OmpSs User Guide

Release

BSC Programming Models

Nov 27, 2018

CONTENTS

Note: There is a PDF version of this document at http://pm.bsc.es/ompss-docs/user-guide/OmpSsUserGuide.pdf

CHAPTER

ONE

INSTALLATION OF OMPSS

1.1 Preparation

You must first choose a directory where you will install OmpSs. In this document this directory will be referred to as `TARGET`. We recommend you to set an environment variable `TARGET` with the desired installation directory. For instance:

```
$ export TARGET=$HOME/ompss
```

1.2 Installation of Extrae (optional)

This is just a quick summary of the installation of Extrae. For a more detailed information check Extrae Homepage

1. Get Extrae from https://tools.bsc.es/downloads (choose *Source tarball* of Extrae tool).
2. Unpack the tarball and enter the just created directory:

   ```
   $ tar xfz extrae-xxx.tar.gz
   $ cd extrae-xxx
   ```

3. Configure it:

   ```
   $ ./configure --prefix=$TARGET
   ```

   Note: Extrae may have a number of dependences. Do not hesitate to check the Extrae User's Manual

4. Build and install:

   ```
   $ make
   $ make install
   ```

Now you can proceed to the *Installation of Nanos++*. Do not forget to pass `--with-extrae=$TARGET` to Nanos++ `configure`.

1.3 Installation of Nanos++

1. First, make sure that you fulfill all the *Nanos++ build requirements*.

2. Get the latest Nanos++ *tarball* (`nanox-version-yyyy-mm-dd.tar.gz`) from https://pm.bsc.es/ompss-downloads Unpack the file and enter the just created directory:

```
$ tar xfz nanox-version-yyyy-mm-dd.tar.gz
$ cd nanox-version
```

3. Run `configure`. There are a number of flags that enable or disable different features of Nanos++. Make sure you pass `--prefix` with the destination directory of your OmpSs installation (in our case `TARGET`).

   Important: If you want instrumentation support, you must add `--with-extrae=dir` to your *configure-flags* (see below), where `dir` is the directory where you installed Extrae (usually `TARGET`).

   Check *Nanos++ configure flags* for more information about *configure-flags*.

```
$ ./configure --prefix=$TARGET configure-flags
```

   Note: You can pass an empty set of *configure-flags*.

   Hint: `configure` prints lots of output, but a small summary of enabled features is printed at the end. You may want to check it to ensure you are correctly passing all the flags

4. Build

```
$ make
```

   This may take some time. You may build in parallel using `make -jN` where `N` is the maximum number of threads you want to use for parallel compilation.

5. Install

```
$ make install
```

   This will install Nanos++ in `TARGET`.

Now you can proceed to *Installation of Mercurium C/C++/Fortran source-to-source compiler*.

1.3.1 Nanos++ build requirements

There are a number of software requirements to successfully build Nanos++ from the source code.

Important: Additional software is needed if you compile from the git repository. This section details the requirements when building from a *tarball* (a *tar.gz* file).

- A supported platform running Linux (i386, x86-64, ARM, PowerPC or IA64).
- GNU C/C++ compiler versions 4.4 or better.

If you want CUDA support:

- CUDA 5.0 or better.

If you want to enable the cluster support in Nanos++ you will need:

- GASNet 1.14.2 or better

1.3.2 Nanos++ configure flags

By default Nanos++ compiles three versions: *performance*, *debug* and *instrumentation*. Which one is used is usually governed by flags to the Mercurium compiler. You can speedup the build of the library by selectively disabling these versions.

You can also enable a fourth version: *instrumentation-debug*. That one is probably of little interest to regular users of Nanos++ as it enables debug and instrumentation at the same time.

--disable-instrumentation Disables generation of instrumentation version

--disable-debug Disables generation of debug version

--disable-performance Disables generation of performance version

--enable-instrumentation-debug Enables generation of instrumentation-debug version

Besides the usual shared memory environment (which is called the *SMP* device), Nanos++ supports several devices. Such devices are automatically enabled if enough support is detected at the host. You can disable them with the following flags.

--disable-gpu-arch Disables CUDA support

--disable-opencl-arch Disables OpenCL support

Nanos++ includes several plugins that are able to use other software packages. You can enable them using the following flags.

--with-cuda=dir Directory of CUDA installation. By default it checks `/usr/local/cuda`. If a suitable installation of CUDA is found, CUDA support is enabled in Nanos++ (unless you disable it)

--with-opencl=dir Directory of OpenCL installation. By default it checks `/usr`. If a suitable installation of OpenCL is found, OpenCL support is enabled in Nanos++ (unless you disable it)

--with-opencl-include=dir If you use –with-opencl=dir, `configure` assumes that dir/`include` contains the headers. Use this flag to override this assumption.

--with-opencl-lib=dir If you use –with-opencl=dir, `configure` assumes that dir/`lib` contains the libraries. Use this flag to override this assumption.

--with-extrae=dir Directory of Extrae installation. **This is mandatory if you want instrumentation**. Make sure you have already installed Extrae first. See *Installation of Extrae (optional)*

--with-mpitrace=dir This is a deprecated name for `--with-extrae`

--with-nextsim=dir Directory of NextSim installation

--with-ayudame=dir Directory of Ayudame installation

--with-hwloc=dir Directory of Portable Hardware Locality (hwloc) installation. This is highly recommended for NUMA setups

--with-chapel=dir Directory of Chapel installation

--with-mcc=dir Directory of Mercurium compiler. *This is only for testing Nanos++ itself and only useful to Nanos++ developers*.

1.4 Installation of Mercurium C/C++/Fortran source-to-source compiler

You can find the build requirements, the configuration flags and the instructions to build Mercurium in the following link: https://github.com/bsc-pm/mcxx/blob/master/README.md

Once you complete all the steps listed in the link above you should be ready to *Compile OmpSs programs*.

CHAPTER

TWO

COMPILE OMPSS PROGRAMS

After the *Installation of OmpSs* you can now start compiling OmpSs programs. For that end you have to use the Mercurium compiler (which you already installed as described in *Installation of Mercurium C/C++/Fortran source-to-source compiler*).

See sections *Compiling an OmpSs program with CUDA tasks* and *Compiling an OmpSs program with OpenCL tasks* for more information. .. index:

```
double: Mercurium; drivers
```

2.1 Drivers

The list of available drivers can be found here: https://github.com/bsc-pm/mcxx/blob/master/doc/md_pages/profiles.md

2.1.1 Common flags accepted by Mercurium drivers

Usual flags like `-O`, `-O1`, `-O2`, `-O3`, `-D`, `-c`, `-o`, ... are recognized by Mercurium.

Almost every Mercurium-specific flag is of the form `--xxx`.

Mercurium drivers are deliberately compatible with `gcc`. This means that flags of the form `-fXXX`, `-mXXX` and `-Wxxx` are accepted and passed onto the backend compiler without interpretation by Mercurium drivers.

Warning: In GCC a flag of the form `-fXXX` is equivalent to a flag of the form `--XXX`. This is **not** the case in Mercurium.

2.1.2 Help

You can get a summary of all the flags accepted by Mercurium using `--help` with any of the drivers:

```
$ mcc --help
Usage: mcc options file [file..]
Options:
  -h, --help               Shows this help and quits
  --version                Shows version and quits
  --v, --verbose           Runs verbosely, displaying the programs
                           invoked by the compiler
...
```

2.1.3 Passing vendor-specific flags

While almost every `gcc` of the form `-fXXX` or `-mXXX` can be passed directly to a Mercurium driver, some other vendor-specific flags may not be well known or be misunderstood by Mercurium. When this happens, Mercurium has a generic way to pass parameters to the backend compiler and linker.

`--Wn,<comma-separated-list-of-flags>` Passes comma-separated flags to the native compiler. These flags are used when Mercurium invokes the backend compiler to generate the object file (`.o`)

`--Wl,<comma-separated-list-of-flags>` Passes comma-separated-flags to the linker. These flags are used when Mercurium invokes the linker

`--Wp,<comma-separated-list-of-flags>` Passes comma-separated flags to the C/Fortran preprocessor. These flags are used when Mercurium invokes the preprocessor on a C or Fortran file.

These flags can be combined. Flags `--Wp,a --Wp,b` are equivalent to `--Wp,a,b`. Flag `--Wnp,a` is equivalent to `--Wn,a --Wp,a`

Important: Do not confuse `--Wl` and `--Wp` with the `gcc` similar flags `-Wl` and `-Wp` (note that `gcc` ones have a single `-`). The latter can be used with the former, as in `--Wl,-Wl,muldefs`. That said, Mercurium supports `-Wl` and `-Wp` directly, so `-Wl,muldefs` should be enough.

2.2 Compiling an OmpSs program for shared-memory

For OmpSs programs that run in SMP or NUMA systems, you do not have to do anything. Just pick one of the drivers above.

2.2.1 Your first OmpSs program

Following is a very simple OmpSs program in C:

```
/* test.c */
#include <stdio.h>

int main(int argc, char *argv[])
{
      int x = argc;
#pragma omp task inout(x)
      {
          x++;
      }
#pragma omp task in(x)
      {
          printf("argc + 1 == %d\n", x);
      }
#pragma omp taskwait
      return 0;
}
```

Compile it using *mcc*:

```
$ mcc -o test --ompss test.c
Nanos++ prerun
Nanos++ phase
```

Important: Do not forget the flag `--ompss` otherwise the OmpSs directives will be ignored.

And run it:

```
$ ./test
argc + 1 == 2
```

2.3 Compiling an OmpSs program with CUDA tasks

To compile an OmpSs + CUDA program you must use one of the OmpSs/OmpSs-2 profiles (see *Drivers*) and the `--cuda` flag. Note that the Nanos++ installation that was provided to Mercurium must have been configured with CUDA (see *Nanos++ configure flags*).

The usual structure of an OmpSs program with CUDA tasks involves a set of C/C++ files (usually `.c` or `.cpp`) and a set of CUDA files (`.cu`). You can use one of our profiles to compile the CUDA files, but the driver will simply call the NVIDIA Cuda Compiler `nvcc`. We discourage mixing C/C++ code and CUDA code in the same file when using Mercurium. While this was supported in the past, such approach is regarded as deprecated. The recommended approach is to move the CUDA code to a `.cu` file.

Note: When calling kernels from C, make sure the `.cu` file contains an `extern "C"`, otherwise C++ mangling in the name of the kernels will cause the link step to fail.

Warning: Fortran support for CUDA is experimental. Currently you cannot pass a `.cu` file to Mercurium: compile them separatedly using `nvcc`.

Consider the following CUDA kernels:

```
/* cuda-kernels.cu */

extern "C" {  // We specify extern "C" because we will call them from a C code

  __global__ void init(int n, int *x)
  {
      int i = blockIdx.x * blockDim.x + threadIdx.x;
      if (i >= n)
          return;
      x[i] = i;
  }

  __global__ void increment(int n, int *x)
  {
      int i = blockIdx.x * blockDim.x + threadIdx.x;
      if (i >= n)
          return;
      x[i]++;
  }

} /* extern "C" */
```

You can use the following OmpSs program in C to call them:

```
/* cuda-test.c */

#include <stdio.h>
#include <assert.h>

#pragma omp target device(cuda) copy_deps ndrange(1, n, 1)
#pragma omp task out(x[0 : n-1])
__global__ void init(int n, int *x);

#pragma omp target device(cuda) copy_deps ndrange(1, n, 1)
#pragma omp task inout(x[0 : n-1])
__global__ void increment(int n, int *x);

int main(int argc, char *argv[])
{
    int x[10];

    init(10, x);
    increment(10, x);

#pragma omp taskwait

    int i;
    for (i = 0; i < 10; i++)
    {
        fprintf(stderr, "x[%d] == %d\n", i, x[i]);
        assert(x[i] == (i+1));
    }

    return 0;
}
```

Compile these files using `mcc`:

```
$ mcc -o cuda-test cuda-test.c  cuda-kernels.cu  --ompss --cuda
cuda-test.c:5: note: adding task function 'init'
cuda-test.c:9: note: adding task function 'increment'
cuda-test.c:16: note: call to task function 'init'
cuda-test.c:17: note: call to task function 'increment'
cudacc_cuda-test.cu: info: enabling experimental CUDA support
```

Note: You can compile this example separatedly, as usual: using `-c` and linking the intermediate `.o` files.

Now you can run it:

```
$ ./cuda-test
x[0] == 1
x[1] == 2
x[2] == 3
x[3] == 4
x[4] == 5
x[5] == 6
x[6] == 7
x[7] == 8
x[8] == 9
x[9] == 10
```

```
MSG: [12] GPU 0 TRANSFER STATISTICS
MSG: [12]     Total input transfers: 0 B
MSG: [12]     Total output transfers: 0 B
MSG: [12]     Total device transfers: 0 B
MSG: [13] GPU 1 TRANSFER STATISTICS
MSG: [13]     Total input transfers: 0 B
MSG: [13]     Total output transfers: 40 B
MSG: [13]     Total device transfers: 0 B
```

Note that due to the caching mechanism of Nanos++ only 40 bytes (10 integers) have been transferred from the GPU to the host. See *Running OmpSs Programs* for detailed information about options affecting the execution of OmpSs programs.

2.4 Compiling an OmpSs program with OpenCL tasks

To compile an OmpSs + OpenCL program you must use one of the OmpSs/OmpSs-2 profiles (see *Drivers*) and the `--opencl` flag. Note that the Nanos++ installation that was provided to Mercurium must have been configured with OpenCL (see *Nanos++ configure flags*).

Consider this OpenCL file with kernels similar to the CUDA example above:

```
/* ocl_kernels.cl */
__kernel void init(int n, int  __global * x)
{
    int i = get_global_id(0);
    if (i >= n)
        return;
    x[i] = i;
}

__kernel void increment(int n, int  __global * x)
{
    int i = get_global_id(0);
    if (i >= n)
        return;
    x[i]++;
}
```

We can call these kernels from the OmpSs program:

```
#include <stdio.h>
#include <assert.h>

#pragma omp target device(opencl) copy_deps ndrange(1, n, 8) file(ocl_kernels.cl)
#pragma omp task out(x[0 : n-1])
void init(int n, int *x);

#pragma omp target device(opencl) copy_deps ndrange(1, n, 8) file(ocl_kernels.cl)
#pragma omp task inout(x[0 : n-1])
void increment(int n, int *x);

int main(int argc, char *argv[])
{
    int x[10];

    init(10, x);
```

```
    increment(10, x);

#pragma omp taskwait

    int i;
    for (i = 0; i < 10; i++)
    {
        fprintf(stderr, "x[%d] == %d\n", i, x[i]);
        assert(x[i] == (i+1));
    }

    return 0;
}
```

Note: OpenCL files must be passed to the driver if they appear in `file` directives. This also applies to separate compilation using `-c`.

