // Each thread adds 5 to a[0].
  gpu_uint k = atomic_add(a[0], 5);
}
```

2.10.1 Atomic Functions

The following atomic functions are supported.

In these functions, **REF** is a reference to a location in global or local memory, and may reference to a value within a user-defined struct. **T** is the value type of the memory and must be one of "gpu_int", "gpu_uint", "gpu_int64", "gpu_uint64".

T atomic_add(REF, T value)

Adds *value* to the memory location referenced by *REF*. The original value is returned.

T atomic_sub(REF, T value)

Subtracts *value* to the memory location referenced by *REF*. The original value is returned.

T atomic_min(REF, T value)

Calculates the minimum value. The original value is returned.

T atomic_max(REF, T value)

Calculates the maximum value. The original value is returned.

T atomic_and(REF, T value)

Calculates bitwise and. The original value is returned.

T atomic_or(REF, T value)

Calculates bitwise or. The original value is returned.

T atomic_xor(REF, T value)

Calculates bitwise xor. The original value is returned.

T atomic_xchg(REF, T value)

Exchanges *value* with the memory location referenced by *REF*. Returns the original value. *T* must be one of "gpu_int", "gpu_uint", "gpu_float", "gpu_int64", "gpu_uint64", or "gpu_double".

T atomic_cmpxchg(REF, T cmp, T value)

Compares *cmp* with the memory location referenced by *REF*. If both are the same, write *value* to the memory location. The original value is returned. *T* must be one of "gpu_int", "gpu_uint", "gpu_int64", "gpu_uint64".

Chapter 3

Operators and Functions

All the usual C++ operators and math functions have been overloaded for the GPU types and may freely be used. Only the conditional operator ("`a ? b : c`") must be replaced by the conditional function (see section [sec:cond]).

3.1 Operators

3.1.1 Operators for Floating Point numbers and Integers

- Arithmetic: "+", "-", "*", "/"
- Increment/Decrement: "++", "--"
- Comparison: "==", "!=", ">", ">=", "<", "<="
- Assignment: "="
- Compound assignment operators: "+=", "-=", "*=", "/="

3.1.2 Operators for Integers

- shift operators:
  - "<<"
  - ">>"

  Left/right shift a by b bits. b must be smaller than the number of bits in a.
- "&" (binary and), "|" (binary or), "^" (binary xor)
- "%" (modulo operator)
- Compound assignment operators: "<<=", ">>=", "&=", "|=", "^=", "%=" %=

3.1.3 Boolean Operators

- && (boolean and)
- "||" (boolean or)
- "!" (boolean not)

3.2 Floating Point Functions

Mathematical functions on GPU types can be used in the same way as they are used on CPU types, for example:

```
gpu_float x = 0.5;
gpu_float s = exp(x);
```

Goopax will use the best implementation for the given video card, based on performance measurements. It will select between native functions, OpenCL implementations, and Goopax implementations.

3.2.1 Unary Functions

- cos, cospi, acos, acospi
- sin, sinpi, asin, asinpi
- tan, tanpi
- sinh, asinh
- cosh, acosh
- tanh
- exp, exp2
- log, log2
- sqrt, cbrt
- erf, erfc
- tgamma
- ceil, floor, round
- isfinite, isinf, isnan, isnormal

3.3 Integer Functions

T clz(T)

Returns the number of leading zeros.

gpu_uint popcount(T)

Counts the number of bits set to 1.

T rotl(T a, gpu_int bits)

Rotate left. *bits* can be any number, positive, negative, or zero.

T rotr(T a, gpu_int bits)

Rotate right. *bits* can be any number, positive, negative, or zero.

3.4 Functions for Integers and Floats

gpu_T min(gpu_T a, gpu_T b)

Returns the minimum value of a and b.

gpu_T max(gpu_T a, gpu_T b)

Returns the maximum value of a and b.

gpu_TU abs(gpu_T a)

Returns the absolute value of *a*. If a is a signed integer, the result is an unsigned integer.

3.5 Work-Group Functions

gpu_bool work_group_any(gpu_bool x)

Returns true if x is true for any thread in the work-group.

gpu_bool work_group_all(gpu_bool x)

Returns true if x is true for all threads in the work-group.

T work_group_reduce_add(T x)

Returns the sum of all values of x in the work-group.

T work_group_reduce_min(T x)

Returns the minimum value of x in the work-group.

T work_group_reduce_max(T x)

Returns the maximum value of x in the work-group.

T work_group_broadcast(T x, gpu_uint local_id)

Broadcasts value x to thread local_id in the work-group.

work_group_scan_inclusive_(T x)
work_group_scan_exclusive_(T x)

Does a prefix-sum operation over all threads in the work-group. may be sum, min, or max. Example:

local_id	0	1	2	3	4 ...
x	1	0	2	5	1
work_group_scan_inclusive_add	1	1	3	8	9
work_group_scan_exclusive_add	0	1	1	3	8

3.6 Random Numbers

Goopax provides a WELL512 random number generator. It is very fast – all the calculation is done in registers – at the expense that the random numbers should be used in blocks.

To use the random numbers, it is first necessary to create an object of type `WELL512`, which contains a buffer object to store the seed values. Then, in the GPU kernel program, an object of type `WELL512lib` should be created, which will provide the random numbers. The constructor of `WELL512lib` takes the `WELL512` object as input parameter.

The function `WELL512lib::gen_vec<`*type*`>()` returns a vector of random numbers of the specified type (which can be of floating point or integral type). The size of this vector depends on its type. It is 8 for 64-bit values, 16 for 32-bit values, and larger for smaller integer types. The best performance is achieved if all values in the vector are used before a new vector is generated. Floating point random numbers are in the range $0..1$, unsigned integers from 0 to the maximum value, integers from the largest negative value to the largest positive value.

See the following example:

