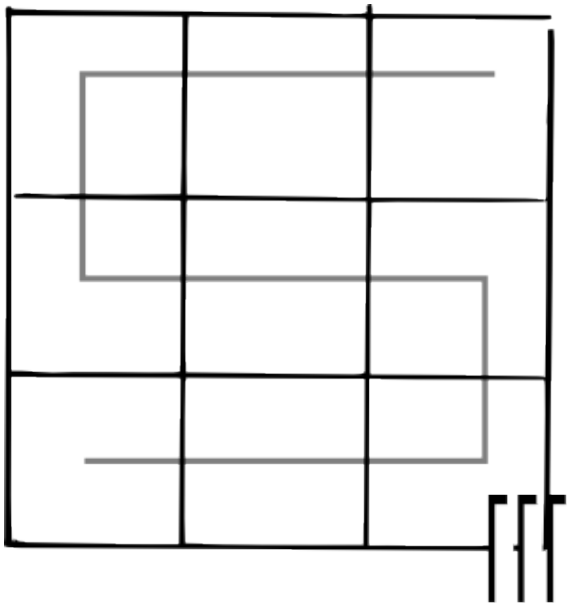


Peano Cookbook

www.peano-framework.org

Dr. rer. nat. Tobias Weinzierl

October 8, 2015

1 Preamble

Peano is an open source C++ solver framework. It is based upon the fact that spacetrees, a generalisation of the classical octree concept, yield a cascade of adaptive Cartesian grids. Consequently, any spacetree traversal is equivalent to an element-wise traversal of the hierarchy of the adaptive Cartesian grids. The software Peano realises such a grid traversal and storage algorithm, and it provides hook-in points for applications performing per-element, per-vertex, and so forth operations on the grid. It also provides interfaces for dynamic load balancing, sophisticated geometry representations, and other features. Some properties are enlisted below.

Peano is currently available in its third generation. The development of the original set of Peano codes started around 2002. 2005-2009, we merged these codes into one Peano kernel (2nd generation). In 2009, I started a complete reimplementation of the kernel with special emphasis on reusability, application-independent design and the support for rapid prototyping. This third generation of the code is subject of the present cookbook.

Dependencies and prerequisites

Peano is plain C++ code and depends only on MPI and Intel's TBB or OpenMP if you want to run it with distributed or shared memory support. There are no further dependencies or libraries required. C++ 11 is used. GCC 4.2 and Intel 12 should be sufficient to follow all examples presented in this document. If you intend to use Peano, we provide a small Java tool to facilitate rapid prototyping and to get rid of writing glue code. This Peano Development Toolkit (PDT) is pure Java and uses DaStGen. While we provide the PDT's sources, there's also a jar file available that comprises all required Java libraries and runs stand alone. To be able to use DaStGen—we use this tool frequently throughout the cookbook—you need a recently new Java version.

We recommend to use Peano in combination with

- Paraview (`www.paraview.org`) or VisIt (`https://wci.llnl.gov/simulation/computer-codes/visit/`) as our default toolboxes create vtk files.
- make and `awk`.

But these software tools are not mandatory.

The whole cookbook assumes that you use a Linux system. It all should work on Windows and Mac as well, but we haven't tested it in detail.

Who should read this document

This cookbook is written similar to a tutorial in a hands-on style. Therefore, it also contains lots of source code snippets. If you read through a chapter, you should immediately be able to re-program the presented details in your code and use the ideas.

Therefore, this cookbook is written for people that have a decent programming background as well as scientific computing knowledge. Some background in the particular application area's algorithms for some chapters also is required. If you read about the particle handling in Peano, e.g., the text requires you to know at least some basics such as linked-cell methods. The text does not discuss mathematical, numerical or algorithmic background. It is a cookbook after all.

What is contained in this document

This book covers a variety of problems I have tackled with Peano when I wrote scientific papers. There is no overall read thread through the document. I recommend to start reading some chapters and then jump into chapters that are of particular interest. Whenever something comes to my mind that should be added, I will add it. If you feel something is urgently missing and deserves a chapter or things remain unclear, please write me an email and I'll see whether I can provide some additional text or extend the cookbook.

October 8, 2015
Tobias Weinzierl

Contents

1 Preamble **i**

2 Quickstart **1**
2.1 Download and install . . . 1
2.1.1 Download the archives from the website . . . 1
2.1.2 Access the repository directly . . . 2
2.1.3 Prepare your own project . . . 2
2.2 Create an empty Peano project . . . 3
2.3 A first spacetree code . . . 3
2.4 Some real AMR . . . 4
2.5 A tree within the spacetree . . . 6

3 Basic Programming Course **9**
3.1 Grid creation . . . 9
3.1.1 On the power of loosing control . . . 10
3.1.2 What happens . . . 13
3.1.3 Multiscale data . . . 14
3.2 Logging, statistics, assertions . . . 16
3.2.1 The user interface . . . 16
3.2.2 Repository fields . . . 16
3.2.3 Log filter . . . 16
3.2.4 Using logging and tracing . . . 16
3.2.5 Statistics . . . 16
3.2.6 Assertions . . . 16
3.3 DaStGen primer and the heap . . . 16
3.4 Filling the element-wise traversal with life . . . 16

4 Applications **17**
4.1 A patch-based heat equation solver . . . 17
4.1.1 Preparation . . . 17
4.1.2 Setting up the patches . . . 20

5 High Performance Computing **21**
5.1 MPI . . . 21

6 Tuning **23**
6.1 Performance analysis . . . 23
6.2 Reducing the MPI grid setup and initial load balancing overhead . . . 24

6.3 MPI quick tuning 28
6.3.1 Filter out log statements 28
6.3.2 Switch off load balancing 28
6.4 Reduce MPI Synchronisation 29
6.4.1 The smell 29
6.4.2 Weaken synchronisation with global master 29
6.4.3 Postpone master-worker and worker-master data exchange . . 30
6.4.4 Skip worker-master data transfer locally/sporadically 30
6.5 Other ideas 30

2 Quickstart

Time: Should take you around 15 minutes to get the code up and running. Then another 15 minutes to have the first static adaptive Cartesian grid.
Required: No previous knowledge, but some experience with the Linux command line and Paraview is advantageous.

2.1 Download and install

To start work with Peano, you need at least two things.

1. The Peano source code. Today, the source code consists of two important directories. The `peano` directory holds the actual Peano code. An additional `tarch` directory holds Peano's technical architecture.
2. The Peano Development Toolkit (PDT). The PDT is a small Java archive. It takes away the cumbersome work to write lots of glue code, i.e. empty interface implementations, default routines, ..., so we use it quite frequently.

For advanced features, you might want to use some **toolboxes**. A toolbox in Peano is a small collection of files that you store in a directory and adopt all pathes accordingly. From a user's point of view, when we use the term toolbox we actually mean this directory with all its content.

> **Remark:** Originally, we hoped that Peano's technical architecture (`tarch`) might become of value for several projects, i.e. projects appreciate that they do not have to re-develop things such as logging, writing of output files, writing support for OpenMP and TBB, and so forth. To the best of our knowledge, the tarch however is not really used by someone else, so we cannot really claim that it is independent of Peano. Nevertheless, we try to keep it separate and not to add anyting AMR or grid-specific to the tarch.

There are two ways to get hold of Peano's sources and tools. You either *download the archives from the website* or you *access the repository directly*. Both variants are fine. We recommend to access the respository directly.

2.1.1 Download the archives from the website

If you don't want to download Peano's whole archive, change to Peano's webpage `http://www.peano-framework.org` and grab the files

- `peano.tar.gz` and
- `pdt.jar`

from there. If you do so, please skip the first two lines from the script before. Otherwise, load down the important files with `wget`. Independent of which variant you follow, please unpack the `peano.tar.gz` archive. It holds all required C++ sources.

```
> wget http://sourceforge.net/projects/peano/files/peano.tar.gz
> wget http://sourceforge.net/projects/peano/files/pdt.jar
> tar -xzvf peano.tar.gz
```

There's a couple of helper files that we use IN the cookbook. They are not necessarily required for each Peano project, but for our examples here they are very useful. So, please create an additional directory `usrtemplates` and grap these files

```
> mkdir usrtemplates
> cd usrtemplates
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTKMultilevelGridVisualiserImplementation.template
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTKMultilevelGridVisualiserHeader.template
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTKGridVisualiserImplementation.template
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTKGridVisualiserHeader.template
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTK2dTreeVisualiserImplementation.template
> wget http://sourceforge.net/projects/peano/files/ \
  usrtemplates/VTK2dTreeVisualiserHeader.template
```

2.1.2 Access the repository directly

Instead of a manual download, you might also decide to download a copy of the whole Peano repository. This also has the advantage that you can do a simple `svn update` anytime later throughout your development to immediately obtain all kernel modifications.

```
> svn checkout http://svn.code.sf.net/p/peano/code/trunk peano
```

Your directory structure will be slightly different than in the example above, but this way you can be sure you grabbed everything that has been released for Peano through the webpage ever.

The archive `pdt.jar` will be contained in `pdt`, while the two source folders will be held by `src`. The directory `usrtemplates` is contained in `pdt`.