Compile the OmpSs program using `mcc`:

```
$ mcc  -o ocl_test ocl_test.c ocl_kernels.cl --ompss
ocl_test.c:5: note: adding task function 'init'
ocl_test.c:9: note: adding task function 'increment'
ocl_test.c:18: note: call to task function 'init'
ocl_test.c:19: note: call to task function 'increment'
```

Now you can execute it:

```
$ ./ocl_test
x[0] == 1
x[1] == 2
x[2] == 3
x[3] == 4
x[4] == 5
x[5] == 6
x[6] == 7
x[7] == 8
x[8] == 9
x[9] == 10
MSG: [12] OpenCL dev0 TRANSFER STATISTICS
MSG: [12]     Total input transfers: 0 B
MSG: [12]     Total output transfers: 0 B
MSG: [12]     Total dev2dev(in) transfers: 0 B
MSG: [13] OpenCL dev1 TRANSFER STATISTICS
MSG: [13]     Total input transfers: 0 B
MSG: [13]     Total output transfers: 40 B
MSG: [13]     Total dev2dev(in) transfers: 0 B
```

Again, due to the caching mechanism of Nanos++, only 40 bytes (10 integers) have been transferred from the OpenCL device (in the example above it is a GPU) to the host.

2.5 Problems during compilation

While we put big efforts to make a reasonably robust compiler, you may encounter a bug or problem with Mercurium.

There are several errors of different nature that you may run into.

- Mercurium ends abnormally with an internal error telling you to open a ticket.
- Mercurium does not crash but gives an error on your input code and compilation stops, as if your code were not valid.
- Mercurium does not crash, but gives an error involving an `internal-source`.
- Mercurium generates wrong code and native compilation fails on an intermediate file.
- Mercurium forgets something in the generated code and linking fails.

2.5.1 How can you help us to solve the problem quicker?

In order for us to fix your problem we need the *preprocessed* file.

If your program is C/C++ we need you to do:

1. Figure out the compilation command of the file that fails to compile. Make sure you can replicate the problem using that compilation command alone.
2. If your compilation command includes `-c`, replace it by `-E`. If it does not include `-c` simply add `-E`.
3. If your compilation command includes `-o file` (or `-o file.o`) replace it by `-o file.ii`. If it does not include `-o`, simply add `-o file.ii`.
4. Now run the compiler with this modified compilation command. It should have generated a `file.ii`.
5. These files are usually very large. Please compress them with `gzip` (or `bzip2` or any similar tool).

Send us an email to pm-tools@bsc.es with the error message you are experiencing and the (compressed) preprocessed file attached.

If your program is Fortran just the input file may be enough, but you may have to add all the `INCLUDE`d files and modules.

CHAPTER

THREE

RUNNING OMPSS PROGRAMS

Nanos++ is an extensible runtime library designed to serve as a runtime support in parallel environments. It is mainly used to support OmpSs (an extension to the OpenMP programming model) developed at BSC. Nanos++ also has modules to support OpenMP and Chapel.

With Nanox++ also comes a small application offering a quick help about the runtime common options. It lists the available configuration options and modules. For each module also lists all configuration flags available. This application is installed in the NANOX_INSTALL_DIR/bin directory and it is called 'nanox'. It accepts several command line options:

--help Shows command line and environment variables available.

--list-modules Shows the modules that are available with the current installation of Nanos++.

--version Shows the installed version of Nanos++ and the configure command line used to install it.

Some parts of the runtime come in the form of modules that are loaded on demand when the library is initialized. Each module can have its own particular options which can be set using also command-line parameters. In order to see the list of available modules and options use the nanox tool. As in the case of the runtime options you can also choose among all available plugins using a environment variable (or the equivalent command-line option like using NX_ARGS).

3.1 Runtime Options

Nanos++ has different configuration options which can be set through environment variables. These runtime parameters can be also passed as a command-line option like through the environment variable NX_ARGS. For example, these two invocations of a Nanos++ application are equivalent:

```
$ export NX_FANCY_OPTION=whatever
$./myProgram

$ export NX_ARGS='--fancy-option=whatever'
$ ./myProgram
```

3.1.1 General runtime options

Nanos++ have several common options which can be specified while running OmpSs's programs.

--smp-cpus=<n> Set number of requested CPUs.

--smp-workers=<n> Set number of SMP worker threads.

--cpus-per-socket=<n> Set number of CPUs per socket.

--num-sockets=<n> Set number of sockets available.

--hwloc-topology=<xml-file> Overrides hwloc's topology discovery and uses the one provided by *xml-file*.

--stack-size=<n> Defines default stack size for all devices.

--binding-start=<cpu-id> Set initial CPU for binding (binding required).

--binding-stride=<n> Set binding stride (binding required).

--disable-binding, --no-disable-binding Disables/enables thread binding (enabled by default).

--verbose, --no-verbose Activate verbose mode (requires Nanos++ debug version).

--disable-synchronized-start, --no-disable-synchronized-start Disables synchronized start (enabled by default).

--architecture=<arch> Defines default architecture. Where *arch* can be one of the following options: `smp`, `gpu`, `opencl` *or* `cluster`.

--disable-caches, --no-disable-caches Disables/enables the use of caches (enabled by default).

--cache-policy=<cache-policy> Set default cache policy. Where *cache-policy* can be one of the following options: `wt` (write through) *or* `wb` (write back).

--summary, --no-summary Enables/disables runtime summary mode, printing statistics at start/end phases (disabled by default)

3.1.2 CUDA specific options

This is the list of CUDA specific options:

NX_ARGS option	Environment variable	Description
–disable-cuda	NX_DISABLECUDA	Enable or disable the use of GPUs with CUDA
–gpus	NX_GPUS	Defines the maximum number of GPUs to use
–gpu-warmup	NX_GPUWARMUP	Enable or disable warming up the GPU before running user's code
–gpu-prefetch	NX_GPUPREFETCH	Set whether data prefetching must be activated or not
–gpu-overlap	NX_GPUOVERLAP	Set whether GPU computation should be overlapped with all data transfers, whenever possible, or not
–gpu-overlap-inputs	NX_GPUOVERLAP_INPUTS	Set whether GPU computation should be overlapped with host –> device data transfers, whenever possible, or not
–gpu-overlap-outputs	NX_GPUOVERLAP_OUTPUTS	Set whether GPU computation should be overlapped with device –> host data transfers, whenever possible, or not
–gpu-max-memory	NX_GPUMAXMEM	Defines the maximum amount of GPU memory (in bytes) to use for each GPU. If this number is below 100, the amount of memory is taken as a percentage of the total device memory
–gpu-cache-policy	NX_GPU_CACHE_POLICY	Defines the cache policy for GPU architectures: write-through / write-back / do not use cache
–gpu-cublas-init	NX_GPUCUBLASINIT	Enable or disable CUBLAS initialization

Following table summarizes valid and default values:

NX_ARGS option	Environment variable	Values	Default
–disable-cuda	NX_DISABLECUDA	yes / no	Enabled
–gpus	NX_GPUS	integer	All GPUs
–gpu-warmup	NX_GPUWARMUP	yes / no	Enabled
–gpu-prefetch	NX_GPUPREFETCH	yes / no	Disabled
–gpu-overlap	NX_GPUOVERLAP	yes / no	Disabled
–gpu-overlap-inputs	NX_GPUOVERLAP_INPUTS	yes / no	Disabled
–gpu-overlap-outputs	NX_GPUOVERLAP_OUTPUTS	yes / no	Disabled
–gpu-max-memory	NX_GPUMAXMEM	positive integer	No limit
–gpu-cache-policy	NX_GPU_CACHE_POLICY	wt/wb/nocache	wb
–gpu-cublas-init	NX_GPUCUBLASINIT	yes / no	Disabled

3.1.3 Cluster specific options

TBD

3.2 Running on Specific Architectures

The specific information related to the different architectures supported by OmpSs is explained here.

3.2.1 CUDA Architecture

Performance guidelines and troubleshooting advices related to OmpSs applications using CUDA are described here.

Tuning OmpSs Applications Using GPUs for Performance

In general, best performance results are achieved when prefetch and overlap options are enabled. Usually, a *write-back* cache policy also enhances performance, unless the application needs lots of data communication between host and GPU devices.

Other Nanox options for performance tuning can be found at *My application does not run as fast as I think it could*.

Running with cuBLAS (v2)

Since CUDA 4, the first parameter of any cuBLAS function is of type `cublasHandle_t`. In the case of OmpSs applications, this handle needs to be managed by Nanox, so `--gpu-cublas-init` runtime option must be enabled.

From application's source code, the handle can be obtained by calling `cublasHandle_t nanos_get_cublas_handle()` API function. The value returned by this function is the cuBLAS handle and it should be used in cuBLAS functions. Nevertheless, the cuBLAS handle is only valid inside the task context and should not be stored in a variable, as it may change over application's execution. The handle should be obtained through the API function inside all tasks calling cuBLAS.

Example:

```
#pragma omp target device (cuda) copy_deps
#pragma omp task inout([NB*NB]C) in([NB*NB]A, [NB*NB]B)
void matmul_tile (double* A, double* B, double* C, unsigned int NB)
{
```

```
    REAL alpha = 1.0;

    // handle is only valid inside this task context
    cublasHandle_t handle = nanos_get_cublas_handle();

    cublas_gemm(handle, CUBLAS_OP_T, CUBLAS_OP_T, NB, NB, NB, &alpha, A, NB, B, NB, &
↪alpha, C, NB);
}
```

GPU Troubleshooting

If the application has an unexpected behavior (either reports incorrect results or crashes), try using Nanox debug version (the application must be recompiled with flag `--debug`). Nanox debug version makes additional checks, so this version may trigger the cause of the error.

How to Understand GPU's Error Messages

When Nanox encounters an error, it aborts the execution throwing a **FATAL ERROR** message. In the case where the error is related to CUDA, the structure of this message is the following:

FATAL ERROR: [*#thread* **]** *what Nanox was trying to do when the error was detected* **:** *error reported by CUDA*

Note that the true error is the one reported by CUDA; not the one reported by the runtime. The runtime just gives information about the operation it was performing, but this does not mean that this operation caused the error. For example, there can be an error launching a GPU kernel, but (unless this is checked by the user application) the runtime will detect the error at the next call to CUDA runtime that will probably be a memory copy.

Recommended Steps When an Error Is Found

1. Compile the application with `--debug` and `-ggdb3` flags, optionally use `-O0` flag, too.
2. Run normally and check if any **FATAL ERROR** is triggered. This will give a hint of what is causing the error. For getting more information, proceed to the following step.
3. Run the application inside gdb or cuda-gdb to find out the exact point where the application crashes and explore the backtrace, values of variables, introduce breakpoints for further analysis or any other tool you consider useful. You can also check *My application crashes. What can I do?* for further information.
4. If you think that the error is a bug in OmpSs, please report it.

Common Errors from GPUs

Here is a list of common errors found when using GPUs and how they can be solved.

- **No cuda capable device detected:** This means that the runtime is not able to find any GPU in the system. This probably means that GPUs are not detected by CUDA. You can check your CUDA installation or GPU drivers by trying to run any sample application from CUDA SDK, like *deviceQuery*, and see if GPUs are properly detected by the system or not.
- **All cuda capable devices are busy or unavailable:** Someone else is using the GPU devices, so Nanox cannot access them. You can check if there are other instances of Nanox running on the machine, or if there is any other application running that may be using GPUs. You will have to wait till this application finishes or frees some GPU devices. The *deviceQuery* example from CUDA SDK can be used to check GPU's

memory state as it reports the available amount of memory for each GPU. If this number is low, it is most likely that another application is using that GPU device. This error is reported in CUDA 3 or lower. For CUDA 4 or higher an *Out of memory* error is reported.

- **Out of memory:** The application or the runtime are trying to allocate memory, but the operation failed because GPU's main memory is full. Nanox, by default allocates around 95% of each GPU device's memory (unless this value is modified using `--gpu-max-memory` runtime option). If the application needs additional device memory, try to tune the amount of memory that Nanox is allowed to use with `--gpu-max-memory` runtime option. This error is also displayed when there is someone else using the GPU device and Nanox cannot allocate device's memory. Refer to *All cuda capable devices are busy or unavailable* error to solve this problem.
- **Unspecified launch failure:** This usually means that a segmentation fault occurred on the device while running a kernel (an invalid or illegal memory position has been accessed by one or more GPU kernel threads). Please, check OmpSs source code directives related to dependencies and copies, compiler warnings and your kernel code.

3.2.2 Offload Architecture

Performance guidelines and troubleshooting advices related to OmpSs applications using offloaded codes are described here.

Offloading

In order to offload to remote nodes, the allocation of the nodes must be performed by the application by using provided API Calls.

From application's source code, the allocation can be done by calling `deep_booster_alloc(MPI_Comm spawners, int number_of_hosts, int process_per_host, MPI_Comm *intercomm)` API function. This function will create number_of_hosts*process_per_host remote workers and create a communicator with the offloaded processes in `intercomm`.

The masters who will spawn each process are the ones who are part of the communicator "spawners". This call is a collective operation and should be called by every process who is part of this communicator. Two values have been tested:

- MPI_COMM_WORLD: All the masters in the communicator will spawn number_of_hosts*process_per_host remote workers. All these workers can communicate using MPI (they are inside the same MPI_COMM_WORLD).
- MPI_COMM_SELF: Each master which calls the spawn will spawn number_of_hosts*process_per_host remote workers, only visible for himself. Workers spawned by the same node will be able to communicate using MPI, but they won't be able to communicate with the workers created by a different master.

This routine has different API calls (both work in C and also in Fortran (MPI_Comm/int/MPI_Comm* = INTEGER in Fortran, int* = INTEGER ARRAY in Fortran)):

- **deep_booster_alloc** *(MPI_Comm spawners, int number_of_hosts, int process_per_host, MPI_Comm *intercomm)*: Allocates process_per_host processes in each host. If there are not enough number of hosts, the call will fail and returned intercomm will be MPI_COMM_NULL.
- **deep_booster_alloc_list** *(MPI_Comm spawners, int pph_list_length, int* pph_list, MPI_Comm *intercomm)*: Provides a list with the number of processes which will be spawned in each node. For example, the following list {0,2,4,0,1,3} will spawn 10 processes split as indicated between hosts 1,2,4,5, skipping host 0 and 3. If there are not enough number of hosts, the call will fail and returned intercomm will be MPI_COMM_NULL.

- **deep_booster_alloc_nonstrict** *(MPI_Comm spawners, int number_of_hosts, int process_per_host, MPI_Comm *intercomm, int* provided)*: Allocates process_per_host processes in each host. If there are not enough number of hosts, the call will allocate as many as available and return the number of processes allocated (available_hosts*process_per_host) in "provided".
- **deep_booster_alloc_list_nonstrict** *(MPI_Comm spawners, int pph_list_length, int* pph_list, MPI_Comm *intercomm, int* provided)*: Provides a list with the number of processes which will be spawned in each node. For example, the following list {0,2,4,0,1,3} will spawn 10 processes split indicated between hosts 1,2,4,5, skipping host 0 and 3. If there are not enough number of hosts, the call will allocate as many as available and return the number of the number of processes allocated available in "provided".

