Journal of Computer Science 3 (1): 47-50, 2007
ISSN 1549-3636

Exploring Strategies for Parallel Computing of RS Data Assimilation with SWAP-GA

[1]Shamim Akhter, [1]Kiyoshi Honda, [1]Yann Chemin and [2]Putchong Uthayopas
[1]School of Advanced Technologies, Asian Institute of Technology, Thailand
[2]Department of Computer Engineering, Kasetsart University, Thailand

Abstract: An agro-hydrological simulation model is useful for agriculture monitoring. One issue in running such model is parameter identification, especially when the target area is large such as provincial or country level. Remote Sensing (RS) provides us with useful information over large area. RS cannot observe input parameters of agro-hydrological models directly. However, a method to estimate input parameters of such model from RS using data assimilation has been proposed by Ines[1] using the SWAP (Soil, Water, Atmosphere and Plant) model. Genetic Algorithm (GA) was used in this optimization process. The combined model of SWAP and GA is called SWAP-GA model. When dealing with sufficiently large and complex processing with RS data, single computers time processing extends to unacceptable limits. It becomes necessary to introduce methods for using higher processing power such as distributed computing. Cluster based computing support both high performance and load balancing parallel or distributed applications. Implementing SWAP-GA in Cluster computers will remove the computational time constraint, with this hypothesis three different parallel SWAP-GA approaches are proposed in this study. Distributed population (where GA will work on distributed manner), Distributed pixel (Pixels are processed in parallel) and Mixed of distributed population and pixel model called Hybrid model. The technical considerations of implementing such methodologies are visited here.

Key words: SWAP, genetic algorithms, data assimilation, cluster computing

INTRODUCTION

Remote Sensing plays a vital role in agriculture monitoring especially when considering large cropping areas. Agriculture monitoring activities with Remote Sensing can be done at a regular interval to which 15 days is reasonable for medium to large size pixel satellite sensors. Evapotranspiration (ETa) is one of the indicators of crop productivity and can be estimated from satellite Remote Sensing. SWAP[2] is a simulated model that can serve as crop productivity monitoring tool. The assimilation of satellite images ETa and SWAP produces output with detailed crop productivity parameters. Due to the changing in input parameters of SWAP model for pixel-to-pixel basis, the assimilation search space is very large, evolutionary search algorithms like the genetic algorithm[3] are performing well in such conditions. Similar work by Ines[1] used some remotely sensed information combined with a binary GA and SWAP model for optimizing soil hydraulic parameters. A real-coded genetic algorithm[4], was applied by Chemin *et al.*[5]. The real-coded GA is removing one layer of programming that is the coding/decoding to-from binary strings.

Eventhough, the methodological issue of finding the input parameters of SWAP has been solved by bridging a genetic algorithm onto it, the data assimilation problem has a practical dimension to be solved: it requires computation time. Operating on these assimilating procedures with a single computer system takes an unacceptable time. It becomes necessary to introduce methods for using higher processing power such as distributed computing.

The main purpose of parallel or distributed processing is to perform computations faster than can be done with a single processor by using number of processors concurrently. Cluster is a type of parallel and distributed processing system, which consists of a collection of interconnected stand-alone computers working together as a single, integrated computing resource. Cluster based computing has been traditionally associated with high performance computing of massive CPU bound applications. With a hypothesis that Cluster style computing will remove computational time constraints for SWAP-GA, different SWAP-GA Cluster implementing procedures for remotely sensed images can be considered. The AIT Cluster computers (http://optima.ait.ac.th) and Kasetsart University Cluster computers (http://magi.cpe.ku.ac.th/scmsweb/scms_home.html and http://maeka.ku.ac.th/) can be used for the purpose of these experiments.

To implement the parallel SWAP-GA module using Cluster computers, three (3) approaches are proposed. 1) All populations will be distributed properly among available slaves by the Master computer. Slaves do the evaluation, generate the fitness and send back the population (with fitness) to the Master. 2) All pixels will be distributed among the available slaves.