2.1.3 Prepare your own project

From hereon, we recommend that you do not make any changes within Peano repositories but use your own directory `peano-projects` for your own projects. We refer to one of these projects generically from hereon as `myproject`. Within `peano-projects`, we will need to access the directories `peano` and `tarch`. It is most convenient to create symbolic links to these files. Alternatively, you also might want to copy files around or adopt makefiles, scripts, and so forth. I'm too lazy to do so and rely on OS links.

```
> mkdir peano-projects
> cd peano-projects
> ln -s <mypath>/peano peano
> ln -s <mypath>/tarch tarch
> ls
```

```
  peano tarch
```

2.2 Create an empty Peano project

Peano projects require four files from the very beginning:

- A **specification** file is kind of the central point of contact. It defines which data models are used and which operations (algorithmic phases) do exist in your project. And it also specifies the project name, namespace, and so forth.
- A **vertex definition** file specifies which data is assigned to vertices in your grid.
- A **cell definition** file specifies which data is assigned to cells in your grid.
- A **state definition** file specifies which data is held in your solver globally.

We will use these files and modify them all the time. For our first step, they are basically empty. As mentioned before, we suggest to have one directory per project. Rather than creating the files as well as the directory manually, we can use the PDT for this:

```
> java -jar <mypath>/pdt.jar --create-project myproject myproject
> ls
  myproject peano tarch
```

If you are interested in the semantics of the magic arguments, call jar file without any argument and you will obtain a brief description. A quick check shows that the aforementioned four files now have been created:

```
> ls -al myproject
drwxr-xr-x 2 ... .
drwxr-xr-x 5 ... ..
-rw-r--r-- 1 ... Cell.def
-rw-r--r-- 1 ... project.peano-specification
-rw-r--r-- 1 ... State.def
-rw-r--r-- 1 ... Vertex.def
```

The PDT typically is used only once with the `--create-project` argument. From hereon, it serves different purposes. That is ...

2.3 A first spacetree code

... it helps us to write all the type of code parts that we don't want to write: **glue code** that does nothing besides gluing the different parts of Peano together.

We postpone a discussion of the content of the generated files to Chapter 3 and continue to run a first AMR example. For this, we call the PDT again. However, this time, we use the generated specification file as input and tell the tool to create all glue code.

```
> java -jar <mypath>/pdt.jar --generate-gluecode \
  myproject/project.peano-specification myproject \
  <mypath>/usrtemplates
```

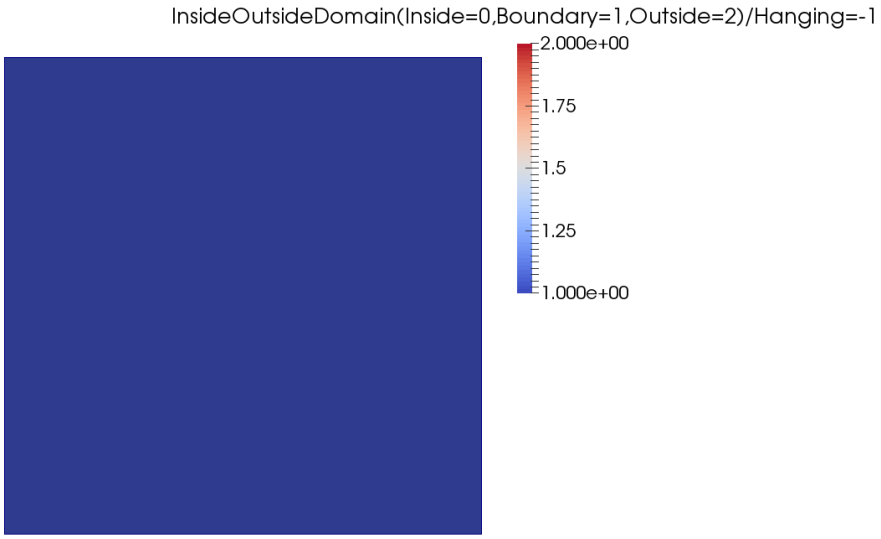
By default, the autogenerated, (almost) empty four files require the `usrtemplates`. We reiterate that many projects later won't need them. If we again study the content of our directory, we see that lots of files have been generated. For the time being, the `makefile` is subject of our interest. Depending on your compiler, you should be able to call `make` straight away. If it doesn't work, open your favourite text editor and adopt the makefile accordingly.

```
> ls myproject
  adapters Cell.cpp Cell.def
  Cell.h dastgen main.cpp
  makefile mappings project.peano-specification
  records repositories runners
  State.cpp State.def State.h
  tests Vertex.cpp Vertex.def
  Vertex.h VertexOperations.cpp VertexOperations.h
> make -f myproject/makefile
> ls
  files.mk myproject peano peano-YourProjectName-debug tarch
```

There it is: the first Peano executable. We can run it straight away:

```
> ./peano-YourProjectName-debug
> ls
  files.mk grid-0.vtk myproject peano
  peano-YourProjectName-debug tarch
```

We see that it has produced a vtk file. So it is time to startup Paraview or VisIt and see what is inside.



Congratulations: We have created the simplest adaptive Cartesian grid in 2d that does exist. A single square!

2.4 Some real AMR

We now set up something slightly more complicated. First of all, we switch to a 3d setup rather than 2d. For this, open the makefile (`myproject/makefile`) and alter the content of the `DIM` variable.

```
# Set Dimension
# -------------
#DIM=-DDim2
```

```
DIM=-DDim3
#DIM=-DDim4
```

If you clean your project (`make -f myproject/makefile clean`) and rebuild your code, you see that the individual files are translated with the compile switch

```
g++ ... -DDim3 ....
```

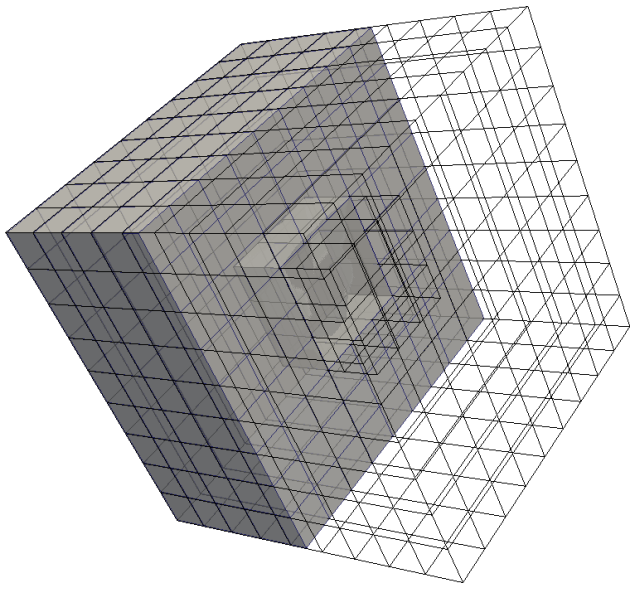
Indeed, this is all that's required for Peano to run a 3d experiment rather than a 2d setup.

> **Remark:** We do support currently up to 10-dimensional setups. If you require higher dimensions, you might even be able to extend Peano accordingly by changing solely the file `peano/utils/Dimensions.h`. But have fun with your memory requirements exploding.

Next, we will edit the file `myproject/mappings/CreateGrid.cpp`. Open it with your favourite text editor and search for the operation `createBoundaryVertex`. Change it into the code below:

```
void myproject::mappings::CreateGrid::createBoundaryVertex(
 myproject::Vertex& fineGridVertex,
 const tarch::la::Vector<DIMENSIONS,double>& fineGridX,
 const tarch::la::Vector<DIMENSIONS,double>& fineGridH,
 myproject::Vertex *const coarseGridVertices,
 const peano::grid::VertexEnumerator& coarseGridVerticesEnumerator,
 myproject::Cell& coarseGridCell,
 const tarch::la::Vector<DIMENSIONS,int>& fineGridPositionOfVertex
) {
  logTraceInWith6Arguments("createBoundaryVertex(...)", ...);
    // leave this first line as it is

  if (coarseGridVerticesEnumerator.getLevel()<2) {
    fineGridVertex.refine();
  }

  logTraceOutWith1Argument("createBoundaryVertex(...)",fineGridVertex);
}
```

If you compile this code and run the executable, you will (besides lots of debug output) obtain a way bigger vtk file. If you visualise it this time, we observe that the code refines towards the cube's boundary. You may want to play around with magic `2` in the operation above. Or you might want to continue to our final example.