Deallocation of the nodes will be performed automatically by the runtime at the end of the execution, however its strongly suggested to free them explicitly (it can be done at any time of the execution). This can be done by using the API function, `deep_booster_free(MPI_Comm *intercomm)`.

Communication between master and spawned processes (when executing offloaded tasks) in user code works, but it's not recommended.

Example in C:

```
int main (int argc, char** argv)
{
        int my_rank;
        int mpi_size;
        nanos_mpi_init(&argc,&argv);
        MPI_Comm_size(MPI_COMM_WORLD,&mpi_size);
        MPI_Comm_rank(MPI_COMM_WORLD,&my_rank);
        MPI_Comm boosters;
        int num_arr[2000];
        init_arr(num_arr,0,1000);
        //Spawn as one worker per master
        deep_booster_alloc(MPI_COMM_WORLD, mpi_size  , 1  , &boosters);
        if (boosters==MPI_COMM_NULL) exit(-1);

        //Each master will offload to the same rank than him, but in the workers
        #pragma omp task inout(num_arr[0;2000]) onto(boosters,my_rank)
        {
                        //Workers SUM all their num_arr[0,1000] into num_arr[1000,
↪2000]
                        MPI_Allreduce(&num_arr[0],&num_arr[1000],1000,MPI_INT,MPI_SUM,
↪MPI_COMM_WORLD);
        }
        #pragma omp taskwait

        print_arr(num_arr,1000,2000);

        deep_booster_free(&boosters);
        nanos_mpi_finalize();
        return 0;
}
```

Example in Fortran:

```
    PROGRAM MAIN
INCLUDE "mpif.h"
                    EXTERNAL :: SLEEP
                    IMPLICIT NONE
                    INTEGER :: BOOSTERS
                    INTEGER :: RANK, IERROR,MPISIZE, PROVIDED
```

```
                    INTEGER, DIMENSION(2000) :: inputs
                    INTEGER, DIMENSION(2000) :: outputs

                    !Initialize MPI with THREAD_MULTIPLE (required for offload)
                    call mpi_init_thread(MPI_THREAD_MULTIPLE,provided,ierror)
                    CALL MPI_COMM_RANK(MPI_COMM_WORLD, RANK, ierror)
                    CALL MPI_COMM_SIZE(MPI_COMM_WORLD, MPISIZE, ierror)

                    !!Spawn as one worker per master
                    CALL deep_booster_alloc(MPI_COMM_WORLD, MPISIZE, 1, BOOSTERS)
                    IF (boosters==MPI_COMM_NULL) THEN
                                    CALL exit(-1)
                    END IF
                    INPUTS=5
                    OUTPUTS=999

                    !!Each master will offload to the same rank than him, but in the␣
↪workers
                    !$OMP TASK IN(INPUTS) OUT(OUTPUTS) ONTO(BOOSTERS,RANK)
                                    CALL MPI_ALLREDUCE(INPUTS,OUTPUTS,2000,MPI_INT,
↪MPI_SUM,MPI_COMM_WORLD, IERROR);
                    !$OMP END TASK
                    !$OMP TASKWAIT
                    PRINT *, outputs(2)

                    CALL DEEP_BOOSTER_FREE(BOOSTERS)
                    CALL MPI_Finalize(ierror)
    END PROGRAM
```

Compiling OmpSs offload applications

OmpSs offload Applications can be compiled by using mercurium compilers, “mpimcc/mpimcxx” and “mpimfc” with –ompss flag (E.G.: mpimcc –ompss test.c), which are a wrapper for the current MPI implementation available in users PATH.

User MUST and provide use a multithread MPI implementation when offloading (otherwise nanox will give a explanatory warning and crash at start or hang) and link all the libraries of his program with the same one, -mt_mpi is automatically set for Intel MPI by our compiler.

If no OFFL_CC/OFFL_CXX/OFFL_FC environment variables are defined. OmpSs will use Intel (mpiicc, mpiicpc or mpiifort) compilers by default, and if they are not available, it fall back to OMPI/GNU ones (mpicc,mpicxx and mpicpc).

WARNING: When setting OFFL_CC/OFFL_FC, make sure that OFFL_CXX points to the same MPI implementation than OFFL_CC and OFFL_FC (as one C++ MPI-Dependant part of nanox is automatically compiled and linked with your application)

Naming Convention

MPI Offload executables have to follow a naming convention which identifies the executable and the architecture. Both executables have to be generated from the same source code. This convention is XX.YY, where XX is your executable name and must be the same for every architecture and YY is the architecture name (as given by ‘uname -p’ or ‘uname -m’ (both upper or lower case names are supported)). For x86_64 and k1om (MIC), two default alias are providen, intel64/INTEL64 and mic/MIC.

Example dual-architecture (MIC+Intel64) compilation:

```
mpimcc --ompss test.c -o test.intel64
mpimcc --ompss --mmic test.c -o test.mic
```

Offload hosts

In machines where there is no job manager integration with our offload, the hosts where to offload each process have to be specified manually in a host-file. The path of this file can be specified to Nanox in the environment variable `NX_OFFL_HOSTFILE`, like `NX_OFFL_HOSTFILE=./offload_hosts`

The syntax of this file is the following: hostA:n<env_var1,env_var2... hostB:n<env_var4,env_var1...

Example hostfile:

```
knights4-mic0:30<NX_SMP_WORKERS=1,NX_BINDING_STRIDE=4
knights3-mic0
knights2-mic0
knights1-mic0:3
knights0-mic0
```

Debugging your application

If you want to see where your offload workers are executed and more information about them, use the environment variable `NX_OFFL_DEBUG=<level>` level is a number in the range 1-5, higher numbers will show more debug information.

Sometimes offloading environment can be hard to debug and prints are not powerful enough, below you can find some techniques which will allow you to do so. You should be able to debug it with any MPI debugging tool if you point it to the right offload process ID. In order to do this you have two options.

1. Obtaining backtrace:

```
1- Compile your application with -k and -g and with --debug flag
2- Set "ulimit -c unlimited", after doing so, when the application crashes it␣
↪will dump the core.
3- Open it with "gdb ./exec corefile" and you will be able to see
some information about the crash (for example: see backtrace with "bt" command).
```

2. Execution time debugging:

```
1- Compile your application with -k and -g and with --debug flag
2- Add a sleep(X) at the start of the offload section which gives you enough time␣
↪to do the next steps:
3- Execute your application and make sure it is able to execute the offload␣
↪section/sleep.
4- ssh to the remote machine or execute everything in localhost (for debugging␣
↪purposes).
5- Look for the allocated/offloaded processes. (NX_OFFL_DEBUG=3 will help you to␣
↪identify their hostname and PID).
6- Attach to one of the processes with "gdb -pid YOURPID" (use your preferred␣
↪debugger)
        -Your offload tasks will be executed by the main thread of the␣
↪application.
7- You are now debugging the offload process/thread, you can place breakpoints␣
↪and then write "continue" in gdb.
```

Offloading variables

When offloading, a few rules are used in order to offload variables (also arrays/data-regions):

- *Local variables**: They will be copied as in regular OmpSs/OMP tasks.
- *Non-const Global variables**: They will be copied to and from the offload worker (visible inside the task body and also in external functions).
- *Const global variables**: As they are non-writable, they can only be in/firstprivate (copy of host value will be visible inside task body, external functions will see initial value).
- *Global variables which are never specified in copy/firstprivate* : They will not be modified by OmpSs, so users may use them to store "node-private" values under their responsibility.

Global variables are C/C++ global variables and Fortran Module/Common variables.
Task region refers to the code between the task brackets/definition, external functions are functions which are called from inside task region.

*OmpSs offload does not provide any guarantee about the value or the allocated space for these variables after the task which copies/uses them has finished.

Obtaining traces of offload applications

In order to trace an offload application users must follow these steps:

- Compile with –instrument. An installation using Nanox instrumentation version (–with-extrae) is required.
- Set "export EXTRAE_HOME=/my/extrae/path" and "export EXTRAE_HOME_MIC=/my/extrae/path/mic"
- Optional: if you want to use a configuration file, export EXTRAE_CONFIG_FILE=./extrae.xml. Do not enable merging of traces inside the xml file, as is not supported. EXTRAE_DIR environment variable has to be set if this file points extrae intermediate files to a different folder than current directory ($PWD).
- Call your program with "offload_instrumentation.sh mpirun . . . " instead of "mpirun . . . "

"offload_instrumentation.sh" script can be found in $NANOX_HOME/bin in case its not available through the PATH. It will setup the environment and then call the real application.

Three configuration environment variables are available for this script:

- *EXTRAE_DIR*: Directory where intermediate files will be generated (if specified with .xml file, this variable has to point to the same folder). [Default value: $PWD]
- *AUTO_MERGE_EXTRAE_TRACES*: If enabled, after the application finishes, the script will check the directory pointed by *EXTRAE_DIR* and merge the intermediate files into a final trace (.prv) [Default value: YES]. If disabled, users can merge these files manually by calling mpi2prv with the right parameters or by executing the script with no arguments (it will merge intermediate files located in *EXTRAE_DIR*).
- *CLEAR_EXTRAE_TMP_FILES*: If enabled, intermediate files will be removed after finishing the execution of the script [Default value: NO].

Offload Troubleshooting

- If you are offloading tasks which use MPI communications on the workers side and they hang, make sure that you launched as many tasks as nodes in the communicator (so every rank of the remote communicator is

executing one of the tasks), otherwise all the other tasks will be waiting for that one if they perform collective operations.

- If you are trying to offload to Intel MICs, make sure the variable I_MPI_MIC is set to yes/1 (export I_MPI_MIC=1).
- If your application behaves differently by simply compiling with OmpSs offload, take into account that OmpSs offload initializes MPI with MPI_THREAD_MULTIPLE call.
- If your application hangs inside DEEP Booster alloc calls, check that the hosts provided in the hostfile exist.
- Your application (when using MPI_COMM_SELF as first parameter when calling `deep_booster_alloc` from multiple MPI processes) has to be executed in a folder where you have write permissions, this is needed because Intel MPI implementation of MPI_Comm_spawn is not inter-process safe so we have to use a filesystem lock in order to serialize spawns so it does not crash. As a consequence of this problem spawns using MPI_COMM_SELF will be serialized. This behaviour can be disabled by using the variable `NX_OFFL_DISABLE_SPAWN_LOCK=1`, but then MPI Implementation has to support concurrent spawns or you will have to guarantee inside your application code that multiple calls to `DEEP_Booster_alloc` with MPI_COMM_SELF as first parameter are not performed at the same time. This behaviour is disabled when using OpenMPI as everything works as it should in this implementation.
- If you are trying to offload code which gets compiled/configured with CMake, you have to point the compilers to ones provided by the offload, you can do this with `FC=mpimfc CXX=mpimcxx CC=mpimcc cmake ..` If you are using FindMPI, you should disable it (recommended). Setting MPI and non-MPI compilers pointing to our compiler with `FC=mpimfc CXX=mpimcxx CC=mpimcc cmake . -DMPI_C_COMPILER=mpimcc -DMPI_CXX_COMPILER=mpimcxx -DMPI_Fortran_COMPILER=mpimfc` (CXX compiler is mandatory even if you are not using C++) should be enough.

Apart from offload troubleshooting, there are some problems with Intel implementation of MPI_COMM_SPAWN_MULTIPLE on MICs:

- (Fixed when using Intel MPI 5.0.0.016 or higher) If the same host (only on MICs) appears twice, the call to DEEP_BOOSTER_ALLOC/MPI_COMM_SPAWN may hang, in this case try specifying thread binding manually and interleaving those hosts with other machines.

Example crashing hostfile:

```
knights3-mic0
knights3-mic0
knights4-mic0
knights4-mic0
```

Example workaround *fixed* hostfile:

```
knights3-mic0
knights4-mic0
knights3-mic0
knights4-mic0
```

- More than 80 different hosts can't be allocated by using a single file, in this case DEEP_BOOSTER_ALLOC/MPI_COMM_SPAWN will throw a Segmentation Fault. In order to handle this problem, we made a workaround, in offload host-file, instead of specifying one host, you can specify a host-file which will contain hosts.

Example hostfile using sub-groups in files:

```
#Hostfile1 contains 64 hosts, and all of them will have the same offload variables
#(path should be absolute or relative to working directory of your app)
./hostfile1:64<NX_SMP_WORKERS=1,NX_BINDING_STRIDE=4
```

```
knights3-mic0
./hostfile2:128
```

Tuning OmpSs Applications Using Offload for Performance

In general, best performance results is when *write-back* cache policy is used (default).

When offloading, one thread will take care of sending tasks from the host to the device. These threads will only wait for a few operations (for most operations, it just sends orders to the remote nodes) this is enough for most applications, but if you feel that these threads are not enough to offload all your independent tasks. You can increase the number of threads which will handle each alloc by using the variable `NX_OFFL_MAX_WORKERS` as in `NX_OFFL_MAX_WORKERS=12`, which by default has a value of 1.

Formula for N_THREADS_PER_ALLOC = min(NX_OFFL_MAX_WORKERS,number of offload processes/number of master processes).

If you know that you application is not uniform and it will have transfers from one remote node to other remote node while one of them is executing tasks, you may get a small performance improvement by setting `NX_OFFL_CACHE_THREADS=1`, which will start an extra cache thread at worker nodes. If this option is not enabled by the user, it will be enabled automatically only when this behaviour is detected.

Each worker node can receive tasks from different masters but it will execute only one at a time. Try to avoid this kind of behaviour if you application needs to be balanced.

Other Nanox options for performance tuning can be found at *My application does not run as fast as I think it could*.

Setting environment variables (i.e. number of threads) for offload workers

In regular applications, number of threads will be defined by regular variables, for example NX_SMP_WORKERS=X which sets the number of SMP workers/OMP threads in the master. In offload workers, default value will be the same than cores on the machine that the worker is running, if you want to specify them manually, as you may have more than one architecture, number of threads can be different for each node/architecture. In order to do this we provide a few ways to do it, in case of conflicts, the latest one in this list, will be used:

- Specify it individually on the hostfile on each node as an environment variable (NX_SMP_WORKERS or OFFL_NX_SMP_WORKERS). This value overwrites global and architecture variable for that node/group of nodes.
- Specify it globally using the variable "OFFL_VAR", this means that once the remote worker starts, VAL will be defined for every offload worker with the value specified in OFFL_VAR.
- Specify it per-architecture using the variable "YY_VAR", being YY the suffix of the executable for that architecture (explained in *Naming Convention*.) and VAR the variable you want to define in that architecture. For example, for Intel mic architecture, you can use "MIC_NX_SMP_WORKERS=240" in order to set number of threads to 240 for this architecture. This value overwrites global variable for the concrete architecture.