```
struct random_example :
  kernel<random_example>
{
  WELL512 rnd;                  // Object for the seed values

  void program()
  {
```

```
    WELL512lib rnd(this->rnd);  // Instantiate the random number generator

    vector<gpu_float> rnd_values = rnd.gen_vec<float>(); // Generate random floats
    ...
  }
};
```

Chapter 4

Error Checking Mechanisms

4.1 Overview

Goopax offers extensive support for automatic error checking. This includes

- Checking for overflow of memory resources
- Checking that all variables are properly initialized
- Detecting race conditions

These error checking mechanisms can be enabled to look for bugs in the program. They are not intended for use in release mode, because they cannot be run on the video card, and they also reduce performance for CPU mode.

4.2 Enabling Error Detection Mechanisms

For error detection mechanisms, special debug data types are provided in the namespace goopax::debug::types. Debug types are prefixed by "T", e.g., "`Tfloat`", "`Tuint64_t`", or "`Tbool`". In difference to the corresponding intrinsic types, debug types will detect and report the use of uninitialized variables. If used in buffers, race conditions will be detected.

4.2.1 Using the Debug Namespace

The proposed way is to import the namespace `goopax::debug::types` in debug mode and the namespace `goopax::release::types` in release mode.

Your program could start like this:

```
#include <goopax>

using namespace goopax;
#if USE_DEBUG_MODE
using namespace goopax::debug::types;
#else
using namespace goopax::release::types;
#endif
```

Here, the debug mode is enabled by the compiler switch "USE_DEBUG_MODE".

4.2.2 Enabling Error Checking in Kernels

To enable error checks in the kernels, all `buffer`, `resource`, and `local_mem` types should use `T`-prefixed types, like this:

```
struct my_kernel :
  kernel<my_kernel>
{
  buffer<Tdouble> A;

  void program()
  {
    resource<Tdouble> A(this->A);
    local_mem<Tint> B(2*local_size());
    ...
  }
  ...
};
...
```

This will enable all error checks in the kernels, if the data types from the debug namespace are used.

4.2.3 Extending the Error Checks to the CPU Program

Error checking mechanism can also be used in the CPU program, by using the T-prefixed data types throughout the program, instead of intrinsic c++ data types. This will provide extensive checks for variable initialization.

This should work in most cases. However, be aware that this may cause compilation errors from time to time (for example, the `main` function requires plain C++ types, as do constant expressions). Such errors have to be resolved by explicit type conversions or by reverting to the C++ intrinsic types.

Warning: Be cautious about `sizeof`! The sizes of debug variables are larger than the sizes of the original data types. Use

```
gettype<}{\it type}}>::size
```

instead, it will return the size of the original, intrinsic C++ type, and can be used on GPU types as well:

```
sizeof(int)             // returns 4
sizeof(Tint)            // undefined, don't use!
sizeof(gpu_int)         // undefined, don't use!
gettype<Tint>::size     // returns 4
gettype<gpu_int>::size // returns 4
```

4.3 Running Programs in Debug Mode

The debug mode is only available on the CPU device. The CPU device can be selected with `env_CPU` when calling the functions `default_device` or `devices`. By default, the number of threads is equal to the number of available CPU cores, and the number of groups is 1. The number of threads can be

changed by calling the `force_local_size` and `force_num_group` member functions of the device. This may be helpful to mimic the behavior of the video card.

4.4 Debugging Errors

When an error is detected, an appropriate error message is generated and the program is terminated. A debugger can be used to pin down the point where the error occurs.

To do this, it is helpful to disable optimization and to enable debugging symbols by passing appropriate compiler options with the `GOOPAX_CXXADD` environment variable and to disable coroutines with `GOOPAX_COROUTINES=0`.
Example:

```
GOOPAX_CXXADD='-O0 -ggdb3' GOOPAX_COROUTINES=0 gdb ./my_program
```

Chapter 5

OpenCL Interoperability

Goopax can be used in conjunction with existing OpenCL code. For this to work, the same OpenCL platform, context, and device must be used in Goopax and in your OpenCL code, and the same OpenCL queue should be shared between OpenCL and Goopax. Memory buffers can then be shared between Goopax and OpenCL.

To use OpenCL Interoperability, the header file <goopax_cl> must be included:

```
#include <goopax_cl>
```

To see how OpenCL interoperability is applied, also see the example programs "cl_interop_1" and "cl_interop-2".

5.1 Accessing Goopax Resources from OpenCL

The following functions provide access to Goopax resources from OpenCL code.

Platform:

```
cl_platform_id get_cl_platform()
```

Returns the OpenCL platform that is used by Goopax.

Context:

```
cl_context get_cl_context()
cl::Context get_cl_cxx_context()
```

Returns the OpenCL context that is used by Goopax.

Device:

```
cl_device_id get_cl_device()
```

Returns the OpenCL device that is used by Goopax.

Buffers and Images:

```
template <class BUF> inline cl_mem get_cl_mem(const BUF& buf)
template <class BUF> inline cl::Buffer get_cl_cxx_buf(const BUF& buf)
```

Returns the OpenCL memory handle for the Goopax buffer or image.

Chapter 6

OpenGL Interoperability

Goopax can share memory resources with OpenGL.

6.1 OpenGL Initialization

To enable OpenGL support, the goopax device must be retrieved from the `get_device_from_gl` function.

6.2 Sharing Buffers and Images with OpenGL

Goopax images and buffers can be created from existing OpenGL objects by using the static member functions "`create_from_gl`".

```
buffer::create_from_gl(goopax_device device, GLuint GLres, uint64_t cl_flags=CL_MEM_READ_WRITE)
  image_buffer::create_from_gl(GLuint GLres, uint64_t cl_flag=CL_MEM_READ_WRITE,
                               GLuint GLtarget=GL_TEXTURE_2D, GLint miplevel=0)
```

GLres: The OpenGL object ID.

cl_flags: The OpenCL access mode.

GLtarget: The OpenGL object type.

GLint: The OpenGL miplevel

For example, if "`gl_id`" is the ID of a 2-dimensional OpenGL texture with 32 bit data size, then