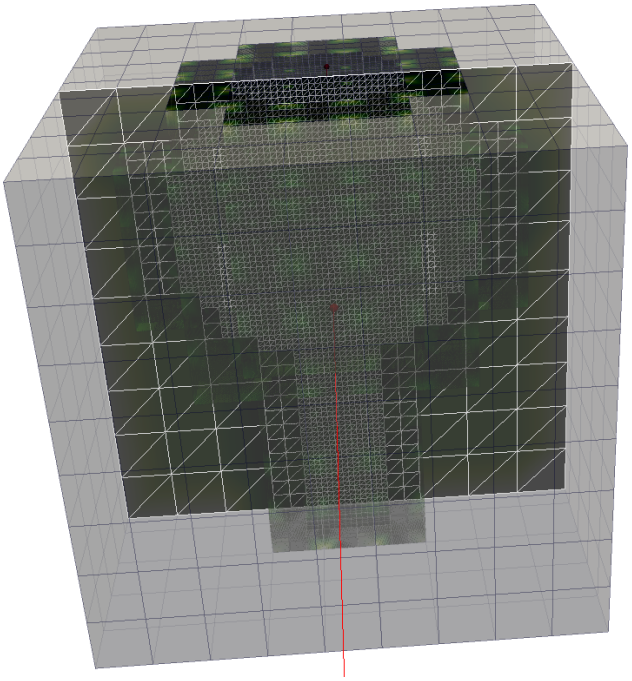
2.5 A tree within the spacetree

In the final example we create a slightly more interesting setup. We solely edit the operation `createInnerVertex` within the file `myproject/mappings/CreateGrid.cpp`, recompile it and have a look at the result. When you study source code, please note the similarity to Matlab when we work with vectors in Peano; as well as that the indices start with 0. If your want to get rid of all the debug statements and are sick of long waiting times, remove the `-DDebug` statement in the line `PROJECT_CFLAGS = -DDebug -DAsserts` within the makefile. There are more elegant ways to filter out log statements that we will discuss later.

```
void myproject::mappings::CreateGrid::createInnerVertex(
 myproject::Vertex& fineGridVertex,
 const tarch::la::Vector<DIMENSIONS,double>& fineGridX,
 const tarch::la::Vector<DIMENSIONS,double>& fineGridH,
 myproject::Vertex *const coarseGridVertices,
 const peano::grid::VertexEnumerator& coarseGridVerticesEnumerator,
 myproject::Cell& coarseGridCell,
 const tarch::la::Vector<DIMENSIONS,int>& fineGridPositionOfVertex
) {
 logTraceInWith6Arguments("createInnerVertex(...)",fineGridVertex,...);

 if (
   fineGridVertex.getRefinementControl()==Vertex::Records::Unrefined
   &&
   coarseGridVerticesEnumerator.getLevel()<4
 ) {
   bool trunk = (fineGridX(0)-0.5)*(fineGridX(0)-0.5)
              + (fineGridX(2)-0.5)*(fineGridX(2)-0.5)<0.008;
   bool treeTop = (fineGridX(0)-0.5)*(fineGridX(0)-0.5)
                + (fineGridX(1)-0.7)*(fineGridX(1)-0.7)
                + (fineGridX(2)-0.5)*(fineGridX(2)-0.5)<0.3*0.3;
   if (trunk | treeTop) {
     fineGridVertex.refine();
```

```
    }
  }

  logTraceOutWith1Argument("createInnerVertex(...)",fineGridVertex);
}
```

So here's what I get. Feel free to create better pics:



3 Basic Programming Course

3.1 Grid creation

Time: 15 minutes for the programming but perhaps around 30 minutes for the visualisation.
Required: Chapter 2.

In this section, we study a 2d example. Please adopt your makefile accordingly. Furthermore, we use the files `VTKMultilevelGridVisualiserHeader` and `...Implementation` as well as `VTK2dTreeVisualiser...`. If you have downloaded the whole Peano repository, these files can be found in `pdt/usrtemplates`. If not, you have to download them manually from the webpage. Please set up an empty project as discussed in Chapter 2 and implement one operation as follows (all other operations can remain empty/only filled with log statements):

```
void myproject::mappings::CreateGrid::touchVertexLastTime(
      myproject::Vertex& fineGridVertex,
      const tarch::la::Vector<DIMENSIONS,double>& fineGridX,
      const tarch::la::Vector<DIMENSIONS,double>& fineGridH,
      myproject::Vertex *const coarseGridVertices,
      const peano::grid::VertexEnumerator& coarseGridVerticesEnumerator,
      myproject::Cell& coarseGridCell,
      const tarch::la::Vector<DIMENSIONS,int>& fineGridPositionOfVertex
) {
  logTraceInWith6Arguments( "touchVertexFirstTime(...)", fineGridVertex, fineGridX, ...

  if (
    coarseGridVerticesEnumerator.getLevel()<5
    &&
    tarch::la::equals( fineGridX, 0.0 )
    &&
    fineGridVertex.getRefinementControl()==Vertex::Records::Unrefined
  ) {
    fineGridVertex.refine();
  }

  logTraceOutWith1Argument( "touchVertexFirstTime(...)", fineGridVertex );
}
```

This source fragment requires some additional explanation. We neglect the enumerator stuff for the time being. That will become clear later throughout the present chapter. The refinement control check says 'well, refine, but do it only on unrefined vertices'. It's just a matter of good style, not to call `refine` on a refined vertex. The middle line uses a function from the tarch's linear algebra namespace. It takes the `fineGridX` vector (the position of the vertex in space) and checks whether all entries equal zero. As we are working with floating point numbers, it is not a bit-wise check. Instead, it uses an interval of machine precision around zero. You may want to

change this notion of machine precision in Peano (file `Scalar` within the `tarch::la` namespace). In general, it would be a good idea to study the content of the `la` component soon—there's lots of useful stuff in there to work with tiny, dense vectors[1].

> **Remark:** Peano realises a **vertex-based**, **logical-or** refinement: You can invoke `refine` on any unrefined vertex. Peano then refines all cells around a vertex in the present or next traversal (it basically tries to do it asap, but sometimes data consistency constraints require it to postpone the actual refinement by one iteration). The other way round, you may read it as follows: A cell is refined if the refinement flag is set for any adjacent vertex.

Whenever you use Peano, you have to do three things:

1. Decide which algorithmic phases do exist and in which order they are called. Examples for algorithmic phases could be: set up grid, initialise all variables, refine regions of interest, perform an iterative solve step, plot some data, compute metrics on the solution, ...
2. Model the data, i.e. decide which data is assigned to the vertices and cells of the grid.
3. Implement the different actions on this data model that are used by the algorithmic phases.

This scheme lacks the bullet point 'run through the grid'. Indeed, Peano applications do never run themselves through the spacetree. They specify which set of operations is to be called throughout a run through the grid, i.e. they say what is done on which data. Afterward, they invoke the iteration and leave it to Peano to run through the grid and invoke these operations in the right order on the right ranks using all the cores you have on your machine[2]. This scheme realises something people call 'The Hollywood Principle':Don't call us, we call you!

> **Remark:** The **inversion of control** is the fundamental difference of Peano to other spacetree-based codes offered as a library. And typically it is the property many users first struggle with. Often, people claim 'I have to run through the grid this and that way'. Often, they are wrong. It can become quite comfortable to leave it to someone else to decide how grid traversals are realised. And it allows the grid traversal in turn to optimise the code under the hook without an application developer to bother.

3.1.1 On the power of loosing control

The algorithmic phases, i.e. what can be done on a grid, are specified in the specification file. Open your project's file. There are two different parts of the document that are of interest to us: An *event mapping* is an algorithmic step that you have to implement yourself. In this chapter's example, we want to do two things: create a grid and count all the vertices. Furthermore, we want to plot our grid, but let's keep in mind that Peano has some predefined actions as well. So we augment our mapping set as follows:

```
// Creates the grid
event-mapping:
```

[1] Peano someday should perhaps be rewritten to use boost linear algebra or some fancy template library. Feel free to do so. Right at the moment, it is all plain hand-craftet routines.

[2] This statements requires explanation, and indeed it is not *that* straightforward. But the idea is phrases correctly: the application codes specifies what is to be done and then outsources the scheduling and the responsibility to use a multicore machine to Peano.

```
  name: CreateGrid

// Counts all the vertices within the grid
event-mapping:
  name: CountVertices
```

Event mappings cannot be used directly. Instead, we have to specify adapters. Adapters take the tree traversal and invoke for each grid part a set of events. As we distinguish adapters which basically just glue together (multiple) events from the events themselves, we will be able to do the following later: we write a fancy visualisation routine, a routine that adopts the grid to a new data set and some compute routines. As we have done this in three different event sets, we can then combine these events in various ways: compute something and at the same time plot, compute only, plot and afterward adopt the grid, and so forth. For the time being, we use the following adapters:

```
adapter:
  name: CreateGrid
  merge-with-user-defined-mapping: CreateGrid

adapter:
  name: CountVertices
  merge-with-user-defined-mapping: CountVertices

adapter:
  name: CreateGridAndPlot
  merge-with-user-defined-mapping: CreateGrid
  merge-with-predefined-mapping: VTKGridVisualiser(finegrid)
  merge-with-predefined-mapping: VTKMultilevelGridVisualiser(grid)

adapter:
  name: CountVerticesAndPlot
  merge-with-user-defined-mapping: CountVertices
  merge-with-predefined-mapping: VTKMultilevelGridVisualiser(grid)

adapter:
  name: Plot
  merge-with-predefined-mapping: VTKGridVisualiser(finalgrid)
```

The first two adapters are trivial: They basically delegate to one event set. The next two take one event set each and invoke it. Furthermore, they also use a predefined event set. They will call `CreateGrid` or `CountVertices`, respectively, and at the same time plot. If you create all code with

```
java -jar <mypath>/pdt.jar --generate-gluecode
myproject/project.peano-specification myproject <mypath>/usrtemplates
```

it is the directory `usrtemplate` where the PDT searches for the predefined event sets. The last adapter by the way is a trivil one, too: It invokes only one of the events that ship with Peano.