MPI implementation usually exports all those variables automatically, if this is not the case, users must configure it so variables which are needed are exported.

Other Offload variables

Other offload configuration variables can be obtained by using "nanox –help":

Offload Arch:

Environment variables

NX_OFFL_HOSTS = <string> Defines hosts file where secondary process can spawn in DEEP_Booster_Alloc Same format than NX_OFFLHOSTFILE but in a single line and separated with ';' Example: hostZ hostA<env_vars hostB:2<env_vars hostC:3 hostD:4

NX_OFFL_ALIGNTHRESHOLD = <positive integer> Defines minimum size (bytes) which determines if offloaded variables (copy_in/out) will be aligned (default value: 128), arrays with size bigger or equal than this value will be aligned when offloaded

NX_OFFL_CACHE_THREADS = <true/false> Defines if offload processes will have an extra cache thread, this is good for applications which need data from other tasks so they don't have to wait until task in owner node finishes. (Default: False, but if this kind of behaviour is detected, the thread will be created)

NX_OFFL_ALIGNMENT = <positive integer> Defines the alignment (bytes) applied to offloaded variables (copy_in/out) (default value: 4096)

NX_OFFL_HOSTFILE = <string> Defines hosts file where secondary process can spawn in DEEP_Booster_Alloc The format of the file is: One host per line with blank lines and lines beginning with # ignored Multiple processes per host can be specified by specifying the host name as follows: hostA:n Environment variables for the host can be specified separated by comma using hostA:n<env_var1,envar2... or hostA<env_var1,envar2...

NX_OFFL_EXEC = <string> Defines executable path (in child nodes) used in DEEP_Booster_Alloc

NX_OFFL_MAX_WORKERS = <positive integer> Defines the maximum number of worker threads created per alloc (Default: 1)

NX_OFFL_LAUNCHER = <string> Defines launcher script path (in child nodes) used in DEEP_Booster_Alloc

NX_OFFL_ENABLE_ALLOCWIDE = <0/1> Alloc full objects in the cache. This way if you only copy half of the array, the whole array will be allocated. This is good/useful when OmpSs copies share data between offload nodes (Default: False)

NX_OFFL_CONTROLFILE = <string> Defines a shared (GPFS or similar) file which will be used to automatically manage offload hosts (round robin). This means that each alloc will consume hosts, so future allocs do not oversubscribe on the same host.

3.3 Runtime Modules (plug-ins)

As the main purpose of Nanos++ is to be used in research of parallel programming environments it has been designed to be extensible by means of plugins. Runtime plugins can be selected for each execution using the proper runtime option (e.g. NX_SCHEDULE=cilk). These are:

3.3.1 Scheduling policies

The scheduling policy defines how ready tasks are executed. Ready tasks are those whose dependencies have been satisfied therefore its execution can start immediately. The scheduling policy has to decide the order of execution of tasks and the resource where each task will be executed.

When you inject a dependant task into the runtime system actually you are inserting this task into the dependant task graph. Once all dependencies for a given task are fulfilled this task will be inserted into a ready queue.

Before taking into account any scheduling criteria, we must consider four different scheduler modifiers. They are:

- Throttling policy at task creation. Just when we are going to create a new task, the runtime system determines (according to a throttling policy) whether to create it and push it into the scheduler system or just create a

minimal description of the task and execute it right away in the current task context. Throttle mechanism can be configured as a completely independent plugin using the throttle option. See *Throttling policies*.

- Task Priority. When inserting a new task in a priority queue, if the new task has the same priority as another task already in the queue, the new one will be inserted before/after the existent one (according with the FIFO/LIFO behaviour between *tasks with the same priority*). The priority is a number >= 0. Given two tasks with priority A and B, where A > B, the task with priority A will be fetched earlier than the one with B from the queue. Priority is also accumulate from parent to children. When a task T with priority A creates a task Tc that was given priority B by the user, the priority of Tc will be added to that of its parent. In other words, the priority of Tc will be A + B. Smart priority behaviour also propagates the priority to the immediate preceding tasks (when dependant tasks). Ready task priority queues are only available in some schedulers. Check specific scheduler parameters and/or description for more details.
- Immediate successor. Releasing last dependency when exiting a task. When a task is in the dependency graph and its last immediate predecessor finishes execution, we are able to run the blocked tasks immediately instead of adding it to the ready queue. Sometimes immediate successor is not a configurable option. Check specific scheduler description and/or parameters for more details.
- Parent continuation when last child has been executed. When executing nested OpenMP/OmpSs applications (explicit tasks creating more explicit tasks) it may happens that a given task becomes blocked due the execution of a taskwait or taskwait on directive. When the last child has finished its execution the parent will be promptly resumed instead of being push back into the ready queue (which potentially could delay its execution). Parent continuation is only available in wf scheduler. Check specific scheduler description for more details.

When a thread has no work assigned or when it is waiting until a certain condition is met (e.g. all task's children completes in a taskwait), threads are in idle loop and conditional wait respectively. Users can also specify which is the behaviour of these thread duties through the following options (for more info see *Thread Manager*):

--checks=<n>	Set number of checks on a conditional wait loop before spinning on the ready task pool (default = 1).
--spins=<n>	Set number of spin iterations before yielding on idle loop and conditional wait loop (default = 1).
--enable-yield	Enables thread yielding on idle loop and conditional wait loop (disabled by default).
--yields=<n>	Set number of yields before blocking on idle loop and conditional wait loop (default = 1).
--enable-block	Enables thread blocking on idle loop and conditional wait loop (disabled by default).

A complete idle loop cycle starts with spinning many times as spins parameter specifies After spinning, if yield is enabled, it will yield once and it will start again spinning. The whole spin/yield process will be repeated as many time as the yields option specifies and, finally, if thread block is enabled, the thread will block. Yield and block can not happen in the same cycle.

The conditional wait loop will start with checking the wait condition as many times as the checks parameter specifies. After that number of checks it will start a idle loop cycle (described in the previous paragraph).

The scheduler plug-in can be specified by the `NX_SCHEDULE` environment variable or the `--schedule` option included in the `NX_ARGS` environment variable:

```
$export NX_SCHEDULE=<module>
$./myProgram

$export NX_ARGS="--schedule[ \|=]<module> ..."
$./myProgram
```

--schedule=<module> Set the scheduling policy to be used during the execution. The argument *module* can be one of the following options: `bf`, `dbf`, `wf`, `socket`, `affinity`, `affinity-smartpriority` or `versioning`. The most suitable scheduling policy can depend on the application and architecture used.

Nanos++ implements several schedule plug-ins. The following sections will show a description of the available plug-ins with respect to scheduling policies. Some of these modules also incorporate specific options.

Breadth First

Important: This is the default scheduler

- *Configuration string:* `NX_SCHEDULE=bf` *or* `NX_ARGS="--schedule=bf"`
- *Description:* This scheduler policy only implements a single/global ready queue. When creating a task with no dependences (or when a task becomes ready after all its dependences has been fulfilled) it is placed in this ready queue. Ready queue is ordered following a FIFO (First In First Out) algorithm by default, but it can be changed through parameters. Breadth first implements immediate successor mechanism by default (if not in conflict with priority).
- *Parameters:*

  --bf-stack, --no-bf-stack Retrieving from queue's back or queue's front respectively. Using –bf-stack parameters transform ready queue into a LIFO (Last In First Out) structure. Effectively, this means that tasks are inserted in queue's back and retrieved from the same position (queue's back).(default behaviour is –no-bf-stack)

  --bf-use-stack, --no-bf-use-stack This is an alias of the previous parameter.

  --schedule-priority, --no-schedule-priority Use priorities queues.

  --schedule-smart-priority, --no-schedule-smart-priority Use smart priority queues.

Distributed Breadth First

- *Configuration string:* `NX_SCHEDULE=dbf` *or* `NX_ARGS="--schedule=dbf"`
- *Description:* Is a breadth first algorithm implemented with thread local queues instead of having a single/global ready queue. Each thread inserts its created tasks into its own ready queue following a FIFO policy (First In First Out, inserting in queue's front and retrieving from queue's back) algorithm. If thread local ready queue is empty, it tries to execute current task's parent (if it is queued in any other thread ready queue). In the case parent task cannot be eligible for execution it steals from next thread ready queue. Stealing retrieves tasks from queue's front (i.e. the opposite side from local retrieve).

  --schedule-priority, --no-schedule-priority Use priorities queues.

  --schedule-smart-priority, --no-schedule-smart-priority Use smart priority queues.

Work First

- *Configuration string:* `NX_SCHEDULE=wf` *or* `NX_ARGS="--schedule=dbf"`
- *Description:* This scheduler policy implements a local ready queue per thread. Once a task is created it chooses to continue with the new created task, leaving current task (creator) into current thread's ready queue. Default

behaviour is implemented through FIFO access to local queue, LIFO access on steal and steals parent if available which actually is equivalent with the cilk scheduler behaviour.

- *Parameters:*

  --wf-steal-parent, --no-wf-steal-parent If local ready queue is empty, it tries to steal parent task if available (default: wf-steal-parent).

  --wf-local-policy=<string> Defines queue access for local ready queue. Configuration string can be:

  - `FIFO`: First In First Out queue access (default).
  - `LIFO`: Last In First Out queue access.

  --wf-steal-policy=<string> Defines queue access for stealing. Configuration string can be:

  - `FIFO`: First In First Out queue access (default).
  - `LIFO`: Last In First Out queue access.

Socket-aware scheduler

- *Configuration string:* `NX_SCHEDULE=socket` *or* `NX_ARGS="--schedule=socket"`
- *Description:* This scheduler will assign top level tasks (depth 1) to a NUMA node set by the user before task creation while nested tasks will run in the same node as their parent. To do that, the user must call the nanos_current_socket function before executing tasks to set the NUMA node the task will be assigned to. The queues are sorted by priority, and there are as many queues as NUMA nodes specified (see num-sockets parameter). Besides that, changing the binding start and stride is not supported. Work stealing is optional. By default, work stealing of child tasks is enabled. Upon initialisation, the policy will create lists of valid nodes to steal. For each node, the policy will only steal from the closest nodes. Use numactl –hardware to print the distance matrix that the policy will use. For each node, a pointer to the next node to steal is kept; if a node steals a task, it will not affect where other nodes will steal. There is an option to steal from random nodes as well, and to steal top level tasks instead of child tasks.
- *Parameters:* There are important parameters as the number of sockets and the number of cores per socket. If Nanos++ is linked against the hwloc library, it will be used to get that information automatically. Besides that, if the number of cores per socket, sockets and number of processing elements do not make sense, the number of sockets will be adjusted.

  --num-sockets Sets the number of NUMA nodes.

  --cores-per-socket Sets the number of hardware threads per NUMA node.

  --socket-steal, --no-socket-steal Enable work stealing from the ready queues of other NUMA nodes (default).

  --socket-random-steal, --no-socket-random-steal Instead of round robin stealing, steal from random queues.

  --socket-steal-parents, --no-socket-steal-parents Steal depth 1 tasks instead of child tasks (disabled by default).

  --socket-steal-spin Number of spins before attempting work stealing.

  --socket-smartpriority, --no-socket-smartpriority Enable smart priority propagation (disabled by default).

  --socket-immediate, --no-socket-immediate Use getImmediateSuccessor when prefetching (disabled by default). Note: this is not completely implemented (atBeforeExit).

--socket-steal-low-priority, --no-socket-steal-low-priority Steal low priority tasks from the other nodes' queues. Note: this is currently broken.

Affinity

Warning: This policy has been discontinued in master development branch, but it is still used in the *cluster* development branch

- *Configuration string:* `NX_SCHEDULE=affinity` *or* `NX_ARGS="--schedule=affinity"`
- *Description:* Take into account where the data used by a task is located. Meaningful only in Multiple Address Space architectures (or SMP NUMA). Affinity implements immediate successor by default.

Affinity Smart Priority

Warning: This policy has been discontinued in master development branch, but it is still used in the *cluster* development branch

- *Configuration string:* `NX_SCHEDULE=affinity-smartpriority` *or* `NX_ARGS="--schedule=affinity-smart-priority"`
- *Description:* Affinity policy with smart priority support. Works exactly like the affinity policy (same number of queues, it also has work stealing) but uses sorted priority queues with priority propagation to immediate preceding tasks. When inserting a new task in the queue, if the new task has the same priority as another task already in the queue, the new one will be inserted after the existent one (note that, unlike in the priority scheduling policy, there is no option to change this behaviour). Affinity implements immediate successor by default.

Versioning

- *Configuration string:* `NX_SCHEDULE=versioning` *or* `NX_ARGS="--schedule=versioning"`
- *Description:* This scheduler can handle multiple implementations for the same task and gives support to the `implements` clause. The scheduler automatically profiles each task implementation and chooses the most suitable implementation each time the task must be run. Each thread has its own task queue, where task implementations are added depending on scheduler's decisions. The main criteria to assign a task to a thread is that it must be the earliest executor of that task (this means that the thread will finish task's execution at an earliest time). In some cases (where the number of tasks in the graph is big enough), idle threads are also allowed to run task implementations even though they are not the fastest executors. Specifically, the profile for each task implementation includes its execution time and the data set size it uses and it is used to estimate the behavior of future executions of the same implementation. Versioning implements immediate successor by default.
- *Parameters:*

  --versioning-stack, --no-versioning-stack Retrieving from queue's back or queue's front respectively. Using –versioning-stack parameters transform ready queue into a LIFO (Last In First Out) structure. Effectively, this means that tasks are inserted in queue's back and retrieved from the same position (queue's back). (default behaviour is –no-versioning-stack)

  --versioning-use-stack, --no-versioning-use-stack This is an alias of the previous parameter.

Note: A detailed explanation of this scheduler can be found at the following conference paper: Judit Planas, Rosa Badia, Eduard Ayguade, Jesus Labarta, "Self-Adaptive OmpSs Tasks in Heterogeneous Environments", Proceedings of 27th IEEE International Parallel and Distributed Processing Symposium (IEEE IPDPS) (2013).

