Parallel Computing 40 (2014) 144–158

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

Tuning and hybrid parallelization of a genetic-based multi-point statistics simulation code

Oscar Peredo [a,b,*], Julián M. Ortiz [b,c], José R. Herrero [d], Cristóbal Samaniego [a]

[a] *Barcelona Supercomputing Center (BSC-CNS), Department of Computer Applications in Science and Engineering, Edificio NEXUS I, Campus Nord UPC, Gran Capitán 2-4, 08034 Barcelona, Catalunya, Spain*

[b] *Advanced Laboratory for Geostatistical Supercomputing, Advanced Mining Technology Center, University of Chile, Chile*

[c] *Department of Mining Engineering, University of Chile, Av. Tupper 2069, Santiago 837-0451, Chile*

[d] *Computer Architecture Department, Universitat Politècnica de Catalunya (UPC-BarcelonaTech), Campus Nord, Mòdul C6, Desp. 206, C/Jordi Girona 1-3, 08034 Barcelona, Catalunya, Spain*

ARTICLE INFO

Article history:
Received 15 February 2013
Received in revised form 19 January 2014
Accepted 18 April 2014
Available online 29 April 2014

Keywords:
Geostatistics
Stochastic simulation
Multi-point statistics
Code optimization
Parallel computing
Genetic algorithms

ABSTRACT

One of the main difficulties using multi-point statistical (MPS) simulation based on annealing techniques or genetic algorithms concerns the excessive amount of time and memory that must be spent in order to achieve convergence. In this work we propose code optimizations and parallelization schemes over a genetic-based MPS code with the aim of speeding up the execution time. The code optimizations involve the reduction of cache misses in the array accesses, avoid branching instructions and increase the locality of the accessed data. The hybrid parallelization scheme involves a fine-grain parallelization of loops using a shared-memory programming model (OpenMP) and a coarse-grain distribution of load among several computational nodes using a distributed-memory programming model (MPI). Convergence, execution time and speed-up results are presented using 2D training images of sizes $100 \times 100 \times 1$ and $1000 \times 1000 \times 1$ on a distributed-shared memory supercomputing facility.

1. Multi-point statistics simulation

Numerical modeling with geostatistical techniques aims at characterizing natural phenomena by summarizing and using the spatial correlation of collected data in order to measure the uncertainty at unsampled locations in space. As explained by Deutsch [15], in simulation techniques, this spatial correlation is imposed into a model commonly constructed on a regular lattice. The models must reproduce the statistical (histogram) and spatial distribution (variogram or other spatial statistics) and their quality is often judged in terms of the reproduction of geological features.

Conventional techniques in geostatistics address the modeling using statistical measures of spatial correlation that quantify the expected dissimilarity (transition to a different category) between locations separated by a given vector distance, in reference to a given attribute, such as the facies, rock type, porosity, grade of an element of interest, among others. This is done using the variogram. Limitations of these techniques have been pointed out in that they only account for two locations at a time when defining the spatial structure [27]. Much richer features can be captured by using multi-point statistics (MPS)

---

[*] Corresponding author at: Barcelona Supercomputing Center (BSC-CNS), Department of Computer Applications in Science and Engineering, Edificio NEXUS I, Campus Nord UPC, Gran Capitán 2-4, 08034 Barcelona, Catalunya, Spain. Tel.: +34 934016817.

E-mail address: operedo@alges.cl (O. Peredo).