Next, please create all glue code and have a quick look into the file `runners/Runner.cpp`. This file is the starting point of Peano. The C++ main routine does some setup steps and then creates an instance of the Runner (see the source code yourself if you don't believe). It then invokes `run()` which in turn continues to `runAsMaster` or `runAsWorker()`. The latter will play a role once we use MPI. For the time being, let's focus on the master's routine. Here, we see the following:

```
int myproject::runners::Runner::runAsMaster(myproject::repositories::Repository& repository) {
```

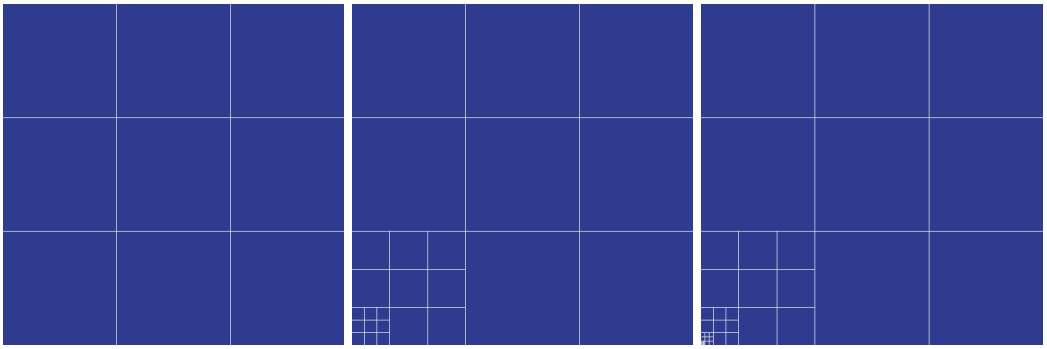
```
  peano::utils::UserInterface userInterface;
  userInterface.writeHeader();

  // @todo Insert your code here

  // Start of dummy implementation

  repository.switchToCreateGrid(); repository.iterate();
  repository.switchToCountVertices(); repository.iterate();
  repository.switchToCreateGridAndPlot(); repository.iterate();
  repository.switchToCountVerticesAndPlot(); repository.iterate();
  repository.switchToPlot(); repository.iterate();



  repository.logIterationStatistics();
  repository.terminate();
  // End of dummy implementation

  return 0;
}
```

The PDT cannot know what exactly we do, so it basically runs all the adapters we have specified. This is the place where we implement our overall algorithm, i.e. the big picture. Lets change it as follows:

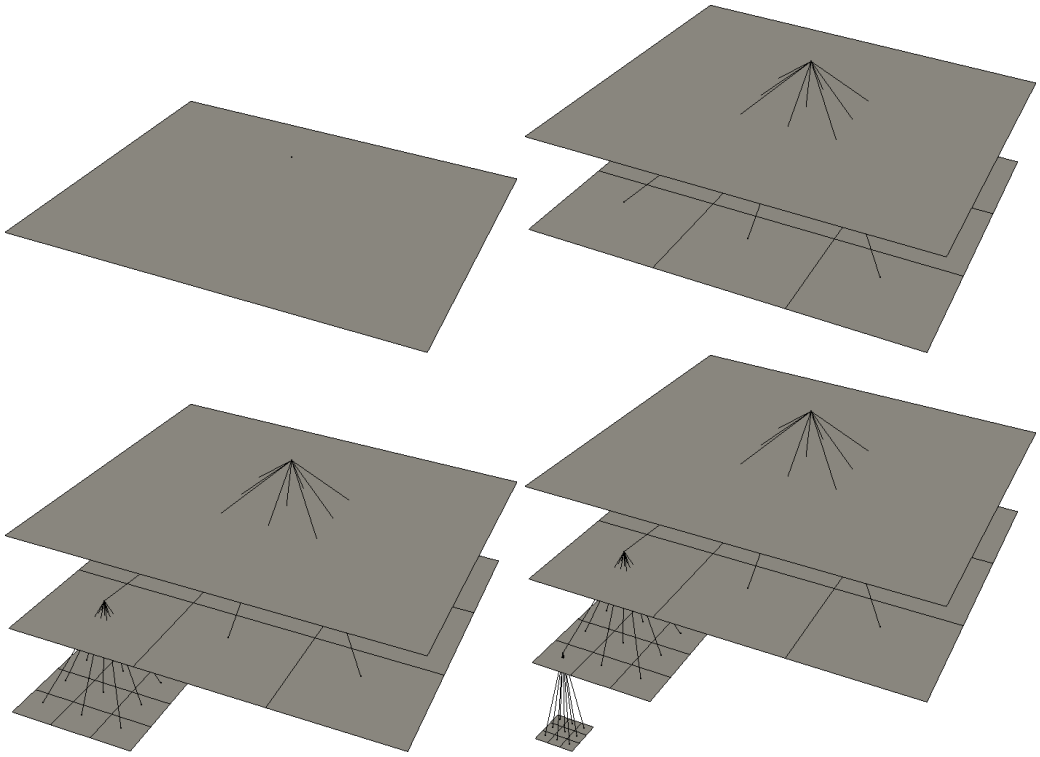
```
  peano::utils::UserInterface userInterface;
  userInterface.writeHeader();

  repository.switchToCreateGridAndPlot();
  for (int i=0; i<10; i++) repository.iterate();
  repository.switchToCountVertices(); repository.iterate();

  repository.logIterationStatistics();
  repository.terminate();

  return 0;
```

This algorithm says that we want to create a grid and at the same time plot it. We want to do this ten times in a row. Afterward, we switch to our vertex counting and want to run through the grid once more. This time, nothing shall be plotted. We just want to know how many vertices there are.

We really do not care about how the code runs through the grid. We also do not really care how all vertices, cells, whatever are processed. We say what is done in which order from a bird eye's perspective. If you compile the code and run it now, you should end up with a sequence of vtk files. You might want to make a video (take the files that are called `finegrid-`something). Below are some screenshots:



3.1.2 What happens



To understand what is happening, we can read all the outputs written to the terminal by Peano (see the next chapter how to remove them/filter them). Or we can just give a quick sketch:

1. We first create a repository. A repository is basically the spacetree, i.e. the grid, plus all the adapters we've defined. It also provides some statistics the can read out.

2. The code runs into the Runner's `runAsMaster()` routine and selects which adapter to use. This is the switch statement. In a parallel code, all involved ranks would now immediately active this adapter.

3. In the runner, we then call the `iterate` operation on the repository. The repository now starts to run through the whole grid. Actually, it runs through the spacetree in a kind of top-down way.

4. Whenever it encounters an interesting situation (it loads a vertex for the very first time, e.g.), it triggers an event. Event means that it calls an operation on the adapter.

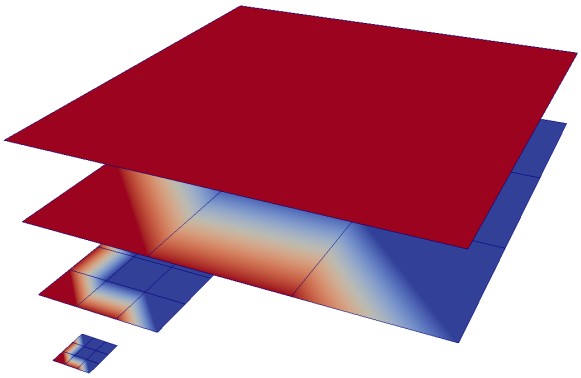
5. An adapter may fuse multiple events; both predefined and user-defined. Per event, it calls all these mappings' implementations one after another. The order is the same you have used in your specification file.

> **Remark:** This document does not run through the list of available events, in which order they are called and so forth. You may want to have a look into any event's header. The PDT augments the header with a quite verbose explanation what event is called when.

It might now be the right time to look into one of these mapping headers to get a first impression what is available. Basically, they describe all important plug-in points that you might want to use in any element-wise multiscale grid traversal.

3.1.3 Multiscale data

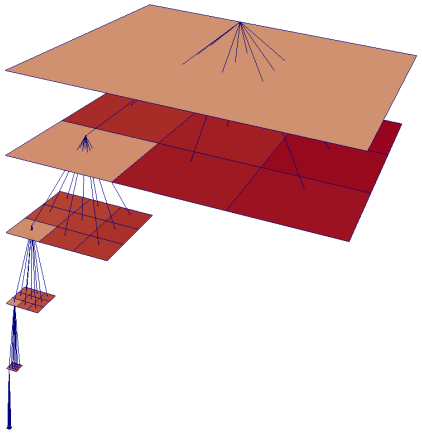
Prior to a discussion of Peano's data model, it might make sense to load all the files `grid-level-?-0.vtk` into your favourite visualisation software. Dilate them according to their level encoded in the name. In Paraview, you have to select each file, apply Filters/Alphabetical/Transform, and dilate each file by -0.2 along the z-axis per level. You might end up with something alike:



Again, feel free to create a video that shows how additional levels are added in each step. There's a more elegant way to end up with a similar picture. I once created a graph visualiser that is today also available as predefined mapping. Just modify your adapters as follows:

```
adapter:
  name: CreateGridAndPlot
  merge-with-user-defined-mapping: CreateGrid
  merge-with-predefined-mapping: VTK2dTreeVisualiser(tree,getLevel)
```

This time, creating a video should be straightforward.



We see that speaking of Peano in terms of a spacetree software is only one way to go. We also could speak of it of a software managing Cartesian grids embedded into each other. The latter point of view reveals an important fact: Peano handles vertices and cells. Cells are embedded into each other as they form the tree and there is a clear parent-child relation. Vertices connect cells. A vertex is unique due to its position in space plus its level, i.e. there might be some coordinates that host multiple vertices. In math, we would call this generating system. There are two types of vertices: the standard ones are adjacent to 2^d cells on the same level with d being the dimension of the problem. All other vertices are hanging nodes.