```
  image_buffer<2, uint32_t> A = image_buffer<2, uint32_t>::create_from_gl(gl_id)
```

creates an image "`A`" that can be used with Goopax.

Chapter 7

Example Programs

The source code of these programs is included in the goopax package.

7.1 pi

This program approximates the value of π in a very simple way: It uses a WELL512 random number generator to produce points in a 2-dimensional space $0 < x < 1$ and $0 < y < 1$ and counts, how many of those points lie within a circle of radius 1 from the origin, i.e. $x^2 + y^2 < 1$. This fraction multiplied by 4 approximates the value π. This program supports MPI and can be run with the `mpirun` command to combine the computing power of several hosts or video cards.

7.2 Mandelbrot

This program calculates a Mandelbrot image and uses OpenGL to draw it on the screen. It is an interactive program that can be controlled with the left mouse button and the forward and backward arrow keys.

7.3 Deep Zoom Mandelbrot

This is a somewhat more complex Mandelbrot program, combining arbitrary-precision arithmetics on the CPU with a special algorithm of our design. It allows to zoom in to large magnification levels (factor 10^{80} and more), and still display the result in real time.

7.4 fft

This program applies Fourier transforms on live camera images, and filters out high frequencies or low frequencies.

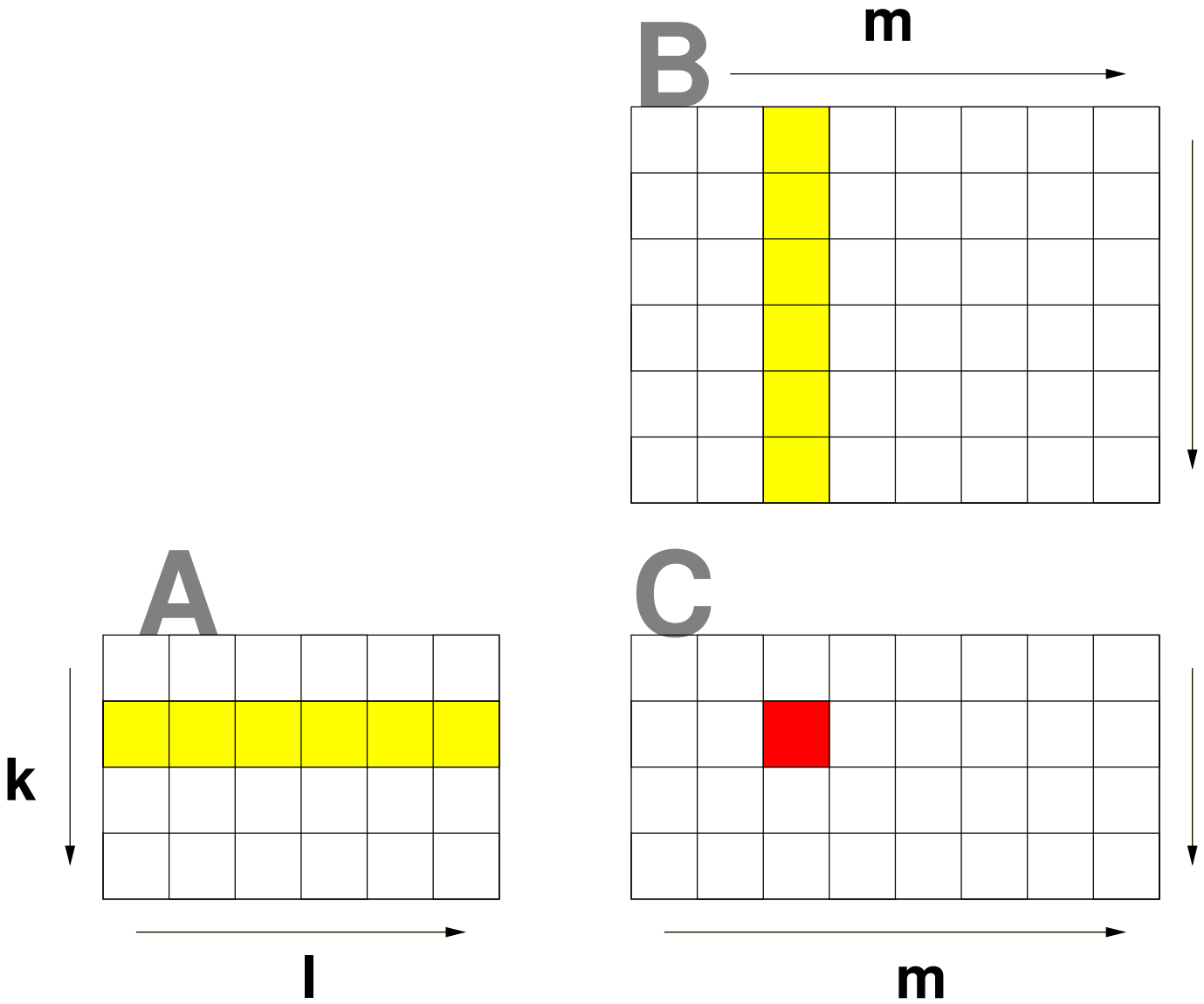
7.5 nbody

N-body simulation. Two colliding galaxies are simulated, each consisting of particles. Each particle interacts via gravity with all other particles. The particles are displayed via OpenGL.

7.6 matmul

The example program '`matmul`' performs matrix multiplications. It uses the nave multiplication algorithm. For productive purposes, faster algorithms such as Strassen's algorithm should be considered as well. The major performance bottleneck is the memory access. Three different algorithms are implemented that use different techniques to reduce the access to global memory.

7.6.1 Nave Implementation

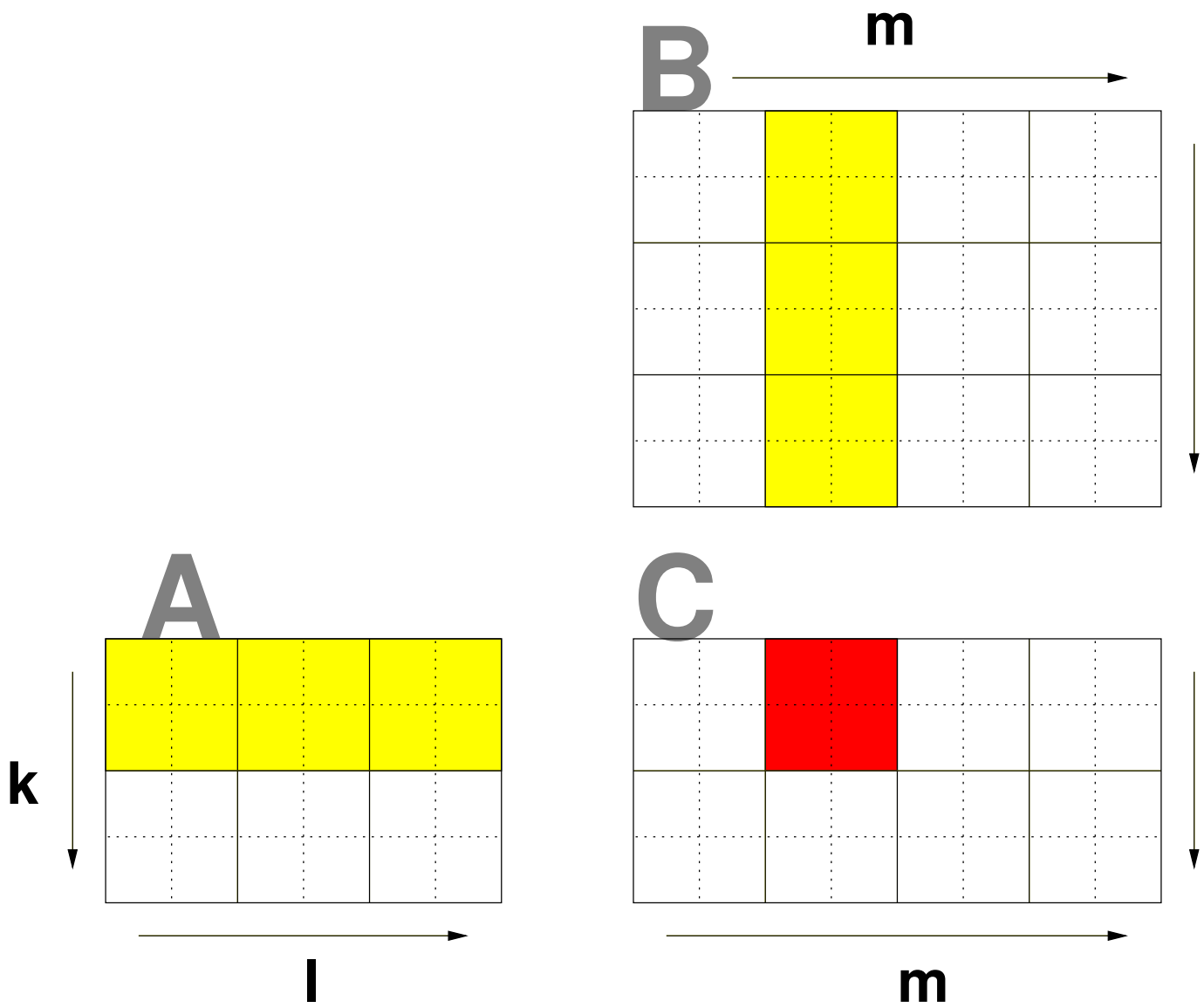


In the nave implementation, a 2-dimensional loop is executed over all matrix elements in the result matrix C, with indices k and m. This loop is parallelized over all threads, so that every element in

C is handled by exactly one thread. For every matrix element, the value $C[k][m]$ is calculated as a dot product of the k'th row of matrix A and the m'th column of matrix B.