Corresponding Author: Shamim Akhter, School of Advanced Technologies, Asian Institute of Technology, Thailand

Each slave will evaluate a total sequential SWAP-GA procedure inside itself for one pixel at a time and will produce a total assimilation result for that pixel in a file in their local Hard Disk. 3) The combined model of distributed pixel and distributed population is called Hybrid model.

The objectives of this study are to expose the different methodologies developed to tackle the HPC-based optimization at stake. The three methods are detailed in terms of theory, coding structure and hardware interaction.

Background: Serial SWAP-GA is a combination of SWAP crop model, Real Coded GA and ETa simulated data. ETa data is the input and GA is used for finding the best parameters for SWAP model (to find the crop growth rate) so that the output of SWAP will be assimilated with ETa data. The procedure of assimilating is shown in Fig. 1. The difference of SWAP-GA ETa data and actual ETa RS data is the cost value (Equation 1) and by inversing the summation of cost and constraints, a fitness value (equation 4) will be generated. The highest fitness value will provide better assimilation. "Y" indicates the fitness value is reached to a predefine threshold value and "N" indicates that the assimilation procedure is not complete yet. So the whole procedure will be repeated again.

The parameters under optimization are the starting date of cropping, the time extent of cropping and the groundwater depth in 1st January and in 31st December. It is expected to use rice pixels with double cropping as a case study.

The search domains for the dates of starting of cropping will require a non-overlapping restriction of about 90-100 days. The time extent of the cropping season will be between 3 to 5 months. The groundwater level maybe ranging from 0 to 500cm depth.

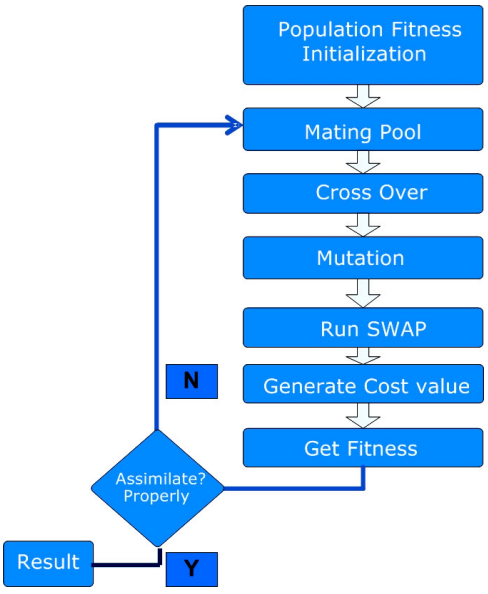

Fig. 1: Serial SWAP-GA model structure diagram

Consider C the cost function, having (*x,y,d*) parameters, *x* the *longitude [0-180/E-W]*, *y* the *latitude [0-90/N-S]*, *d* the *date [yyyymmdd]* .With *d* = [i,...,j], with i to j being the different satellite overpass dates, n is the sum of i to j. C holds the value of the average absolute difference of Etas (Satellite Image Etas and Simulated Etas).

$$C_{xy} = \frac{1}{n}\sum_{i}^{j} \left| ETa_{xyd} - ETa_{SWAP_{xyd}} \right| \text{ mm} \qquad (1)$$

The fitness of an individual having *xy* pixel location characteristics will be the inverse of the cost function times the constraint aiming at minimizing the differences between SWAP simulation and target ETa.

$$F_{xy} = \frac{1}{\left(C_{xy} * (1.0 + Constra\,int)\right)} \text{ -} \qquad (2)$$

Methodologies for parallel or distributed SWAP-GA: To parallelize the SWAP-GA the tool used is MPI[6]. MPI stands for "Message Passing Interface". MPI is a library of functions (in C) or subroutines (in FORTRAN) that one can insert into the source code to perform data communication between processors[7].

A Master-Slave procedure is maintained to distribute the jobs inside a cluster. In Master-Slave terminology, there will be a front-end node (processor, called Master), which will communicate with users and command the other nodes (called slave or working nodes). Whenever any job submitted to cluster, first all available nodes or processors are ranked from "0" to available processors number. "0" is the rank for master node. Thereby, jobs will be distributed to slaves by Master node.