Remark: Hanging nodes are not stored persistently. They are created and destroyed in each traversal upon request and never stored in-between two iterations. You can never be sure how often a hanging nodes is constructed per traversal. It could be up to $2^d - 1$ times. You only know it is created once at least.

Cells and vertices hold data. This data is described in the files `Vertex.def` and `Cell.def`. The .def files feed into our tool DaStGen that translated them internally into a standard C++ class (though it does some more stuff: it autogenerates all MPI data types that we need later on for parallel codes, and it in particular compresses the data such that a bool field is really mapped onto a single bit, e.g.). There's an alternative way to hold data on the grid that is called heap: in this case, we do not assign a fixed number of properties to vertices or cells. Both storage variants are discussed in Chapter **??**.

3.2 Logging, statistics, assertions

3.2.1 The user interface

3.2.2 Repository fields

3.2.3 Log filter

3.2.4 Using logging and tracing

3.2.5 Statistics

3.2.6 Assertions

3.3 DaStGen primer and the heap

So what we might do now is the following. Open the `Cell.def` and edit it accordingly. Rerun the PDT, recompile, and work.

```
Packed-Type: short int;


class myproject::dastgen::Cell {
  discard parallelise int myNonPersistentInteger;
  persistent parallelise double myValue;
  persistent parallelise bool myBool1;
  persistent parallelise bool myBool2;
};
```

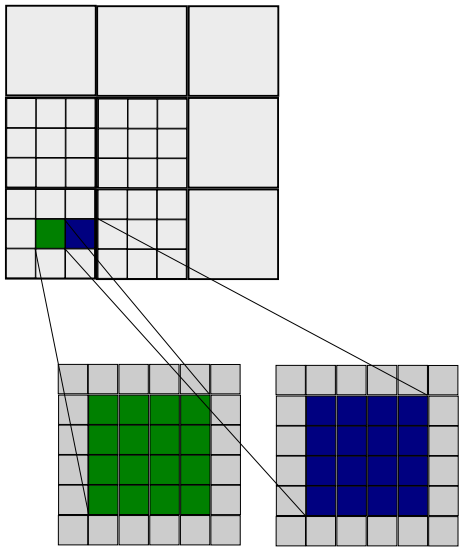
3.4 Filling the element-wise traversal with life

4 Applications

4.1 A patch-based heat equation solver

Time: 60 minutes.
Required: Chapter 2.

In this section, we sketch how to realise a simple explicit heat equation solver that is based upon patches. The idea is that we embed a small regular Cartesian grid into each individual spacetree leaf. This patch is surrounded by a halo/ghost cell layer holding copies from neighbouring patches.



In the sketch, we enbed 4×4 patches into the spacetree cells. Two cells (blue and green) are illustrated. Each patch is surrounded by a halo layer of size one. We will write the code to work within the patches, while the layers around the patches will hold copies of the neighbouring patches and thus couple them.

For this endeavour, we need the toolbox `multiscalelinkedcell` that you can download from the Peano webpage. We assume that the whole toolbox is unzipped into a directory `multiscalelinkedcell` which is held by your source directory.

4.1.1 Preparation

We start with the creation of a file `PatchDescription.def` in your project's root directory. In our example, each patch solely shall hold one array of unknowns u that are associated to the vertices of the patch. So each patch will have exactly $(4+2+1)^2$ unknowns (the four is the size, there's two halo cells along each coordinate axis, and then there's finally one more vertex than there are cells). Besides the unknowns called u, we also store the position and size as well as the level with each patch—well-aware that level and size are kind of redundant.

```
#include "peano/utils/Globals.h"

Packed-Type: int;

Constant: DIMENSIONS;

/**
 *A cell description describes one individual patch of the overall grid, i.e.
 *it holds pointers to the actual data of the patch (arrays) and its meta data
 *such as time stamps. Each unrefined node of the spacetree, i.e. each leaf,
 *holds exactly one instance of this class.
 */
class myprojectname::records::PatchDescription {
  /**
   *Two pointers to float arrays.
   */
  parallelise persistent int u;
  /**
   *I need level and offset to be able to determine the source and image in
   *the adaptive case.
   */
  parallelise persistent int level;
  parallelise persistent double offset[DIMENSIONS];
  parallelise persistent double size[DIMENSIONS];
};
```

Please note that u is modelled as integer. Actually, we do not hold the data directly within the patch description but we make the patch description hold a pointer to the actual data. The data will be managed by Peano on the heap. The heap uses integers as pointers. They are actually hash map indices.

Our system design is as follows:

- Each cell holds a pointer to one `PatchDescription`.
- The `PatchDescription` holds a pointer to the actual patch data and comprises some additional meta data (such as the level).
- Each vertex holds 2^d pointers to the `PatchDescription` instances belonging to the adjacent cells.

Whenever we enter a cell, we can thus take its patch description, and get the actual data from this description. Alternatively, we can use the cell's 2^d adjacent vertices. As they know the adjacent patch descriptions, we can also get the data associated to cell neighbours and thus befil the ghost layers, e.g.

Take the Peano description file of our project ensurte that it contains the following lines:

```
heap-dastgen-file: PatchDescription.def

[...]

vertex:
  dastgen-file: Vertex.def
  read vector2PowD(int): PatchIndex
  write vector2PowD(int): PatchIndex

[...]
```

```
event-mapping:
  name: Mapping1


event-mapping:
  name: Mapping2

[...]

adapter:
  name: Adapter1
  merge-with-user-defined-mapping: Mapping1
  merge-with-predefined-mapping: MultiscaleLinkedCell(PatchIndex)

adapter:
  name: Adapter2
  merge-with-user-defined-mapping: Mapping2
  merge-with-predefined-mapping: MultiscaleLinkedCell(PatchIndex)
```

Managing all the adjaceny data (making each vertex point to the right patch) obviously is a tedious task. The `multiscalelinkedcell` toolbox fortunately does most of the stuff for us, if we augment each adapter with a predefined mapping, tell this mapping what the attribute for the patch handling will be (`PatchIndex`), and augment the vertex accordingly. Finally, open `Vertex.def` and augment it accordingly:

```
#include "peano/utils/Globals.h"

Packed-Type: int;

Constant: TWO_POWER_D;

class myprojectname::dastgen::Vertex {
  /**
   *These guys are pointers to the adjacent cells. Actually, they do not point
   *to the neighbouring cells but to the heap indices associated to these cells.
   *These heap indices reference one or several instances of PatchDescription.
   */
  expose persistent int patchIndex[TWO_POWER_D];

  [...]
};
```

We run the translation process and add the toolbox directory to the PDT call:

```
java -jar <mypath>/pdt.jar <mypath>/project.peano-specification <mypath> \
<mypath>/usrtemplates:<mypath>/multiscalelinkedcell
```

The PDT in collaboration with the toolbox will now create code that makes each vertex track the `patchIndex` value of the adjacent cells. If you change your grid, the indices are updated automatically, as long as you merge `MultiscaleLinkedCell` into your adapters. To make the code compile, you finally have to add a routine

```
  int getPatchIndex() const;
```

to your Cell class. Make the routine return the value of an attribute `persistent int patchIndex` that you add to your `Cell.def`. Set this field to -1 in the default constructor.

4.1.2 Setting up the patches

Before we start any coding, we have to specify which heaps we want to use to administer the patch description objects and the actual u data. One option is to define this centrally in the `Cell.h` file that is generated by the PDT:

```
#include "peano/heap/Heap.h"
#include "<mypath>/records/PatchDescription.h"

namespace myprojectsnamespace {
  class Cell;

  typedef peano::heap::PlainHeap< myprojectsnamespace::records::PatchDescription >
    PatchDescriptionHeap;
  typedef peano::heap::PlainDoubleHeap DataHeap;
}
```

In this setup, we use the plain heap from Peano's heap directory to administer both the data and the patch descriptions. There are several other, more sophisticated, heap implementations available. While they allow you to tune your code for special purposes, the plain heap typically is a good starting point.

To set up the patches, we create plug into the mapping creating our grid. Alternatively, we can first create the grid and then outsource the patch initialisation into an additional mapping. In any case, I strongly encourage you to initialise the heap as a first step. This is however optional:

```
void myprojectsnamespace::mappings::InitPatches::beginIteration(
  ...
) {
  logTraceInWith1Argument( "beginIteration(State)", solverState );

  PatchDescriptionHeap::getInstance().setName( "patch-description-heap" );
  DataHeap::getInstance().setName( "data-heap" );

  logTraceOutWith1Argument( "beginIteration(State)", solverState);
}
```

So far, each cell points to index -1 as patch description index, and each vertex knows that all adjacent cells point to -1. We change this now as we plug into `enterCell` and introduce a new operation in Cell:

```
void myprojectsnamespace::mappings::InitPatches::enterCell(
  ...
) {
  fineGridCell.init( ... ); // please pass through the level, the offset and the size
}


void myprojectsnamespace::Cell::initCellInComputeTree( ... ) {
  const int newPatchIndex = PatchDescriptionHeap::getInstance().createData(1);
  _cellData.setPatchIndex( newPatchIndex );
  assertion( newPatchIndex>=0 );
}
```

5 High Performance Computing

5.1 MPI

6 Tuning

6.1 Performance analysis

Time: Less than 10 minutes unless you postprocess a big file.
Required: You may work with the plain output that Peano writes to the terminal. If you use log filters (cmp. Chapter **??**), it is important that you know how to switch particular logging infos on. You also need a working Python installation.