Bottom level-aware scheduler

- *Configuration string:* `NX_SCHEDULE=botlev` *or* `NX_ARGS="--schedule=botlev"`
- *Description:* This scheduler targets single-ISA heterogeneous machines that maintain two kinds of cores (fast and slow, such as ARM big.LITTLE). The scheduler detects dynamically the longest path of the task dependency graph and assigns the tasks that belong to this path (critical tasks) to the fast cores of the system. The detection of the longest path is based on the computation and the usage of bottom-level longest-path priorities, that is the length of the longest path in the dependency chains from each node to a leaf node. There are two queues for the ready tasks, one per processor kind. Fast cores retrieve tasks from their unique queue and slow cores retrieve tasks from the other queue. The queues are sorted according to the task priority (bottom level). Work stealing is enabled by default for fast cores: a fast core steals a task from the slow-cores' queue if it is idle. Optionally, work stealing can be performed by both sides if the parameter NX_STEALB is enabled. The policy can be flexible or strict, meaning that the flexible policy considers more tasks as critical, while the strict limits the number of critical tasks.
- *Parameters:*

  –update-freq or NX_BL_FREQ Sets the update frequency of the priorities. By default it is set to zero which means update every time a new dependency occurs. It can be set with any positive integer value and affects the priorities of the tasks. By setting this parameter to a positive value the bottom levels of the tasks are less accurate but the policy has slightly less scheduling overheads. –numSpins or NX_NUM_SPINS Sets the number of spins before attempting work stealing (by default it is set to 300). –from or NX_HP_FROM Sets the thread id of the first fast core. –to or NX_HP_TO Sets the thread id of the last fast core. –strict or NX_STRICTB Defines whether the policy is strict or flexible (–strict=0 for flexible or –strict=1 for strict). –steal or NX_STEALB Defines whether the policy uses uni- or bi-directional work stealing (–steal=0 for uni-directional or –steal=1 for bi-directional).

3.3.2 Throttling policies

The throttle policy determines when tasks are created as entities, that can be scheduled and executed asynchronously, or are executed immediately, as if the code were not meant to be executed in parallel. Applications may be sensitive to this behavior so different throttling policies are provided by Nanos++ in order to allow a more precise tuning of the system.

Usage

Description: Sets the throttling policy that will be used during the execution.
Type: string value.
Environment variable: `NX_THROTTLE=<string>`
Command line flag: `--throttle[ |=]<string>`

List of throttling plugins

Throttle policies are provided in the form of plug-ins. Currently Nanos++ comes with the following throttling policies:

Hysteresis

Important: This is the default throttling policy

- *Configuration string:* `NX_THROTTLE=hysteresis` *or* `NX_ARGS="--throttle=hysteresis"`
- *Description:* Based on the hysteresis mechanisms, once we reach to a upper limit we stop creating tasks but we start to create them again when we reach a lower limit. Upper limit and lower limit are configurable through command line flags:
- *Parameters:*

  --throttle-upper Defines the maximum number of tasks per thread allowed to create new first level's tasks. (default value 500).

  --throttle-lower Defines the number of tasks per thread to re-active first level task creation. (default value 250).

  --throttle-type Defines the target counter when taking into account the number of tasks. User can choose among `ready` or `total`. (default value total).

Task depth

- *Configuration string:* `NX_THROTTLE=taskdepth` *or* `NX_ARGS=--throttle=taskdepth"`
- *Description:* This throttle mechanism is based on the task depth. The runtime will not create tasks further than the nested level specified by the limit.
- *Parameters:*

  --throttle-limit Defines maximum depth for tasks. (default value 4).

Ready tasks

- *Configuration string:* `NX_THROTTLE=readytasks` *or* `NX_ARGS="--throttle=readytasks"`
- *Description:* Defines the throttle policy according with the number of ready tasks while creating a new task.
- *Parameters:*

  --throttle-limit Defines the number of ready tasks we use as limit for task creation. When creating a task, if the number of ready tasks is greater than limit task will not be created, task creation will block and issued to scheduler decision. If number of ready tasks is less or equal than limit, task will be created normally. (default value 100).

Idle threads

- *Configuration string:* `NX_THROTTLE=idlethreads` *or* `NX_ARGS="--throttle=idlethreads`
- *Description:* This throttle policy take the decision of create (or not) a task according with the number of idle threads we have at task creation instant.
- *Parameters:*

--throttle-limit Defines the number of idle threads we use as limit for task creation. When creating a task, if the number of idle threads is less than this value the task will not be created, task creation will block and issued to scheduler decision. If number of idle threads is greater or equal than this limit, the task will be created normally. (default value is 0).

Number of tasks

- *Configuration string:* `NX_THROTTLE=numtasks` *or* `NX_ARGS=--thottle=numtasks`
- *Description:* Defines the throttle policy according with the existing number of tasks at task creation.
- *Parameters:*

--throttle-limit Defines the number of tasks (total tasks already created and not completed, being ready or not) we use as limit for task creation. When creating a task, if the number of on-fly-tasks is greater than limit the task will not be created, task creation will be blocked and issued to scheduler decision. If number of on-fly-tasks is less or equal than this limit, the task will be created normally. (default value 100).

Dummy

- *Configuration string:* `NX_THROTTLE=dummy` *or* `NX_ARGS="--throttle=dummy"`
- *Description:* This throttle policy always takes the decision of create (or not) tasks.
- *Parameters:*

--throttle-create-tasks Set the behaviour of the dummy throttle to force task creation and deferred execution. (default option).

--no-throttle-create-tasks Set the behaviour of the dummy throttle to avoid task creation and force undeferred execution.

Throttle policy examples

If we want to execute our program with 4 threads and using a throttling policy which create tasks only if we have at least one thread idle, we will use the following command line:

```
NX_ARGS="--threads=4 --throttle=idlethreads --throttle-limit=1" ./myProgram
```

If we want to execute our program with 8 threads and using a throttling policy which create tasks only if we do not have at least 50 ready tasks per thread, we will use the following command line:

```
NX_ARGS="--threads=8 --throttle=readytasks --throttle-limit=50" ./myProgram
```

3.3.3 Instrumentation modules

The instrumentation flag selects among all available instrumentation plug-ins which one to use. An instrumentation plug-in generates output information which describes one or more aspects of the actual execution. This module is in charge of translate internal Nanos++ events into the desired output: it may be just showing the event description in the screen (as soon as they occur), generates a trace file which can be processed by other software component, build a dependence graph (taking into account only dependence events), or just call an external service when a given

type of event occurs. In general the events are taking place inside the runtime but also they can be generated by the application's programmers or introduced by the compiler (e.g. outlined functions created by Mercurium compiler).

The instrumentation plug-in can be specified by the `NX_INSTRUMENTATION` environment variable or the `--instrumentation` option included in the `NX_ARGS` environment variable:

```
$export NX_INSTRUMENTATION=<module>
$./my_program

$export NX_ARGS="--instrumentation[ \|=]<module> ..."
$./my_program
```

--instrumentation=<module> It sets the instrumentation backend to be used during the execution. Module can be one of the following options: `empty`, `print_trace`, `extrae`, `graph`, `ayudame` or `tasksim tdg`.

By default, only a set of Nanos++ events are enabled (see Nanos++ events section for further details). Additionally you can also use following options in order to specify the set of events that will be instrumented for a given execution.

--instrument-default=<mode> Set the instrumentation event set by default. Where `mode` can be one of the following options: `none` disabling all events, or `all` enabling all events. Additionally we can choose one instrumentation profile: `advanced`, `developer` and `user`. Interval and punctual event table shows which of these events are generated when activating one of these levels. (Default value is `user`).

--instrument-enable=<event-type> Add events to the current instrumentation event list. Where `event-type` can be one of the following options: `state` enabling all the **state** events, `ptp` enabling all the **point-to-point** events, or an specific *event-prefix* enabling those **interval** and/or **punctual** events whose name matches with *event-prefix*. Users can also combine several instrument-enable options in the same execution environment (e.g. `NX_ARGS="... --instrument-enable=state --instrument-enable=for"`).

--instrument-disable=<event-type> Remove events to the current instrumentation event list. Where `event-type` can be one of the following options: `state` disabling all the **state** events, `ptp` disabling all the **point-to-point** events, or an specific *event-prefix* disabling those **interval** and/or **punctual** events whose name matches with *event-prefix*. Users can also combine several instrument-disable options in the same execution environment (e.g. `NX_ARGS="... --instrument-disable=state --instrument-disable=for"`).

--instrument-cpuid, --no-instrument-cpuid Add cpuid event when binding is disabled (expensive).

Nanos++ implements several instrumentation plug-ins. The following sections will show a description of each of the available plug-ins with respect to that feature.

Plug-in description

Empty Trace

Important: This is the default instrumentation plugin.

- *Configuration String:* `NX_INSTRUMENTATION=empty_trace` *or* `NX_ARGS="--instrumentation=empty_trace"`
- *Description:* It does not generate any output.

Extrae

- *Configuration String:* `NX_INSTRUMENTATION=extrae` *or* `NX_ARGS="--instrumentation=extrae"`
- *Description:* It generates a Paraver trace using Extrae library

Paraver is a flexible parallel program visualization and analysis tool based on an easy-to-use Motif GUI. Paraver was developed responding to the need of having a qualitative global perception of the application behavior by visual inspection and then to be able to focus on the detailed quantitative analysis of the problems. Paraver provides a large amount of information useful to decide the points on which to invest the programming effort to optimize an application.

Before getting any trace we will need to compile our application with instrumentation support. Compile your application using Mercurium with the `--instrument` flag:

```
$ mcc --ompss --instrument my_program.c -o my_program
```

In order to get a Paraver's Trace you will need to execute your OmpSs program with the instrumentation plugin Extrae using the environment variable NX_INSTRUMENTATION. That will enable the Extrae plugin which translates all internal events into a Paraver events using the Extrae library. Extrae is completely configured through a XML file that is set through the EXTRAE_CONFIG_FILE environment variable:

```
$ export NX_INSTRUMENTATION=extrae
$ export EXTRAE_CONFIG_FILE=config.xml
$ .\my_program
```

Important: Extrae library installation includes several XML file examples to serve as a basis for the end user (see $EXTRAE_DIR/share/example). Check your Extrae installation to get one of them and devote some time in tuning the parameters. This file is divided in diferent sections in which you can enable and disable some Extrae features.

XML's merge section enabled (recommended)

If the merge section is enabled the merge process will be automatically invoked after the application run. You can enable or disable this Extrae feature by just setting the enabled flag to yes/no. The value of this XML node (in between `<merge>` and `</merge>` tags) is used as the tracefile name (i.e. my_program_trace.prv in the following example). You can check other flags meanings in the Extrae User's Guide:

```
<merge enabled="yes"
   synchronization="default"
   binary="my_program"
   tree-fan-out="16"
   max-memory="512"
   joint-states="yes"
   keep-mpits="no"
   sort-addresses="yes"
   remove-files="yes"
>
```

```
  my_program_trace.prv
</merge>
```

XML's merge section disabled

If the merge section is disabled (or you are not using the XML configuration file), once you have run your instrumented program Extrae library will produce several files containing all the information related with the execution. There will be as many .mpit files as tasks and threads where running the target application. Each file contains information gathered by the specified task/thread in raw binary format. A single .mpits file that contain a list of related .mpit files and, if the DynInst based instrumentation package was used, an addition .sym file that contains some symbolic information gathered by the DynInst library.

In order to use Paraver, those intermediate files (i.e., .mpit files) must be merged and translated into a Paraver trace file format. To proceed with any of these translation all the intermediate trace files must be merged into a single trace file using one of the available mergers in the Extrae bin directory.

Using Hardware Counters (PAPI)

In order to get a Paraver trace files including hardware counters (HWC) we need to verify that our Extrae library has been installed with PAPI support. Once we have verified this we can turn on the generation of PAPI hardware counter events through the XML Extrae configuration file section: counters. Here is an example of the counters section in the XML configuration file:

```
<counters enabled="yes">
  <cpu enabled="yes" starting-set-distribution="1">
    <set enabled="yes" domain="all" changeat-time="1s">
      PAPI_TOT_INS,PAPI_TOT_CYC,PAPI_L1_DCM
      <sampling enabled="yes" period="100000000">PAPI_TOT_CYC</sampling>
    </set>
    <set enabled="yes" domain="all" changeat-time="1s">
      PAPI_TOT_INS,PAPI_TOT_CYC,PAPI_FP_INS
    </set>
  </cpu>
  <network enabled="no" />
  <resource-usage enabled="no" />
</counters>
```

Processor performance counters are configured in the <cpu> nodes. The user can configure many sets in the <cpu> node using the <set> node, but just one set will be active at any given time. In the example two sets are defined. First set will read PAPI_TOT_INS (total instructions), PAPI_TOT_CYC (total cycles) and PAPI_L1_DCM (1st level cache misses). Second set is configured to obtain PAPI_TOT_INS (total instructions), PAPI_TOT_CYC (total cycles) and PAPI_FP_INS (floating point instructions). You can get a list of PAPI Standard Events By Architecture here. In order to change the active set you can use the changeat-time attribute specifying the minimum time to hold the set (i.e. changeat-time="1s" in the example).

See also the list of PAPI Standard Events by Architecture and the Manuals of the BSC Performance Tools.

Task Sim

Warning: This plugin is experimental

- *Configuration String:* `NX_INSTRUMENTATION=tasksim` *or* `NX_ARGS="--instrumentation=tasksim"`
- *Description:* It generates a simulation trace which can be used with TaskSim

Ayudame

Ayudame plug-in allows to interact current program execution with Temanejo graphical debugger. The main goal is to display the task-dependency graph, and to allow simple interaction with the runtime system in order to control some aspects of the parallel execution of an application.

Ayudame plug-in is not included by default in Nanos++ runtime package. In Temanejo's manual (see references at the end of the section) you will find detailed instructions about how to get and install this plug-in.

Before using Ayudame/Temanejo interaction you will need to compile your application with instrumentation support. Compile your application using Mercurium compiler and the `--instrument` flag:

```
$ mcc --ompss --instrument my_program.c -o myProgram
```

At runtime you will need to enable Ayudame plug-in using instrumentation option:

```
$export NX_INSTRUMENTATION=ayudame
$./myProgram

$export NX_ARGS="--instrumentation[ \|=]ayudame ..."
$./myProgram
```

You can find more information about Ayudame/Temanejo in HLRS web site.

Task Dependency Graph

- *Configuration String:* `NX_INSTRUMENTATION=tdg` *or* `NX_ARGS="--instrumentation=tdg"`
- *Description:* It generates a .dot file with all the real dependences produced in that particular execution.

This plugin also admit some extra parameters:

--node-size=<string> Defines the semantic for the node size. Configuration string can be:

- `constant`: the size of all nodes is the same (default).
- `linear`: the size of the nodes increases linearly with the task time.
- `log`: the size of the nodes increases logarithmically with the task time.

This instrumentation plugin generates a dependency graph using Graphviz (http://www.graphviz.org) format.

When compiling with Mercurium, you will need to use the flag `--instrument`. At runtime you will need to enable the `tdg` plugin using the environment variable `NX_ARGS="--instrument=tdg"`. The execution of the program will generate `graph.dot` file, and also (if `dot` program available) a `graph.pdf` file.

Dependences shown in the graph are real dependences, which means that are dependences that actually happen during that specific execution.

Usually one wants all the theoretical dependences in the program. To do this you have to run your program with a single thread and ensure that the schedule creates all the tasks (i.e. does not decide to immediately run them). To achieve this use `NX_ARGS="--smp-workers=1 --throttle=dummy"`, see example below.

This version of the plugin is thread-safe, so you do not need to execute your program with a unique thread to draw the TDG (although it is recommended to draw properly your program dependences).