http://dx.doi.org/10.1016/j.parco.2014.04.005

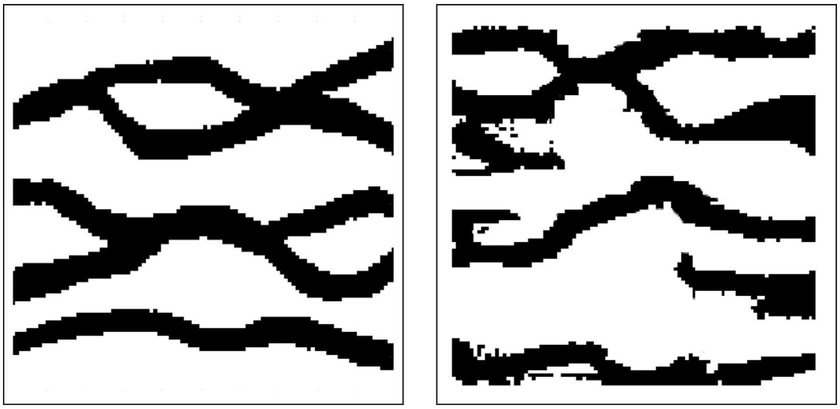

Fig. 1. Training image (left) and simulated realization (right).

that consider the simultaneous arrangement of the attribute of interest at several locations, providing the possibility to account for complex features, such as hierarchy between facies, delay effects, superposition or curvilinearity.

MPS simulation aims at generating realizations that reproduce pattern statistics inferred from some training source, usually a training image. For example, in Fig. 1, left, we can see a training image based on sinuous channels with a simulated realization. These training images are used as a pattern database to generate simulations of the underlying image, as shown in Fig. 1, right. The simulations use those patterns with the aim that the training and simulated images share the same pattern histogram.

There are several approaches to simulate accounting for MPS. Modifications of conventional methods to impose local directions of continuity using the variogram is a simple approach to impose some of the complex geological features [43,44]. Object based methods and methods inspired in the genetic rules and physics of the deposition of sediments in different environments also seek to overcome the limitations of conventional categorical simulation techniques, with significant progress [13,42,36]. Presently, the most popular method is a sequential approach based on Bayes' postulate to infer the conditional distribution from the frequencies of multi-point arrangements obtained from a training image. This method, originally proposed by Guardiano and Srivastava [21], and later efficiently implemented by Strebelle and Journel [41], is called single normal equation simulation (snesim) (see also [40]). This method has been the foundation for many variants such as simulating directly full patterns [1,17] and using filters to approximate the patterns [45]. The use of a Gibbs Sampling algorithm to account directly for patterns has also been proposed [2,29]. A sequential method using a fixed search pattern and a 'unilateral path' also provides good results [8,9,33]. Other approaches available consider the use of neural networks [3,4], updating conditional distributions with multi-point statistics as auxiliary information [30–32] or secondary variable [24]. Recently, a couple of new approaches focused on patching patterns directly to reduce computing time and impose larger scale structures, have been presented [37,18]. These methods have a significant potential for practical applications. Alternatively, the problem can be addressed as an optimization one, using simulated annealing [14] or genetic algorithms [35]. The genetic approach is still under development, but essentially follows the same stochastic strategy as the annealing scheme. This work focuses on code optimizations and parallelization of a genetic-based sequential code that simulates categorical variables to reproduce multi-point statistics. However, many of the techniques and ideas proposed here can be applied to other codes implementing similar simulation algorithms.

In Section 2 we explain the basic ideas about genetic algorithms, parallel architectures and programming models. After that, the main bottlenecks of the genetic-based simulation are detailed in Section 3. A brief explanation of the actual implementation is presented in Section 4, together with the proposed code optimizations and parallelization schemes, in Sections 5 and 6 respectively. Finally, in the last sections we include the results obtained and final conclusions.

2. Genetic algorithms and parallel computing

Genetic algorithms (GA) were developed in the 1970s with the work of Holland [23] and in subsequent decades with De Jong and Goldberg [12,19]. Initially used to find good feasible solutions for combinatorial optimization problems, today they are used in various industrial applications, and recent advances in parallel computing have allowed their development and continuing expansion.

In the canonical approach of GA, typically there is an initial population of individuals, where each individual is represented by a string of bits, as $indiv_k = 000110101$, and a fitness function $fitness(indiv_k)$ which represents the performance of each individual. The fitness function, or objective function, is the objective that must be minimized through the generations over all the individuals. A termination criteria must be defined in order to achieve the desired level of decrement in the