Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 545–554

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Efficient Implicit Parallel Patterns for Geographic Information System

Kevin Bourgeois[1,2], Sophie Robert[1], Sébastien Limet[1], and Victor Essayan[2]

[1] Univ.Orléans, INSA Centre Val de Loire, LIFO EA4022, Orléans, France
(kevin.bourgeois, sophie.robert, sebastien.limet)@univ-orleans.fr
[2] Géo-Hyd (Antea Group), Olivet, France
(kevin.bourgeois, victor.essayan)@anteagroup.com

Abstract

With the data growth, the need to parallelize treatments become crucial in numerous domains. But for non-specialists it is still difficult to tackle parallelism technicalities as data distribution, communications or load balancing. For the geoscience domain we propose a solution based on implicit parallel patterns. These patterns are abstract models for a class of algorithms which can be customized and automatically transformed in a parallel execution. In this paper, we describe a pattern for stencil computation and a novel pattern dealing with computation following a pre-defined order. They are particularly used in geosciences and we illustrate them with the flow direction and the flow accumulation computations.

Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

Keywords: Parallel programming, Implicit parallelism, Performance, GIS

1 Introduction

During the last decades, fast progress in capturing and measuring methods in every scientific domains made the amount of available data grows in a dramatic way. Such a revolution allows the scientists to refine their models and simulations. However computing on such amounts of data requires the use of powerful computers, such as clusters, to get reasonable run-times. One of the main locks in many sciences is to gather in a single man or at least in a single team, high skills in both the specific scientific domain and in high performance computing. One solution to address this issue consists in providing to scientists, who are able to program in sequential way, tools which help them to (semi-)automatically derive efficient parallel programs from a program that hides, more or less, technicalities due to parallelism. Currently, great efforts are done in computer science to define efficient tools to help non-HPC experts to program efficiently parallel computers. There are mainly two families of solutions. On one hand are dedicated libraries that provide turnkey implementations of classical and domain-specific functions. On the other hand, skeleton programming or pattern programming provide an abstract model of an algorithm that can be customized and transformed into a parallel program. Often, these

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.235

approaches are associated with a Domain Specific Language (DSL) to describe, in a familiar way for the scientist, the skeleton and its inputs.

Geosciences are no exception and amount of data is a major concern even for SMEs or public labs working on geographic information systems (GIS) to conduct environmental studies like soil erosion or analysis of the water cycle. The work presented in this paper addresses the implicit parallelism for geoscientists. It focuses on two parallel patterns widely used when computing on large GIS represented by 2D rasters. However this work takes place in a more general project that aims at providing an implicit parallel computing framework dedicated to geoscientists. This framework illustrated Figure 1 relies on implicit parallelism associated to a DSL. It is split into three main layers from a high level user interface to program the algorithm to a low level data system to efficiently manage data on disks. These layers are connected by an implicit parallelism pattern layer. This framework targets computer architectures such as clusters of small or middle size (i.e. from 10 to 100 nodes) as those frequently found in SMEs or scientific labs.

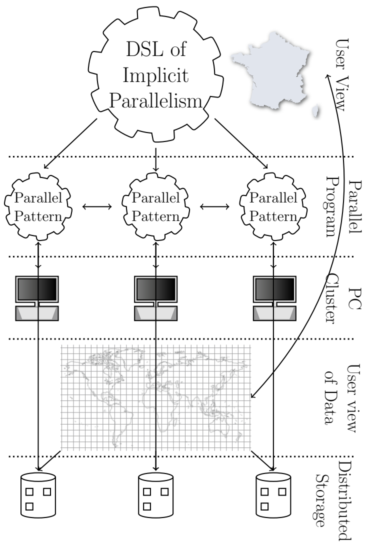

Figure 1: Implicit Parallel Framework for GIS

At the top level, the geoscientist has a classical view of his/her GIS and programs in a classical way using a DSL embedded in Python. This DSL is based on libraries to compute on raster and/or vector datasets using some predefined patterns. At the low level, the datasets are stored in a distributed file system and the framework provides a layer that allows the user to efficiently select data he/she requires as well as to store intermediate results. At the middle, a layer is responsible to make the connection between the two ends of the framework and to produce efficient parallel programs from the ones written by the user. This middle layer relies on the implicit parallel pattern programming interface written in C++ and using MPI parallel library. This middle layer uses the data layer library to access and store data. Different patterns have been implemented corresponding to different classes of algorithms frequently used in geosciences. They are optimized to efficiently parallelize programs according to their class. From the program written by the user with the DSL, it is possible to determine the patterns to apply in the parallel program and a C++ program is automatically derived using them.