In the graph, task nodes are colored according with the function they represent (the plugin includes a legend in the right-top part showing equivalence between colors and functions). The size of the node may also represent time, which means that bigger tasks are more time expensive than the smaller ones. Finally the `tdg` plugin also shows two different kind of arrows. Solid arrows are true dependences, while dashed arrows are used to specify anti/output dependences as defined in the legend. Gray arrows show nested relationships.

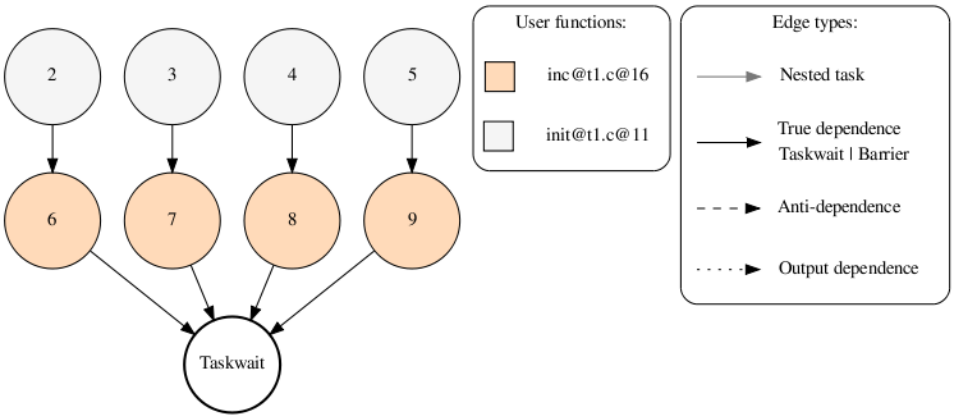
Output Example

Running the following code with this plugin will produce the next graph:

```
int a[N];

for(int i=0; i<N; ++i) {
   #pragma omp task out(a[i]) label(init) firstprivate(i)
   a[i]=0;
}

for(int i=0; i<N; ++i) {
   #pragma omp task inout(a[i]) label(inc) firstprivate(i)
   a[i]++;
}

#pragma omp taskwait

for(int i=0; i<N; ++i) {
   printf("%d,",a[i]);
}
```

Command line:

```
$ NX_ARGS="--smp-workers=1 --throttle=dummy --instrumentation=tdg" ./myprogram
```



Nanos++ events

State events

In Nanos++ programming model a thread can represent whether a POSIX thread or a runtime Work Descriptor according with the instrumentation model we were using.

##	Event	Description
00	NANOS_NOT_CREATED	Thread has not been created yet.
01	NANOS_NOT_RUNNING	Thread is executing nothing (e.g. a Work Descriptor which is not in execution or sub state when state is enabled).
02	NANOS_STARTUP	Thread is executing runtime start up.
03	NANOS_SHUTDOWN	Thread is executing runtime shut down.
04	NANOS_ERROR	Error while instrumenting.
05	NANOS_IDLE	Thread is on idle loop.
06	NANOS_RUNTIME	Thread is executing generic runtime code (not defined in other state).
07	NANOS_RUNNING	Thread is executing user code. Main thread executing main(), or when executing an outlined user function. This state is directly related with "user-code" event which register Work Descriptor Id.
08	NANOS_SYNCHRONIZATION	Thread is executing a synchronization mechanism: team barrier, task wait, synchronized condition (wait and signal), wait on dependencies, set/unset/try locks (through Nanos API), acquiring a single region, waitOnCondition idle loop and waking up a blocked Work Descriptor.
09	NANOS_SCHEDULING	Thread is calling a scheduler policy method, submitting a new task or a Work Descriptor is yielding thread to a new one.
10	NANOS_CREATION	Thread is creating a new Work Descriptor or setting translate function.
11	NANOS_MEM_TRANSFER_IN	Thread is copying data to cache.

Point to Point events

##	Event	Description
0	NANOS_WD_DOMAIN	WD's Identifier
1	NANOS_WD_DEPENDENCY	Data Dependency
2	NANOS_WAIT	Hierarchical Dependency (e.g "omp taskwait")
3	NANOS_WD_REMOTE	Remote Workdescriptor Execution (among nodes)
4	NANOS_XFER_PUT	Transfer PUT (Data Send)
5	NANOS_XFER_GET	Transfer GET (Data Request)

Only in cluster branch:

##	Event	Description
6	NANOS_AM_WORK	Sending WDs (master -> slave)
7	NANOS_AM_WORK_DONE	WD is done (slave->master)

Interval/Punctual events

###	Event	Description	ADV	DEV	USR
001	api	Nanos Runtime API (interval)	X	X	
002	wd-id	Work Descriptor id (interval)	X	X	
003	cache-copy-in	Transfer data into device cache (interval)			
004	cache-copy-out	Transfer data to main memory (interval)			
005	cache-local-copy	Local copy in device memory (interval)			
006	cache-malloc	Memory allocation in device cache (interval)			
007	cache-free	Memory free in device cache (interval)			
008	cache-hit	Hit in the cache (interval)			
009	copy-in	Copying WD inputs (interval)			
010	copy-out	Copying WD outputs (interval)			
011	user-funct-name	User Function Name (interval)	X	X	X
012	user-code	User Code (wd) (interval)			
013	create-wd-id	Create WD Id: (punctual)			
014	create-wd-ptr	Create WD pointer: (punctual)	X		
015	wd-num-deps	Create WD num. deps. (punctual)	X		
016	wd-deps-ptr	Create WD dependence pointer (punctual)	X		
017	lock-addr	Lock address (interval)	X		
018	num-spins	Number of Spins (punctual)	X	X	
019	num-yields	Number of Yields (punctual)	X	X	
020	time-yields	Time on Yield (in nsecs) (punctual)	X	X	
021	user-funct-location	User Function Location (interval)	X	X	X
022	num-ready	Number of ready tasks in the queues (punctual)	X	X	X
023	graph-size	Number tasks in the graph (punctual)	X	X	X
024	loop-lower	Loop lower bound (punctual)	X	X	
025	loop-upper	Loop upper (punctual)	X	X	

###	Event	Description	ADV	DEV	USR
026	loop-step	Loop step (punctual)	X	X	
027	in-cuda-runtime	Inside CUDA runtime (interval)	X	X	
028	xfer-size	Transfer size (punctual)	X	X	
029	cache-wait	Cache waiting for something (interval)			
030	chunk-size	Chunk size (punctual)	X	X	
031	num-sleeps	Number of sleeps (punctual)	X	X	
032	time-sleeps	Time on Sleep (in nsecs) (punctual)	X	X	
033	num-scheds	Number of scheduler operations (punctual)	X	X	
034	time-scheds	Time in scheduler operations (in nsecs) (punctual)	X	X	
035	sched-versioning	Versioning scheduler decisions (punctual)	X		
036	dependence	Dependence analysis, system have found a new dependence (punctual)	X	X	
037	dep-direction	Dependence direction (punctual)	X	X	
038	wd-priority	WD's priority (punctual)	X	X	
039	in-opencl-runtime	In OpenCL runtime (interval)	X	X	
040	taskwait	Taskwait (interval)	X	X	X
041	set-num-threads	Set/change number of threads (punctual)	X	X	X
042	cpuid	CPU id (punctual) - Xdisabled by default	X	X	X
043	dep-address	Dependence address (punctual)	X	X	
044	copy-data-in	Work descriptor id that is copying data in	X	X	
045	cache-copy-dada-in	Work descriptor id that is copying data in	X	X	
046	cache-copy-data-out	Work descriptor id that is copying data out	X	X	
047	sched-affinity-const	Constraint used in affinity scheduler	X	X	
048	in-mpi-runtime	Inside MPI runtime	X	X	
049	wd-ready	Workdescriptor becomes ready	X		
050	wd-blocked	Workdescriptor becomes blocked	X		

###	Event	Description	ADV	DEV	USR
051	parallel-outline-fct	Parallel outline function	X		
052	async-thread	Asynchronous thread state events	X	X	
053	copy-in-gpu	Asynchronous memory copy from host to device			
054	copy-out-gpu	Asynchronous memory copy from device to host			
055	gpu-wd-id	GPU's work descriptor id	X	X	
056	wd-criticality	Work descriptor criticality	X	X	
057	blev-overheads	Total overheads of botlev scheduler	X	X	
058	blev-overheads-break	Overheads of botlev scheduler	X	X	
059	critical-wd-id	A critical work descriptor is submitted	X	X	
060	copy-dir-devices	Asynchronous memory copy between host and device	X	X	X
061	concurrent-task	Number of concurrent task in the ready queue	X	X	
062	network-transfer	Network transfer to node	X	X	
064	thread-numa-node	NUMA node of the worker thread	X	X	
065	wd-numa-node	NUMA node assigned to the work descriptor	X	X	
066	steal	If the work descriptor to be executed is the result of a steal operation	X		

3.3.4 Barrier algorithms

Usage

Description: Selects the barrier algorithm used during the execution.
Type: string value.
Environment variable: `NX_BARRIER=<string>`
Command line flag: `--barrier[ |=]<string>`

List of barrier plugins

Centralized

Important: This is the default barrier algorithm

- *Configuratin string:* `NX_BARRIER=centralized` *or* `NX_ARGS="--barrier=centralized`
- *Description:* All the threads are synchronized following a centralized structure.

Tree

Warning: This barrier is experimental.

- *Configuration string:* `NX_BARRIER=tree` *or* `NX_ARGS="--barrier=tree"`
- *Description:* Threads are synchronized following a tree structure.

3.3.5 Dependence managers

Nanos++ provides several plugins to handle task dependencies, with different performance and features.

Usage

Description: Changes the dependency plugin to use.
Type: string value
Environment variable: `NX_DEPS=<string>`
Command line flag: `--deps[ |=]<string>`

List of dependence plugins

Plain

Important: This is the default dependence plugin

- *Configuration string:* `NX_DEPS=plain` *or* `NX_ARGS="--deps=plain"`
- *Description:* This plugin uses single memory addresses to compute dependences. In most of the cases the programmer can use a single memory point as a sentinel for a whole region.

Regions

- *Configuration string:* `NX_DEPS=regions` *or* `NX_ARGS="--deps=regions"`
- *Description:* This plugin does not partition regions. It is more suitable for task with halos.

Perfect-regions

- *Configuration string:* `NX_DEPS=perfect-regions` *or* `NX_ARSG="--deps=perfect-regions"`
- *Description:* Partitions regions into perfect regions. It is recommended for applications like multisort.

Contiguous regions

Important: This is a test and functional plugin, so it's performance may not be the best

- *Configuration string:* `NX_DEPS=cregions` *or* `NX_ARSG="--deps=cregions"`
- *Description:* Detects dependencies between regions defined by start address and end address.

Contiguous regions nocache

Important: This is a test and functional plugin, so it's performance may not be the best

- *Configuration string:* `NX_DEPS=cregions_nocache` *or* `NX_ARSG="--deps=cregions_nocache"`
- *Description:* Similar to Contiguous regions but small regions go to a dedicated structure (performance may be better in some applications).

Dependence management examples

The default plugin is called plain, and it can only handle simple dependencies that do not overlap. The following is allowed:

```
#pragma omp task inout( [NB]block )
task_1();

#pragma omp task inout( [NB]block )
task_2();

#pragma omp task inout( block[0;NB] )
task_3();
```

The following requires regions support, as provided by the regions and perfect-regions plugins:

```
#pragma omp task inout( block[0;3] )
task_4();

#pragma omp task inout( block[1;3] )
task_5();
```

Here task 5 will not depend on task 4 when using the plain plugin.

Regions are also required when you have non-contiguous memory, such as matrices. Although regions are much more powerful, there is a drawback with the current regions plugins: alignment. Keep in mind that if your data is not properly aligned, there will be a potential task serialisation and therefore a huge performance toll.

3.3.6 Thread Manager

The Thread Manager module controls the amount of working threads needed for a specific amount of workload. It could be useful to block idle threads when they are not needed or even to create auxiliary threads in a heavy work load scenario if we have resources to spare.

Thread Manager options

--thread-manager=<none,nanos,dlb> Select which Thread Manager will be used

--enable-dlb Enable inter-process management with Dynamic Load Balancing (DLB) library

--enable-block Enable thread blocking on idle loop

--enable-sleep Enable thread sleeping on idle loop

--sleep-time=<n> Set the amount of time (in nsec) in each sleeping phase

--enable-yield Enable thread yielding on idle loop

--yields=<n> Set number of yields before blocking

--force-tie-master Force Master WD (user code) to run on Master Thread

--warmup-threads Force the creation of as many threads as available CPUs at initialization time, then block them immediately if needed.

List of thread managers

Nanos++ provides the selection of the thread manager to use through the option `--thread-manager`. The default value is *none*, unless one or more of the options `--enable-dlb`, `--enable-yield`, `--enable-block` or `--enable-sleep` are selected, which in this case the default value is *nanos*.

Thread Manager: Nanos

If `--enable-block` is used, Nanos++ will temporarily block those threads that are not considered useful. As soon as Nanos++ detects an increase in the workload, the threads will be activated again.

Alternatively, if `--enable-sleep` is used, Nanos++ will temporarily sleep for a fixed amount of time those threads that are not considered useful. After that, the threads are activated again.

Thread Manager: DLB

Thread manager DLB (do not confuse with `--enable-dlb`, which only enables some DLB features), yields the complete thread control to the DLB library. Each thread can be individually blocked as in the Nanos Thread Manager but DLB can give the resources associated to this thread to another process.

3.4 Extra Modules (plug-ins)

Plugins are also used in some other library modules. This allows to implement several versions for the same algorithm and also offers an abstraction layer to access them. These extra plugins modules are:

- Slicers: tasks that can be eventually divided in more tasks.
- Worksharings: a list of work items which can be executed in parallel without creating a new task (e.g. in loops, a chunk of iterations will be a work item).

These plug-ins can be set in user source code using the proper OmpSs mechanism. Slicers implement task generating loops constructs (i.e. an OmpSs loop construct). Worksharing implement work distribution among the team of threads encountering the construct (e.g. OpenMP loop constructs).

CHAPTER

FOUR

INSTALLATION OF OMPSS FROM GIT

Sometimes it may happen that you have to compile OmpSs from the git repository.

Note: There is no need to compile from git, you can always use a tarball. See *Installation of OmpSs*.

4.1 Additional requirements when building from git

- Automake 1.9 or better. Get it at http://ftp.gnu.org/gnu/automake
- Autoconf 2.60 or better. Get it at http://ftp.gnu.org/gnu/autoconf
- Libtool 2.2 or better. Get it at http://ftp.gnu.org/gnu/libtool
- Git. Get it at http://git-scm.com/download/linux

Note: It is likely that your Linux distribution or system already contains these tools packaged or installed. Check the documentation of your Linux distribution or ask your administrator.