Prior to any parallelisation or tuning discussion, I want to emphasise that it usually makes sense first of all to have a how Peano is performing from a grid point of view. For this, the framework comes along with a rather useful script.

- Run your code and ensure that `info` outputs from the `peano::performanceanalysis` component are enabled.
- Pipe the output into a file:

```
> ./myExecutable myArguments > outputfile.txt
```

  We call this file `outputfile.txt` from hereon.

- Pass the output file to Peano's performance analysis script written in Python. Besides the script (name), you also have to tell the script how many MPI ranks you have used and how many threads have been enabled. Skip the arguments if you haven't used MPI.

```
> python <mypath>/peano/performanceanalysis/performanceanalysis.py outputfile.txt
```

- Open the web browser of your choice and open the file `outputfile.txt.html`

> **Remark:** Besides the output written by Peano through the component `peano::performanceanalysis`, you also have to use the `CommandLineLogger` (the default), and you have to make this one write out time stamps as well as trace information. If you use your own logger or a modified log format, the Python script will fail.

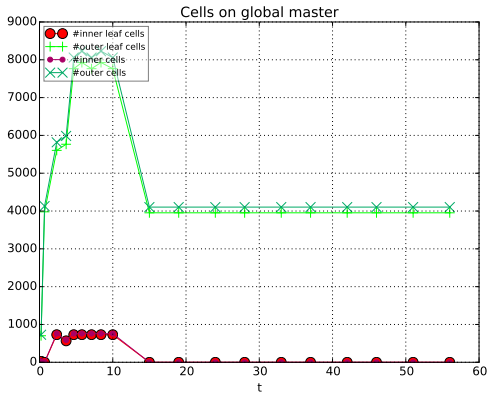
If you browse through your directory, you will notice that all graphs are written both as png and as pdf. You can thus integrate them directly into your LaTeX reports.

6.2 Reducing the MPI grid setup and initial load balancing overhead

Time: Around 30 minutes.
Required: A working MPI code.

In this section, we assume that you've a reasonable load balancing and that you were able to postprocess your performance analysis outputs. We discuss

The smell



If you identify ranks whose local load decreases incrementally, these are ranks that step by step fork more of their work to other ranks. In principle, this is fine and a result of load balancing. For reasonably static setups, it however is irritating: why is there such a long setup phase where obviously solely data is redistributed?

The reason can be found in the semantics of `createVertex` and `touchVertexFirstTime`. Both operations try to refine the grid around the respective vertex immediately. Only if circumstances such as a parallel partitioning running through this vertex—the refinement instruction then first has be distributed to all ranks holding a copy of this vertex—do not allow Peano to realise the refinement immediately, the refinement is postponed to the next iteration. In many parallel codes, all the refinement calls pass through immediately on rank 0 before it can spawn any rank. This leads to the situation that the whole grid is in one sweep built up on the global master and afterwards successively distributed among the ranks.

Such a behaviour is problematic: the global rank might run out of memory, lots of data is transferred, and the sweeps over the whole grid on rank 0 are typically pretty expensive. A distributed grid setup is advantageous.

The solution To facilitate this, it makes sense to switch from an aggressive refinement into an iterative grid refinement strategy (one refinement level per step, e.g.) to allow the rank to deploy work throughout the grid construction and thus build up the grid in parallel and avoid the transfer of whole grid blocks due to rebalancing. Simply move your `refine()` call from the creational or touch first events into `touchVertexLastTime()`: As a consequence, setting up a (rather regular) grid of depth k requires at least k iterations.

To find out when a grid has been constructed and balanced completely, the repository provides an operation. Instead of writing something along the lines

```
repository.switchToSetup();
```

```
repository.iterate();
```

you have to write

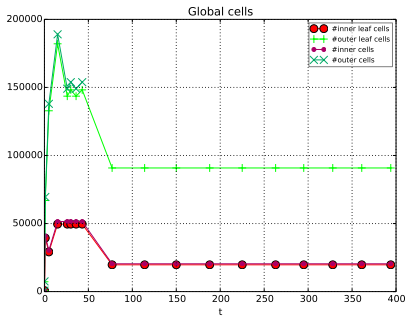
```
repository.switchToSetupExperiment();
do {
  repository.iterate();
} while ( !repository.getState().isGridBalanced() );
```

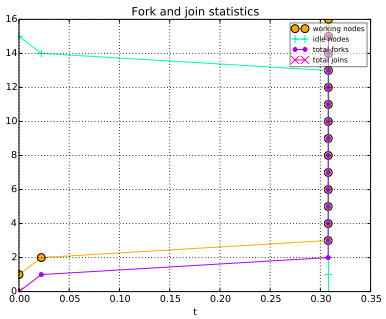
Related pitfalls & ideas As always, the devil is in the details:

- For many load balancing algorithms, it does make sense to create an initial grid of depth $\hat{k} < k$ on your rank 0 before you do any load balancing. This allows the load balancing metric to get a first idea about what the grid will look like and then to switch on load balancing. This can be done with

```
peano::parallel::loadbalancing::Oracle::getInstance().activateLoadBalancing(false);
// set up grid up to a certain level
peano::parallel::loadbalancing::Oracle::getInstance().activateLoadBalancing(true);
```

- Once all ranks have obtained 'their' partition, it does not make sense to continue to build up at most one grid level per sweep. In this case, you have to reliase an inverse pattern compared to the pattern sketched in the first bullet point. Such a situation is easy to spot: it typically materialises in a slow increase of total/global vertices while the fork statistics show that no forks happen anymore. Compare the two plots below:





The grid construction requires about 80s while the last forks are tracked at t=0.3s. Starting from t=0.3s, one could build up the grid one sweep.

> **Remark:** Peano parallel code offers an operation `enforceRefine()` on the vertices that you can use to tackle this problem. Use with care and read through the documentation in code.

- Everytime you rebalance your grid, Peano disables dynamic load balancing for a couple of iterations (three or four). Throughout these iterations, it can recover all adjacency information if the grid itself changes as well. Consequently, it does make sense to add a couple of adapter runs after each grid modification that to not change the grid structure: When you know that you have an adapter that changes the grid, apply afterwards an adapter that does not change the grid for a couple of times. This way, you ensure that no mpi rank runs out of memory. The grid generation does not overtake the rebalancing.

- If you are using the heap data structure, it furthermore makes sense to split up the initialisation into a grid setup and a data struture initialisation. You balance and distribute the grid setup following the recommendations above and then in one additional sweep initialise the heap. You initialise the heap as late as possible and thus avoid unneccesary administrative overhead.

Pattern for static grid setup Most codes at least start form a static grid partitioning and globally know what the initial grid looks like. It then has proven of value to do the following:

1. Determine a certain grid level that should be used to do an initial load balancing. If you have a regular grid, this might be the coarsest grid level that could be deployed among all involved ranks:

```
_coarsestRegularLevelUsedForDD = 0;
int ranksUsedSoFar = 0;
int increment = 1;
while (ranksUsedSoFar < tarch::parallel::Node::getInstance().getNumberOfNodes()) {
  ranksUsedSoFar += increment;
  increment *= THREE_POWER_D;
  _coarsestRegularLevelUsedForDD ++;
}
```

   Typically, I determine this level in `beginIteration()` of the mapping that constructs the initial grid. It is thus determined in parallel on all ranks as soon as a rank joins the game.

2. I make the grid setup refine the grid in `touchVertexLastTime`, i.e. the grid is created with one level per sweep. As this part of the code runs in parallel, we run over the grid k' times, add one level per sweep (so k' becomes the depth of the tree), and at the same time distribute the grid among the ranks. We successively flood the MPI nodes. However, we continue to add new levels if and only if we do not exceed the initial grid depth determined in step 1:

```
...::touchVertexLastTime(...) {
  if (
    shallRefine(fineGridVertex,fineGridH)
    &&
    coarseGridVerticesEnumerator.getLevel() < _coarsestRegularLevelUsedForDD
  ) {
    fineGridVertex.refine();
  }
  ...
}
```

3. I make all the ranks switch off dynamic load balancing the first time the global master runs a step on all ranks out there:

```
void picard::runners::Runner::runGlobalStep() {
  // assertion( !peano::parallel::loadbalancing::Oracle::getInstance().
  // isLoadBalancingActivated() );

  peano::parallel::loadbalancing::Oracle::getInstance().activateLoadBalancing(false);
}
```

   For this, I remove the assertion original put in by PDT. I know what I'm doing, as ...