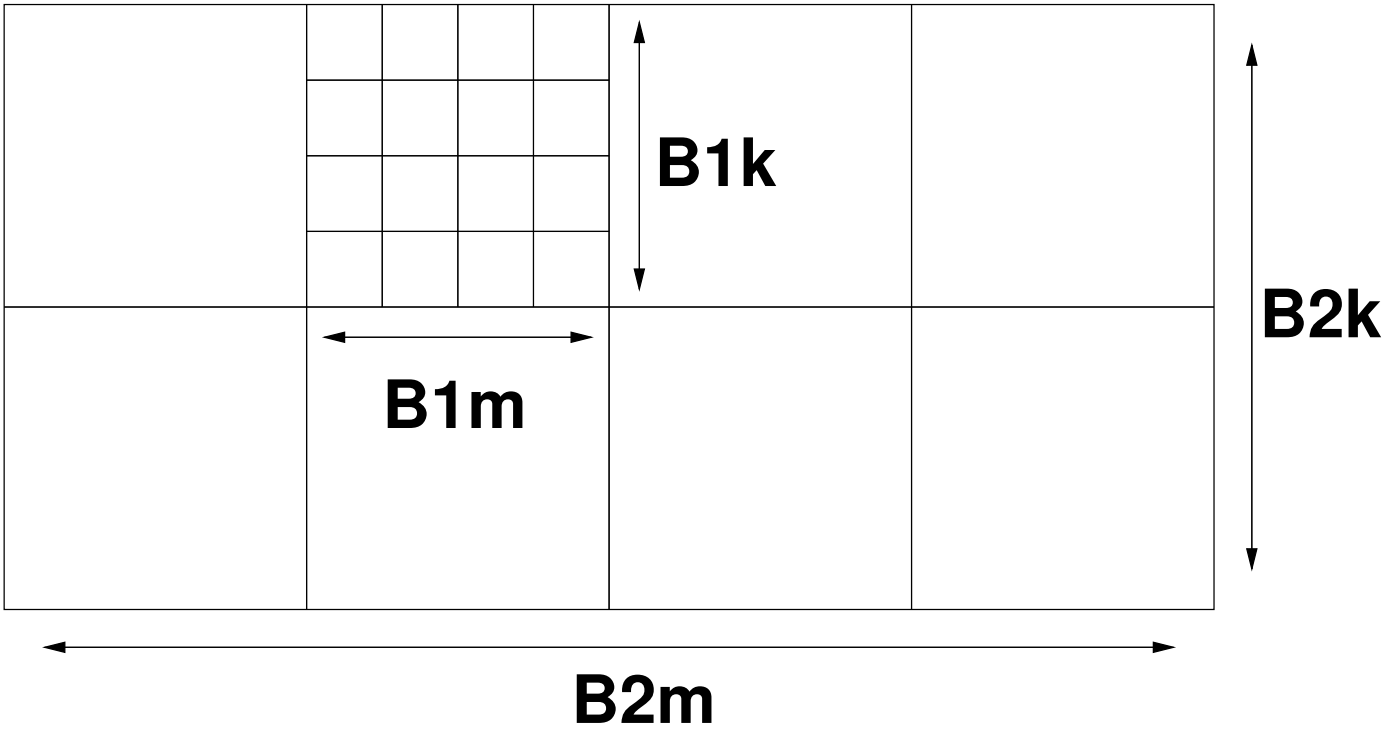
The problem with this nave implementation is the frequent access to global memory.

7.6.2 Caching sub-Blocks in Registers



In the `matmul_reg` implementation, the matrix C is divided into sub-matrices of block size `B1m` x `B1k`. Each block in the result matrix C is handled by exactly one thread. For the calculation of the sub-blocks of matrix C, the values of the corresponding blocks of matrices A and B are stored in registers. This reduces the number of memory accesses to matrix A by a factor `B1m` and to matrix B by a factor `B1k`.

7.6.3 Caching sub-Blocks in Registers and Local Memory



The `matmul_reg_and_localmem` implementation takes the tiling one step further and uses both registers and local memory. Result matrix C is divided into large blocks of size $(B1m \times B2m) \times (B1k \times B2k)$. Each large block is handled by one work-group. This large block is sub-divided into small blocks of size $B1m \times B1k$, one block for every thread in the work-group. Each thread then computes its small block, using registers for the small block calculation, and local memory to fetch the data from within the large block.