4.2 Nanos++ from git

1. Make sure you fulfill the requirements of Nanos++. See *Nanos++ build requirements*.
2. Clone the Nanos++ repository:

```
$ git clone http://pm.bsc.es/git/nanox.git
Cloning into 'nanox'...
remote: Counting objects: 22280, done.
remote: Compressing objects: 100% (7567/7567), done.
remote: Total 22280 (delta 17397), reused 18399 (delta 14290)
Receiving objects: 100% (22280/22280), 3.50 MiB, done.
Resolving deltas: 100% (17397/17397), done.
```

3. Run `autoreconf` in the newly created `nanox` distribution:

```
$ cd nanox
$ autoreconf -fiv
autoreconf: Entering directory `.'
autoreconf: configure.ac: not using Gettext
autoreconf: running: aclocal --force -I m4
```

```
autoreconf: configure.ac: tracing
autoreconf: running: libtoolize --copy --force
libtoolize: putting auxiliary files in `.'.
libtoolize: copying file `./ltmain.sh'
libtoolize: putting macros in AC_CONFIG_MACRO_DIR, `m4'.
libtoolize: copying file `m4/libtool.m4'
libtoolize: copying file `m4/ltoptions.m4'
libtoolize: copying file `m4/ltsugar.m4'
libtoolize: copying file `m4/ltversion.m4'
libtoolize: copying file `m4/lt~obsolete.m4'
autoreconf: running: /usr/bin/autoconf --force
autoreconf: running: /usr/bin/autoheader --force
autoreconf: running: automake --add-missing --copy --force-missing
configure.ac:149: installing './ar-lib'
configure.ac:153: installing './compile'
configure.ac:10: installing './config.guess'
configure.ac:10: installing './config.sub'
configure.ac:15: installing './install-sh'
configure.ac:15: installing './missing'
Makefile.am: installing './INSTALL'
src/apis/c/debug/Makefile.am: installing './depcomp'
autoreconf: Leaving directory `.'
```

4. Now follow steps in *Installation of Nanos++*, starting from the step where you have to run `configure`.

4.3 Mercurium from git

You can find the instructions to build the latest version of Mercurium from our GitHub repository in the following link: https://github.com/bsc-pm/mcxx/blob/master/README.md

CHAPTER

FIVE

FAQ: FREQUENTLY ASKED QUESTIONS

5.1 What is the difference between OpenMP and OmpSs?

5.1.1 Initial team and creation

You must compile with `--ompss` flag to enable the OmpSs programming model. While both programming models are pretty similar in many aspects there are some key differences.

In OpenMP your program starts with a team of one thread. You can create a new team of threads using `#pragma omp parallel` (or a combined parallel worksharing like `#pragma omp parallel for` or `#pragma omp parallel sections`).

In OmpSs your program starts with a team of threads but only one runs the `main` (or `PROGRAM` in Fortran). The remaining threads are waiting for work. You create work using `#pragma omp task` or `#pragma omp for`. One of the threads (including the one that was running `main`) will pick the created work and execute it.

This is the reason why `#pragma omp parallel` is ignored by the compiler in OmpSs mode. Combined worksharings like `#pragma omp parallel for` and `#pragma omp parallel sections` will be handled as if they were `#pragma omp for` and `#pragma omp sections`, respectively.

Mercurium compiler will emit a warning when it encounters a `#pragma omp parallel` that will be ignored.

5.1.2 Worksharings

In OpenMP mode, our worksharing implementation for `#pragma omp for` (and `#pragma omp parallel for`) uses the typical strategy of:

```
begin-parallel-loop
  code-of-the-parallel-loop
end-parallel-loop
```

In OmpSs mode, the implementation of `#pragma omp for` exploits a Nanos++ feature called *slicers*. Basically the compiler creates a task which will create internally several more tasks, each one implementing some part of the iteration space of the parallel loop.

These two implementations are mostly equivalent except for the following case:

```
int main(int argc, char** argv)
{
    int i;
#pragma omp parallel
    {
        int x = 0;
```

```
#pragma omp for
        for (i = 0; i < 100; i++)
        {
            x++;
        }
    }

    return 0;
}
```

In OmpSs, since `#pragma omp parallel` is ignored, there will not be an `x` variable per thread (like it would happen in OpenMP) but just an `x` shared among all the threads running the `#pragma omp for`.

5.2 How to create burst events in OmpSs programs

Burst events are such with begin/end boundaries. They are useful to measure when we are starting to execute a code and when we are finalizing. Lets imagine we have the following code:

```
#include <stdio.h>

#define ITERS 10

#pragma omp task
void f( int n )
{
   usleep(100);

   fprintf(stderr,"[%3d]",n);

   usleep(200);
}

int main (int argc, char *argv[] )
{
   for (int i=0; i<ITERS; i++ )
      f(i);
#pragma omp taskwait
   fprintf(stderr,"\n");
}
```

And we want to distinguish the time consumed in the first `usleep` from the second. We will call them *Phase 1* and *Phase 2* respectively. So we will need one new type of events (Key) and 2 new values with description. So first we need to declare a `nanos_event_t` variable and initializes type, key and value members. The type member will be `NANOS_BURST_START` or `NANOS_BURST_END` depending if we are open or closing the burst. We also will use `nanos_instrument_register_key` and `nanos_instrument_register_value` allowing Nanos++ to generate the appropriate Paraver configuration file. Finally we have to raise the events from the proper place. In our case we want to surround `usleep` function calls.

The code in the function task `f` will be:

```
#pragma omp task
void f( int n )
{
   nanos_event_t e1, e2;
```

```
   // Registering new event key
   nanos_instrument_register_key ( &e1.key, "phases-of-f", "Phases of f()", false );
   nanos_instrument_register_key ( &e2.key, "phases-of-f", "Phases of f()", false );

   // Registering new event values (for key "phases-of-f")
   nanos_instrument_register_value ( &e1.value, "phases-of-f", "phase-1", "Phase 1",␣
↪false);
   nanos_instrument_register_value ( &e2.value, "phases-of-f", "phase-2", "Phase 2",␣
↪false);

   // First phase
   e1.type = NANOS_BURST_START;
   nanos_instrument_events( 1, &e1);

   usleep(100);

   e1.type = NANOS_BURST_END;
   nanos_instrument_events( 1, &e1);

   fprintf(stderr,"[%3d]",n);

   // Second phase
   e2.type = NANOS_BURST_START;
   nanos_instrument_events( 1, &e2);

   usleep(200);

   e1.type = NANOS_BURST_END;
   nanos_instrument_events( 1, &e2);
}
```

Additionally you can use the `nanos_instrument_begin_burst()` and `nanos_instrument_end_burst()` which actually wrap the behaviour of opening and closing the events. And if you don't need to register the values you can use the functions `nanos_instrument_begin_burst_with_val()` and `nanos_instrument_end_burst_with_val`:

```
#pragma omp task
void f( int n )
{
   // Raising open event for phase-1
   nanos_instrument_begin_burst ( "phases-of-f", "Phases of f()", "phase-1", "Phase 1
↪" );

   usleep(100);

   // Raising close event for phase-1
   nanos_instrument_end_burst ( "phases-of-f", "phase-1" );

   fprintf(stderr,"[%3d]",n);

   // Raising open event for phase-2
   nanos_instrument_begin_burst ( "phases-of-f", "Phases of f()", "phase-2", "Phase 2
↪" );

   usleep(200);

   // Raising close event for phase-2
   nanos_instrument_end_burst ( "phases-of-f", "phase-2" );
```

```
   for(int i=0; i<n; i++) {
      nanos_event_value_t ev = i;
      // Raising open event for iteration i
      nanos_instrument_begin_burst_with_val ( "iterations-of-f", "Iterations of f()",␣
↪&ev );
      usleep(100);
      // Raising close event for iteration i
      nanos_instrument_end_burst_with_val ( "iterations-of-f", &ev );
   }
}
```

This mechanism can also be used in Fortran codes as in the follow example:

```
!$OMP TASK
SUBROUTINE F()
    IMPLICIT NONE
    CHARACTER(LEN=*) :: KEY = "phase-of-f" // ACHAR(0)
    CHARACTER(LEN=*) :: KEY_DESCR = "phase of f()" // ACHAR(0)
    CHARACTER(LEN=*) :: VAL = "phase-1" // ACHAR(0)
    CHARACTER(LEN=*) :: VAL_DESCR = "Phase 1" // ACHAR(0)
    INTEGER :: ERROR

    INTEGER, EXTERNAL :: NANOS_INSTRUMENT_BEGIN_BURST
    INTEGER, EXTERNAL :: NANOS_INSTRUMENT_END_BURST

    ERROR = NANOS_INSTRUMENT_BEGIN_BURST(KEY,KEY_DESCR,VAL, VAL_DESCR)
    CALL SLEEP(1)
    ERROR = NANOS_INSTRUMENT_END_BURST(KEY,VAL)
END SUBROUTINE F
```

5.3 How to execute hybrid (MPI+OmpSs) programs

You have an MPI source code annotated with OmpSs directives and you wonder how to compile and execute it. There are some additional steps in order to obtain a Paraver trace file.

5.3.1 Compilation

The idea here is to compile and link the source code with the Mercurium compiler adding any library, include file and flags that are normally used by the MPI compilers. Usually MPI compilers offer some option to get all the information needed. For example, OpenMPI compiler offers the flags `-showme:compile` and `-showme:link` (others have `-compile-info` and `-link-info`, consult your specific compiler documentation). To compile your application you should do something like:

```
$ mcc --ompss $(mpicc -showme:compile) -c hello.c
$ mcc --ompss $(mpicc -showme:link) hello.o -o hello
```

If you need instrumentation just remember to include the `--instrumentation` flag.

5.3.2 Execution

Just use the usual commands to execute an MPI application (consult your specific documentation). Regarding OmpSs the only requirements are that you export all needed variables in each MPI node. For example: To execute the previous

binary with 2 MPI nodes and 3 threads per node using a SLURM queue management:

```
#@ job_name = hello
#@ total_tasks = 2
#@ wall_clock_limit = 00:20:00
#@ tasks_per_node = 1
#@ cpus_per_task = 3

export NX_ARGS="--pes 3"
srun ./hello
```

When running more than one process per node, it is recommended to double-check the CPUs used by each OmpSs process. Job schedulers usually place MPI processes into different CPU sets so they don't overlap, but mpirun does not by default. You can get which CPUs are being used by each process by adding the flag `--summary` to the NX_ARGS environment variable:

```
$ NX_ARGS="--summary" mpirun -n 4 --cpus-per-proc 4 ./a.out 2>&1 | grep CPUs
MSG: [?] === Active CPUs:    [ 4, 5, 6, 7, ]
MSG: [?] === Active CPUs:    [ 8, 9, 10, 11, ]
MSG: [?] === Active CPUs:    [ 12, 13, 14, 15, ]
MSG: [?] === Active CPUs:    [ 0, 1, 2, 3, ]
```

5.3.3 Instrumentation

The required steps to instrument both levels of parallelism are very similar to the process described here: *Extrae*. The only exception is that we have to preload the Extrae MPI library to intercept the MPI events.

When an MPI application is launched through an MPI driver (mpirun) or a Job Scheduler specific command (srun) it is not clear whether the environment variables are exported, or may be you need to define a variable after the MPI launcher has already done the setup but just before your binary starts. This is why we consider a good practice to define the environment variables we wish to set into a shell script file, which each spawned process will run. For instance, we create a script file trace.sh with execution permission that contains:

```
#!/bin/bash
### Other options ###
export NX_ARGS= ...
#####################
export NX_INSTRUMENTATION=extrae
export EXTRAE_CONFIG_FILE=extrae.xml
export LD_PRELOAD=$EXTRAE_HOME/lib/libnanosmpitrace.so  # or libnanosmpitracef.so for␣
↪Fortran
$*
```

We then run our MPI application like this:

```
$ mpirun --some-flags ./trace.sh ./hello_instr
```

If the `extrae.xml` was configured as recommended in the *Extrae* section, a hybrid MPI+OmpSs Paraver trace will be automatically merged after the execution is ended. Otherwise just run:

```
$ mpi2prv -f TRACE.mpits
```

5.4 How to exploit NUMA (socket) aware scheduling policy using Nanos++

In order to use OmpSs in NUMA system we have developed a special scheduling policy, which also supports other OmpSs features as task priorities.

We have tested this policy in machines with up to 8 NUMA nodes (48 cores), where we get about 70% of the peak performance in the Cholesky factorisation. We would appreciate if you shared with us the details of the machine where you plan to use OmpSs.

Important: This scheduling policy works best with the Portable Hardware Locality (hwloc) library. Make sure you enabled it when compiling Nanos++. Check *Nanos++ configure flags*.

Important: Memory is asigned to nodes in pages. A whole page can only belong to a single NUMA node, thus, you must make sure memory is aligned to the page size. You can use the aligned attribute or other allocation functions.

This policy assigns tasks to threads based on either data copy information or programmer hints indicating in which NUMA node that task should run.

You must select the NUMA scheduling policy when running your application. You can do so by defining NX_SCHEDULE=socket or by using –schedule=socket in NX_ARGS. Example:

```
$ NX_SCHEDULE=socket ./my_application
$ NX_ARGS="--schedule=socket" ./my_application
```

5.4.1 Automatic NUMA node discovery

The NUMA scheduling policy has the ability to detect initialisation tasks and track where your data is located. This option is disabled by default and can be selected by supplying –socket-auto-detect in NX_ARGS. Like this:

```
$ NX_ARGS="--schedule=socket --socket-auto-detect" ./my_application
```

This automatic feature has a few requirements and assumptions:

- First-touch NUMA policy is in effect (default in most systems).
- Data is initialised by OmpSs.
- Copies are enabled (either manually, using copy_in|out|inout; or copy_deps, enabled automatically by the compiler).
- Initialisation tasks can only be detected if they are SMP tasks with at least one output whose produced version will be one.

Important: Initialisation tasks will be assigned to NUMA nodes so that your data is initialised in round-robin. If this does not suit you, head over to the programming hints section below.

Some examples:

```
// Default OmpSs configuration: copy_deps enabled by the compiler
#pragma omp task out( [N][N]block )
void init_task( float * block )
```

```
{
}

// Alternate OmpSs configuration: copy_deps manually activated
#pragma omp target device(smp) copy_deps
#pragma omp task out( [N][N]block )
void init_task( float * block )
{
}

// No dependencies defined, copies manually defined
#pragma omp target device(smp) copy_out( [N][N]block )
#pragma omp task
void init_task( float * block )
{
}

// This one will not be detected
#pragma omp task
void not_valid_init_task( float * block )
{
}

// Nor will be this one
#pragma omp task inout( [N][N]block )
void not_init_task( float * block )
{
}
```

The rest of the application would be like a normal OmpSs one:

```
#pragma omp task inout( [N][N] block )
void compute_task( float* block )
{
    for( /* loop here */ )
    {
        // Do something
    }
}

int main(int argc, char* argv[])
{

  // Allocate matrix A

  // Call init tasks
  for( int i = 0; i < nb*nb; ++i )
  {
      init_task( A+i );
  }

  // Now compute
  for( int i = 0; i < nb*nb; ++i )
  {
      compute_task( A+i );
  }

  #pragma omp taskwait
```

```
  return 0;
}
```

5.4.2 Using programmer hints

This approach provides a more direct control on where to run tasks. Data copies are not required although it has the same first touch policy requirement, and your data must be also initialised by an OmpSs task.