4. I run through the grid until it becomes stationary, i.e. does not change anymore and is properly distributed. Due to the variable `_coarsestRegularLevelUsedForDD` this will require a couple of sweeps but will not set up the whole grid. Next, I switch off load balancing globally. Finally, I rerun the grid construction twice. The runner then resembles

```
repository.switchToCreateGrid();
do {
  repository.iterate();
} while ( !repository.getState().isGridBalanced() );

repository.runGlobalStep();
runGlobalStep();

repository.iterate();
repository.iterate();
```

5. So far, the last two iterates do not change the grid anymore and they notably do not build up the whole grid if the grid is truncated by `_coarsestRegularLevelUsedForDD`. I finally return to the mapping's touch vertex last time event and continue to refine if the load balancing is switched off. This refinement will kick in the in first of the two additional `iterate` commands.

```
...::touchLastTime(...) {
  if (
    shallRefine(fineGridVertex,fineGridH)
    &&
    !peano::parallel::loadbalancing::Oracle::getInstance().isLoadBalancingActivated()
  ) {
    fineGridVertex.refine();
  }
}
```

6. This new fragment will make the last `iterate` introduce one additional level that is finer than `_coarsestRegularLevelUsedForDD`. When it invokes the corresponding creational routines, we now use `enforceRefine` to build up the remaining grid parts in one sweep.

```
...::createInnerVertex(...) {
  if (
    shallRefine(fineGridVertex,fineGridH)
    &&
    !peano::parallel::loadbalancing::Oracle::getInstance().isLoadBalancingActivated()
  ) {
    fineGridVertex.enforceRefine();
  }

}
```

Exactly the same has to be done within `createBoundaryVertex`.

6.3 MPI quick tuning

Time: Around 15 minutes.
Required: A working MPI code.

This section collects a couple of really primitive measurements to make your code faster.

6.3.1 Filter out log statements

It is probably to simple to mention, but all our teams from time to time forget this. One of the major things slowing down codes is writing to the terminal. So adding a few additional log filters can significantly speed up your code.

6.3.2 Switch off load balancing

Most of Peano's load balancing algorithms (at least the ones coming along with the standard package) rely on a central node pool. If a rank decides that it would be advantageous to split up its domain, it sends a request to the first rank whether there are any idle nodes available. If your code already uses all ranks, this is a time consuming process that suffers from latency. If you know a prior that the load balancing is static and no further splits of subdomains are possible, it does make sense to switch the load balancing off. There is a routine `activateLoadBalancing` operation on the load balancing oracle to do so.

This operation has to be called on each individual rank, i.e. you can switch the load balancing on and off on a rank-per-rank basis. There are basically two variants/patterns to disable the load balancing:

1. You may introduce a new mapping that does nothing besides switching the load balancing off (typically in `beginIteration`). You then merge this mapping into your other adapters.
2. You add a new bool to your state. In the global runner you set this boolean flag once you want to switch the load balancing off. The state then is successively propagated to the workers. In `beginIteration`, you analyse this bool (in any mapping) and you switch off the load balancing if the flag is set.

Peano also offers the opportunity to invoke a global step on all ranks prior to an `iterate` call. This feature can be used to switch off the load balancing, too:

```
void picard::runners::Runner::runGlobalStep() {
  peano::parallel::loadbalancing::Oracle::getInstance().activateLoadBalancing(false);
}


int picard::runners::Runner::runAsMaster(...) {
  ...

  repository.runGlobalStep(); // on all other ranks
  runGlobalStep(); // and locally, too
}
```

As clarified in the documentation of the operations (see the autogenerated header files of your repository, e.g.), you have to be careful if you follow this variant: You are never allowed to run a global step if any rank is involved in a join or fork.

6.4 Reduce MPI Synchronisation

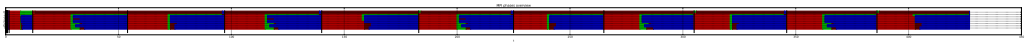
Time: Around 60 minutes.
Required: A working MPI code.

Peano has very strong constraints on the master-worker and worker-master communication as the data exchange between these two is synchronous. It imposes a partial order. If that slows down your application (you see this from the mpianalysis reports), you can kind of weaken the communication constraints. Often, some data is not required immediately, not required globally all the time, or doesn't have to be 100% correct at all algorithmic stages. This chapter discusses some things that you can do then.

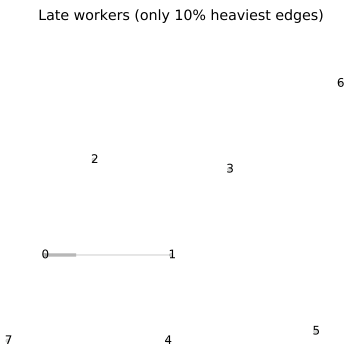
On the following pages, we assume that you have proper load balancing.

6.4.1 The smell

Strong synchronisation materialises in very regular patterns where each rank waits for rank 0 to start up a new traversal.



It also becomes obvious if you study how often a master has to work for its workers. In the picture below, only rank 0 synchronises the other ranks. In this case, you have to weaken the global synchronisation. If multiple of these edges pop up, it is time to weaken all the worker-master synchronisations—unless you can identify that you have a load balancing issue.



6.4.2 Weaken synchronisation with global master

The global master (rank 0) is kind of a pulse generator for the whole code. Whenever the `runAsMaster` operation triggers `iterate`, it tells each rank that handles a partition which adapter to use and to start its traversal or wait for its master to trigger the traversal, respectively. This is a very strong synchronisation. Notably, no rank can continue to work with the next iteration unless rank 0 runs into the next `iterate` as well. There are basically two variants to improve this situation:

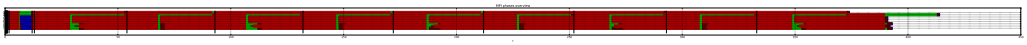
1. Perform more than one time step with the same adapter and settings in a row. For this, use the integer argument of `iterate()`. Note that running multiple time steps switches off load balancing for this phase of the program. Obviously, this version works if and only if you run the same adapter several times.

2. You may alternatively find out that you don't need the rank 0 (that doesn't hold any data anyway) to wait for all the other ranks in each iteration. Often, you run for example a sequence of adapters and you require global data (such as global residual) only after the last run. If you want to realise the second variant, you have to ensure that all mappings you use (also the predefined ones) return false in `prepareSendToWorker(...)`.

> **Remark:** If you want to validate that reductions have been skipped, switch on the log info of `peano::grid::nodes::Node::updateCellsParallel StateBeforeStoreForRootOfDeployedSubtree`.

Please be aware that reduction are skipped by the kernel if and only if all mappings allow the kernel to switch off the reduction. Furthermore, load balancing has to be disabled. If you want to load balance, master and worker ranks have to communicate with each other and may not skip any data/status exchange.

We also observe that the reduction skips often only change the communication profile but do not speed up the computation. Often it is only a preparatory step to switch off boundary data exchange afterwards. Once this is done, you should get a profile as below. It is more or less completely asynchronous, and all data exchange (blue) is hidden in the background, i.e. not visible anymore:



6.4.3 Postpone master-worker and worker-master data exchange

Take the communication specification of each mapping. By default, they are set to the most general case. Adopt it to your algorithmic needs (see documentation of the communication specification class)/

Introduce eager send and late receive

By default, Peano send away data from a local node if and only if it has traversed the whole local tree. In return, it requires all input data before it starts to traverse anything. You may want to tailor this to your needs and send data earlier and receive data later which allows you to overlap computations more aggressively. To do so, you have to adopt the communication specification fields of your mappings. See the documentation of the underlying class for more details. Avoid communication with rank 0

6.4.4 Skip worker-master data transfer locally/sporadically

6.5 Other ideas

, we assume that you've a reasonable load balancing but see that all the

The smell

Solution

Related pitfalls & ideas

Peano takes the computational Domain (the unit square, e.g.) and embeds it into a $3^d patch. This surrounding patch is hel$

foobar

Though rank 0 has deployed all cells to other ranks, still all workers of rank 1 are adjacent to rank 0. If they refine (and they most probably will do as most PDE solvers refine along the domain boundary), there is a pretty huge refined surface that connects each of the eight workers of rank 1 with rank 0. And now rank 0 becomes a bottleneck though rank 0 does no computation at all.

One solution is to extend the computational domain by a halo region. For a unit square, using an offset of [-1/7 x -1/7] and a bounding box size of [9/7 x 9/7] has proven of value. This way, all

halo cells of rank 0 are sufficiently away from the domain's real boundary. So, if a worker of rank 1 refines, it does not share additional Vertices with rank 0. 0 is not a bottleneck anymore.

To identify whether you can benefit from this technique, try a simple run with a regular grid and only two ranks. In this case, you should not see any speedup, as all work is deployed by rank 0 to rank 1. However, you should also not observe a significant runtime penalty. If you do observe, try this fix.

To realise the change, you might change

peano::geometry::Hexahedron geometry(tarch::la::Vector¡DIMENSIONS,double¿(1.0), tarch::la::Vector¡DIMENS
); particles::pit::repositories::Repository* repository = particles::pit::repositories::RepositoryFactory::getInstance().
geometry, tarch::la::Vector¡DIMENSIONS,double¿(1.0), // domainSize, tarch::la::Vector¡DIMENSIONS,double¿(0.0
// computationalDomainOffset);

into

peano::geometry::Hexahedron geometry(tarch::la::Vector¡DIMENSIONS,double¿(1.0), tarch::la::Vector¡DIMENS
); particles::pit::repositories::Repository* repository = particles::pit::repositories::RepositoryFactory::getInstance().
geometry, tarch::la::Vector¡DIMENSIONS,double¿(9.0/7.0), // domainSize, tarch::la::Vector¡DIMENSIONS,double
1.0/7.0) // computationalDomainOffset);

Disable load balancing

Joins and forks are expensive operations and furthermore hinder Peano to use its shared memory parallelisation, i.e. Peano always switches off multithreading if it has to rebalance a rank. As a consequence, it often makes sense to switch the load balancing oracles - to use an oracle not rebalancing at all most of the time but to identify critical steps where another racle rebalancing is us used. Check the load balancing and node weights

One of the first tuning activities is to analyse the load balancing. Peano has a mpianalysis interface and there analysis tools around. However, also the simple Default mpianalysis does the job - at least for stationary partitions, i.e. as long as you don't rebalance. Run your application and switch of the info output of tarch::mpianalysis. Pipe the results into a file and run the Shell/Python script from the mpianalysis directory on the output. This should result in a picture and you can analyse whether the partitioning fits to your expectations.