For instance, you probably have initialisation tasks and want to spread your data over all the NUMA nodes. You must use the function `nanos_current_socket` to specify in which node the following task should be executed. Example:

```
#pragma omp task out( [N][N]block )
void init_task( float * block )
{
  // First touch NUMA policy will assign pages in the node of the current
  // thread when they are written for the first time.
}

int main(int argc, char* argv[])
{
  int numa_nodes;
  nanos_get_num_sockets( &numa_nodes );

  // Allocate matrix A

  // Call init tasks
  for( int i = 0; i < nb*nb; ++i )
  {
      // Round-robin assignment
      nanos_current_socket( i % numa_nodes );
      init_task( A+i );
  }

#pragma omp taskwait
  return 0;
}
```

Now you have the data where you want it to be. Your computation tasks must be also told where to run, to minimise access to out of node memory. You can do this the same way you do for init tasks:

```
int main( int argc, char *argv[] )
{
  // Allocation and initialisation goes above this
  for( /* loop here */ )
  {
      // Once again you must set the socket before calling the task
      nanos_current_socket( i % numa_nodes );
      compute_task( A+i );
  }
#pragma omp taskwait
}
```

5.4.3 Nesting

If you want to use nested tasks, you don't need to (and you should not) call nanos_current_socket() when creating the child tasks. Tasks created by another one will be run in the same NUMA node as their parent. For instance, let's say that compute_task() is indeed nested:

```
#pragma omp task inout( [N][N] block )
void compute_task( float* block )
{
  for( /* loop here */ )
  {
      small_calculation( block[i] );
  }
#pragma omp taskwait
}

#pragma omp task inout( [N]data )
void small_calculation( float* data )
{
  // Do something
}
```

In this case the scheduling policy will run small_calculation in the same NUMA node as the parent compute_task.

5.4.4 Other

Deepening a bit into the internals, this scheduling policy has task queues depending on the number of nodes. Threads are expected to work only on tasks that were assigned to the NUMA node they belong to, but as there might be imbalance, we have implemented work stealing.

When using stealing (enabled by default), you must consider that:

- Threads will only steal child tasks. If you have an application without nested tasks, you must change a parameter to steal from first level tasks (–socket-steal-parents).
- Threads will steal only from adjacent nodes. If you have 4 NUMA nodes in your system, a thread in NUMA node #0 will only steal from nodes 1 and 2 if the distance to 1 and 2 is (for instance) 22, and the distance to node 3 is 23. This requires your system to have a valid distance matrix (you can use numactl –hardware to check it).
- In order to prevent stealing from the same node, it will be performed in round robin. In the above example, the first time a thread in node 0 will steal from node 1, and the next time it will use node 2.

You can get the full list of options of the NUMA scheduling policy here Or using `nanox --help`.

5.5 My application crashes. What can I do?

5.5.1 Does it crash at the beginning or at the end?

Check if Nanos++ was built with the **allocator disabled**.

5.5.2 Get a backtrace

The first step would be to obtain a **backtrace** using `gdb`. (see GNU GDB website).

Ways to run gdb

A typical approach could be:

```
$ gdb --args program parameters
```

If your applications needs some `NX_ARGS` you can pass them like this:

```
$ NX_ARGS="..." gdb --args program parameters
```

or set in the environment using `export` or `setenv` as usual before running `gdb`.

If you are running in a batch queue (not interactively) you will have to use `gdb` in batch mode. Pass the flags `--batch` and `-ex` like this:

```
$ NX_ARGS="..." gdb --batch -ex "run" -ex "some-gdb-command" program parameters
```

You may pass more than one `-ex` command and they will be run sequentially.

Instead of `-ex` you may write the commands in a file and pass the parameter `--command`. Check the GNU GDB documentation.

Backtrace

Once the program crashes inside `gdb` it is time to get a backtrace. If your program has more than one thread the usual `backtrace` (or its short form `bt`) will not give all the information we want. Make sure you use `thread apply all backtrace` command (or its short form `thread apply all bt`).

If you are running in batch, your command will look as follows:

```
$ NX_ARGS="..." gdb --batch -ex "run" -ex "thread apply all bt" program parameters
```

5.5.3 I cannot get a backtrace

If there is no backtrace, this can be a symptom of a **low stack size**. You can increase it with *NX_STACK_SIZE* (see *Running OmpSs Programs*).

If this does not solve your problem, we suggest to follow these steps:

- Comment all tasks.
- Run with only one thread.
- Activate the tasks again, but follow them by a `#pragma omp taskwait`
- Remove the unneeded `taskwait` one by one.
- Increase the number of threads.

Chances are that you will be able to diagnose your problem after one of these steps.

5.6 I am trying to use regions, but tasks are serialised and I think they should not

The two available region plugins have an important limitation: you have to **align memory** to a power of 2. This is an optimisation trade-off (that might be lifted in the distant future).

If you have two matrices of size 1024×1024 (32-bit floats), and you want to work with blocks of 128×128 elements, you must align to $sizeof(float[1024][1024])$. The block size does not affect us in this case.

You must carefully choose a way to align memory, depending on your system. We have found problems with static alignment, while posix_memalign usually works for us.

To see if there is, effectively, aliasing, you may want to run you program with the **graph** instrumentation plugin (*Task Dependency Graph*). If you see lots of anti/output dependencies, you are likely not aligning your memory properly.

Alignment rules: For every dimension, you must ensure that a rule is true:

$$
\begin{aligned}
r(a,i) =& r(a \mod \left(\prod_{j=1}^{i-1} d_j \times b_i\right), i-1) \\
r(a,1) =& a \mod b_1 \\
\forall_{i=1}^{n} r(a,i) =& 0
\end{aligned}
$$

Where a is an address, b_i is the block size **in bytes** in dimension i, d_j the size **in number of elements** in dimension j.

For the following example of 2 dimensions:

- The address of a block module the block width must be equal to 0.
- The address of a block module the total width multiplied by the block height must be also 0.

In some cases, if your application uses overlapping blocks with halos, you'll need to make sure the start position of each block is aligned. Consider the next matrix (6x6):

0	0	0	0	0	0
0	1	1	2	2	0
0	1	1	2	2	0
0	3	3	4	4	0
0	3	3	4	4	0
0	0	0	0	0	0

There are 4 blocks (1, 2, 3 and 4) of 4 elements each. If the application has a halo as an input, and the block as an output dependency, it would look like this for the first block:

X	I	I	X	X	X
I	O	O	I	X	X
I	O	O	I	X	X
X	I	I	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X

X, I, O denotes no dependency, input and output, respectively.

A first approach would be aligning to the next power of 2 of $6 \times 6 \times sizeof(float)$, but it is not aligned to a power of 2.

Besides that, the blocks are not properly aligned. The first block (1,1) would be in address $7 \times sizeof(float) = 28$ (0x1C). $28 \mod (2 \times sizeof(float))$ is 4, not a power of 2.

To solve this, you must pad the data. In this case, we need to allocate an 8x8 matrix and pad data the following way (although we only need 6 rows to comply with the alignment rules):

P	P	P	P	P	P	P	P
P	0	0	0	0	0	0	P
P	0	1	1	2	2	0	P
P	0	1	1	2	2	0	P
P	0	3	3	4	4	0	P
P	0	3	3	4	4	0	P
P	0	0	0	0	0	0	P
P	P	P	P	P	P	P	P

P is the padded data. It does not need to be initialised (thus some those pages will not be used).

Then, for the first block the dependencies will be:

X	X	X	X	X	X	X	X
X	X	I	I	X	X	X	X
X	I	O	O	I	X	X	X
X	I	O	O	I	X	X	X
X	X	I	I	X	X	X	X
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X

The address of the output block (2,2) is $(8 \times 2 + 2) \times sizeof(float) = 72$. Its mod is 0, so the rule is satisfied for the first dimension.

For the second rule, $72 \mod (8 \times (2 \times sizeof(float))) = 8$. Now we apply the first rule for 8 as the address, and its module is 0. It will now work.

5.7 My application does not run as fast as I think it could

There are a few runtime options you can tweak in order to get better results.

- Compile in performance mode (do not use *–instrument* or *–debug*).
- Adjust your program parameters (such as problem size, block size, etc.). Watch out for too fine grain tasks.
- Disable thread sleep, or reduce/increase sleep time.
- Reduce or increase the spin time. If you increase it, this might put more stress on the tasks queue(s).

5.8 How to run OmpSs on Blue Gene/Q?

5.8.1 Installation

Note: If you want IBM compilers support, make sure that the full-name drivers (e.g. powerpc64-bgq-linux-xlc) are in the PATH. You can usually do so by setting the PATH variable before the configuration step:

```
$ export PATH=/bgsys/drivers/ppcfloor/gnu-linux/powerpc64-bgq-linux/bin/:$PATH
```

You don't need to keep this change after the configuration step.

Follow the instructions described here: *Installation of Nanos++* and *Installation of Mercurium C/C++/Fortran source-to-source compiler*, and add these options for each configuration:

- For Nanos++:

```
$ ./configure --host=powerpc64-bgq-linux ...
```

- For Mercurium:

```
$ ./configure --target=powerpc64-bgq-linux ...
```

Finally compile and install as usual.

5.8.2 Compiling your application

Mercurium will install the drivers using the target keyword as the prefix, so to build your application:

```
$ powerpc64-bgq-linux-mcc --ompss test.c -o test
```

For MPI applications, you can use the MPI wrapper as long as the MPI implementation allows to set the native compiler. For example, if using MPICH:

```
$ export MPICH_CC="powerpc64-bgq-linux-xlmcc --ompss"
$ mpicc test.c -o test.c
```

5.8.3 Instrumenting

For non-MPI applications, you can follow the same instructions described here: *Extrae*

For MPI applications, you should follow the same steps but preloading the Extrae MPI library so that it can intercept the MPI calls. Blue Gene/Q does not allow this method so you will have to link your application with Extrae. You will find an example of how to do it in the Extrae installation directory for BG/Q '$EXTRAE_HOME/share/example/MPI/'.

5.9 Why macros do not work in a #pragma?

OpenMP compilers usually expand macros inside `#pragma omp`, so why macros do not work in Mercurium?

The reason is that when Mercurium compiles an OmpSs/OpenMP program we do not enable OpenMP in the native compiler to reduce the chance of unwanted interferences. This implies that the `#pragma omp` is not known to the native preprocessor, which leaves the `#pragma omp` as is. As a consequence, macros are not expanded and they become unusable inside pragmas.

5.9.1 Alternatives

Although macros are not expanded in clauses, most of the cases they are used for integer constants:

```
#define N 100

#pragma omp task out([N]a)
void f(int *a);
```

```
void g()
{
  int a[N];

  f(a);
#pragma omp taskwait
}
```

The example shown above will not work but fortunately in these cases one can avoid macros. A first approach replaces the macro with an enumerator. This is the most portable approach:

```
enum { N = 100 };
// Rest of the code
```

Another approach uses an `const` integer:

```
const int N = 100;
// Rest of the code
```

Note: In C a `const` qualified variable is not a constant expression whereas in C++ it may be (more information). Mercurium, though, allows this case as a constant expression in both languages.

In case you need to change the value at compile time, you can use the following strategy. Here assume that `N_` is a macro that you change through the invocation of the compiler (typically with a compiler flag like `-DN_=100`):

```
enum { N = N_ };
// Rest of the code
```

Another feasible scenario where macros and pragmas are combined is when we want to generate the pragma itself via macros. To do that, we could use the `_Pragma` operator introduced in `C99` and `C++11`:

```
#define STRINGIFY(X) #X
#define MY_PRAGMA(X) _Pragma(STRINGIFY(X))
#define MY_OPENMP_PARALLEL_LOOP omp parallel for

int main() {
    MY_PRAGMA(MY_OPENMP_PARALLEL_LOOP)
    for(int i = 0; i < 10; ++i) {
        // ...
    }
}
```

5.9.2 Why rely on the native preprocessor, then?

We could provide our own preprocessor, but this adds a lot of complication just to support this case (and as shown in *Alternatives* the typical use case can be worked around).

Complications include, but do not limit to:

- it is not trivial to know where are the system headers. While `gcc` provides good support to know which paths it is using, not all compiler vendors provide this information. Even worse, some parameters may change the used paths between invocations of the native compiler

- subtleties among preprocessor implementations. Not all preprocessors behave the same way and sometimes these differences are relevant for the native compiler headers

5.10 How to track dependences for a given task using paraver

Once we have a Paraver trace we can track task execution by filtering the proper events or using the appropriate Paraver configuration file. Nanos also include an script that can dynamically generate a Paraver configuration file filtering a task (using its task id), the dependences it has during the execution and where the task was generated.

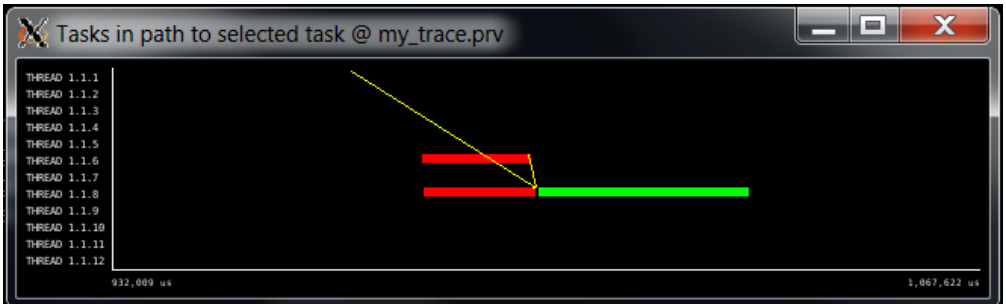
Before using the command you should have the trace loaded in Paraver. You will also need to load the configuration file displaying task numbers and find out the task number of the task you are interested in. This command will help you identify the incoming dependences of the selected task(s).

Usage: track_deps.sh trace.prv list_of_task_numbers

5.10.1 Functionality

The script generates and loads a configuration file that shows only the specified tasks, the tasks on which they depend and the dependences between them. It writes to standard output the numbers of the task so that the command can be used iteratively to explore the full dependency chain.

The following figure shows the results of executing the track_deps.sh script over a Cholesky kernel Paraver trace. We have asked to show the dependences of a task (in the example task id = 25). We can see where the task was created and which tasks it depends on.



- genindex

INDEX

R

S

T

V

W