If it doesn't, you can tune your load balancing. Prior to this, I however recommend that you write down a cost model - how expensive should one cell be to solve? Then you can use the load per cell individually in your mappings and thus guide the load balancing (actually any load balancing you intend later on to use) how costly different subtrees are. Doublecheck the multiscale concurrency

Peano relies on a modified depth-first (dfs) traversal. The parallel variant also is a dfs, but whenever the dfs traversal encounters a remote node, it makes another mpi rank traverse the corresponding spacetree, while it continues itself with the local subtree. Before is ascends again, it checks whether the remote subtree traversals have terminates as well. As a result, it is important to split up the tree on an as coarse level as possible to obtain a high concurrency Level. Let's study a toy problem in 1d:

foobar

In the upper picture, we have forked 2,4,7,8,10 and 12 to remote ranks while we stop the forking on level 1. As a result, our dfs descends into node 1, then forks 3 and 4, waits until they are done, continues with node 5, forks 7 and 8, does ist local 6, waits for 7 and 8 to finish, and continues with 9 forking 11 and 12. Obviously, the maximum concurrency level is two. If we change the decomposition into the lower splitting, the concurrency level is 7 (mind that 8 should have a different colour than 0 and 9, but that's only a visualisation relict).

Now, one has to Keep in mind that Peano forks only subtrees that have a certain regularity: Only nodes (and hence their children) can be forked where all $2^d adjacent vertices are refined. So, if we argue the other way r$
$to refine all vertices up to a certain level independent of your application needs) to allow the load balancer to fork a way sub$

Often, enforcing this kind of regularity is not possible within the mappings, as the mappings work basically inside the computational domain. Given the sketched situations, it can be advantageous however also to refine within obstacles or along complicated boundaries regularly. Therefore, peano::parallel::loadbalancing::OracleForOnePhase holds another attribute that you can uuse to enforce a certain grid regularity and to enable your code to fork more aggressively. Introduce

administrative ranks and reduce algorithmic latency

Throughout the bottom-up traversal, each mpi traversal first receives data from all its children, i.e. data deployed to remote traversals, and afterward sends data to its master in turn. Unfortunately, Peano has to do quite some algorithmic work after the last children record has been received if and only if some subtrees are also to be traversed locally. It hence might make sense to introduce pure administrative ranks that do not take over any computation on the finest grid level. Again, we do a brief 1d toy case study:

foobar

In the upper case, the blue rank triggers the red one to traverse its subtree. The red one in turn tiggers 3 and 4. Afterward, it continues with 2 and then waits for 3 and 4 to finish. After the records from 3 and 4 have been received, it has to send its data to 0 to allow 0 to terminate the global traversal. However, between the last receive and the send, some administrative work has to be done, as the red node also holds local work (it has to run through the embedding cells to get the ordering of the boundary data exchange right, but that's irrelevant from a user point of view). This way, we've introduced an algorithmic latency: Some time elaps between 3 and 4 sending their data and the red one continuing with the data flow up the tree. This latency becomes severe for deep Splittings.

In such a case, it is a better idea to make the red one fork all of its work. See the lower part of the Illustration. In this case, (almost) no local administration is required, i.e. 1 accepts the finished Messages from 2,3 and 4 and almost immediately passes on the token to 0. Now, 1 basically does no work and you introduce a bad balancing here. But you have mpi rank overloading to compensate for this. And a latency reduction usually is more important. Exploit overloading

Peano's parallelisation is based upon tree-splits, i.e. the code can 'only' deploy whole subtrees to other ranks. Imbalances thus are always built-in. They become the more severe the fewer mpi ranks one uses. Thus, one has to check carefully whether mpi overbooking pays off. A general rule of thumb is that the smaller the computational workload the higher the overbooking should be. For codes with an extremely low compute load (just moving data, e.g.), overbooking by a factor of four on SandyBridge seems to be the method of choice. In that case, you start 64 mpi ranks per node. On SuperMUC, you have to restrict yourself to 32 ranks per node due to a load leveler constraint.

Peano provides both mpi and shared memory parallelisation. I currently recommand to use the TBB variant of the latter. Following the overbooking discussion above, it does make sense to equip each mpi rank with t threads though the total number of threads then outnumbers the number of cores by far. This way, some mpi ranks might become idle throughout the computation, but their cores are grapped by other mpi ranks due to their many tbb threads eventually. Pays off in most cases ... as long as the shared memory scales for reasonably fine grids and as long as your grid has such regular subregions. Optimise worker-master commmunication

Given the output of the component mpianalysis (either via text file or the postprocessing script), you can identify whether the masters had to wait for their workers a significant time. If that is the case, i.e. if you observe late workers,

try to balance your workload better, or even try to undersubscribe the workers.

In the latter case, you try to assign nodes acting both as compute nodes and as masters a bigger workload than those without any workers to administer. The rationale is that nodes far away from the global master in the call hierarchy may not delay the time per traversal as any delay there has a huge impact. Consequently, it is better to assign them a smaller workload and to give them time to send away their finished messages (that also have to run through the network). One avoids algorithmic latency. The price to pay is a non-optimal workload balancing.

If a node delays ist master and you cannot change its workload, study its individual runtime profile. If the code spends a signficant time within its boundary exchange, this means that it needs this significant time to wait until mpi has released its last send and receive request and other nodes have delivered their data. Now, if all workload is reasonable balanced, you cannot do anything about the latter fact. However, it might make sense to try a different buffer size. As the code spends all this time to wait for the last peace of data to arrive and to release its current buffer, a smaller buffer size might lead to a situation where the nodes can exchange more data in the

background. Splitting up the buffer into smaller chunks then is an option, i.e. to reduce the buffer size. Be Aware that smaller buffer sizes increase the administrative overhead. A too small buffer size hence slows down the code.

If nodes delay their masters but are not suffering from data exchange, you have to reduce their workload. If that is not possible and if your code runs correctly (read: no deadlocks), it makes sense to try to switch Peano's communication protocols to blocking send. For this, you have to switch the corresponding flag in your compiler specific settings. Send them to sleep

If several ranks are booked to one node, they compete for the network facilities. This competition can slow down the overall system: If one rank is done, it listens for a new message from its master. The other ranks however might still need the network to exchange data and thus are throtteled. For avoid this, you might think either to switch to a real blocking call (given that your MPI implementation realises this internally via an interrupt) or send threads waiting for a synchronous message to sleep.

To do so, you have to go the compiler-specific settings and introduce a sleep penalty. The compiler flags are called ReceiveMasterMessagesBlocking or similar. The allowed values are documented. Switch off reduction

The reduction of data along the spacetree often harms Peano's performance significantly. Check whether you can live (at least for some iterations) without the reduction. Often, e.g., load balancing and reduction are important in time stepping, but one can always do few linear algebra traversals without reducing any global data.

Peano's iterate method can be passed false as argument. Then, the reduction is avoided. And your code should run faster and not suffer from latency and ill-balancing. Not that much at least.

Please note that there are two different types of reduction: Peano by default transers vertices and cells bottom-up and thus, e.g., allows for load balancing. This is the behaviour you can switch off due to the flag. However, note that load balancing relies on reduced data, i.e. if you switch this off, you also disable load balancing. A different story is the user-defined reduction. Many applications reduce data in their services (if they have such global object instances per node) or send data to the master in their events. This is a reduction you have to handle correctly. Diving into implementation details

The file peano::utils::PeanoOptimisation holds a number of different defines that influence Peano's runtime behaviour. With respect to MPI the compile arguments

-DnoParallelExchangePackedRecordsAtBoundary -DnoParallelExchangePackedRecordsBetweenMasterAndWork
-DnoParallelExchangePackedRecordsInHeaps -DnoParallelExchangePackedRecordsThroughoutJoinsAndForks

do have impact on the communication behaviour. They tell Peano not to reduce the memory footprint of the messages prior to sending them away. If you switch them off, you increase the bandwidth required (perhaps only slightly) but you skip the marshalling and unmarshalling steps. This might yield significant speedup but depends strongly on the PDE-specific data exchanged.

Besides the four flags from above, there are some more settings that might interplay with the scaling of your application. But here, everything is trial-and-error. Become topology-aware

Peano uses a node pool strategy to decide which rank assists which other rank if a rank asks for additional workers. By default, this strategy is FCFS and the ranks are just handed out without additional considerations. It thus can happen that all big partitions are assigned to the first p' ranks whereas the remaining p-p' ranks become responsible for rather small subgrids only.

In such a case it often does make sense to implement topology-awareness into your node pool server (it also might make sense to add problem-awareness, i.e. knowledge about your grid, to the oracle, but that is a different story). A simple example for such a generic server can be found in the toolboxes. It assumes that there are k ranks assigned to each compute node. As a result, it assigns work to the ranks in the order 1, k, 2k, 3k, ..., 2, k+1, k+2, and so forth. It realises a modulo work assignment. This reduces the memory required per compute node and it also distributes the communication data footprint evenly.