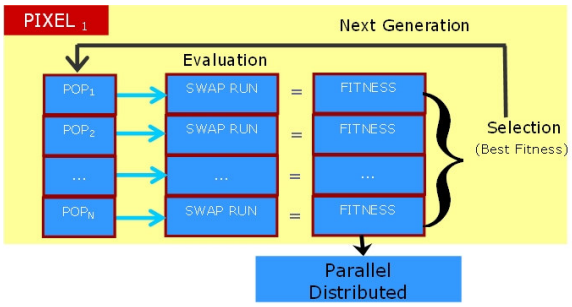

Fig. 2: Domain space of distributed population methodology

Parallel SWAP-GA with distributed population: The whole SWAP-GA model is decomposed into running GA sequences of a pixel. GA needs to calculate each population's fitness value, by executing SWAP one time. For all population's the fitness calculation procedure will be repeated by number of generations.

For each generation, GA serially calculates the populations' fitness, that is the time consuming part in SWAP-GA module. One approach to make SWAP-GA parallel is to distribute each population fitness calculation process to each processor of the Cluster; this approach is called "Distributed population" (Fig. 2).

Each generation, Master distributed all populations to slaves. First all populations are equally distributed among slaves and the rest populations are distributed to slaves sequentially through the rank 1…N slaves. Slaves do the evaluation for number of populations distributed to them, generate their fitness and send back the populations (with fitness) to the Master node (Fig. 3). This same procedure will be repeated by number of generations.

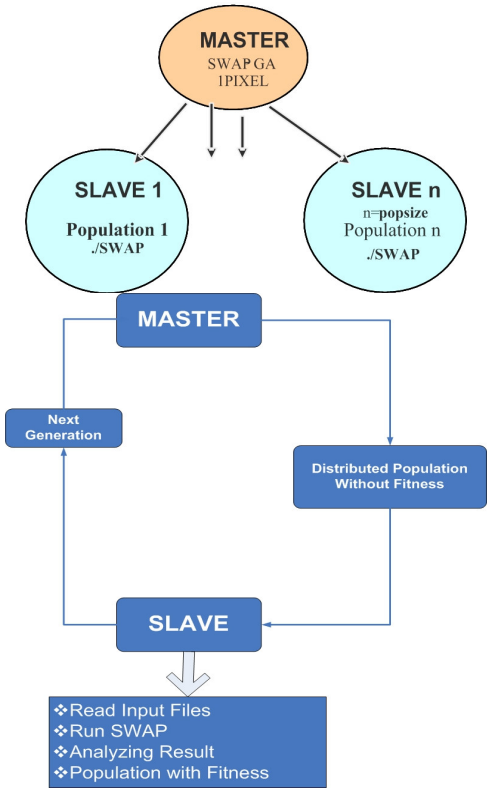

Fig. 3: Distributed population model structure

Parallel SWAP-GA with distributed pixel: Pixels are evaluated sequentially inside a serial SWAP-GA program running. However, there is no internal relation between the pixels in the evaluation procedure. Thereby, each pixel can be evaluated separately in each computing node. This process is called "Distributed pixel" (Fig. 4).

At the very beginning of the running code, it is assumed that the master will partition the total image into several pixels and all separate pixels (a set of pixels) information will be in separate folders (inside master local Hard Disk) maintaining a specific format. All pixels will be distributed among the available slaves.

First all pixels are equally distributed among slaves and the rest pixels are distributed to slaves sequentially through the rank 1…N slaves. At the running period each slave will just copy that specific folder or folders inside its own local Hard disk and evaluate total serial SWAP-GA procedure for that one pixel. It will eventually produce a total assimilation result for that pixel in a file in the local Hard Disk (Fig. 5).

Hybrid model: Hybrid is a combined model of Distributed population and Distributing pixel model (Fig. 6). The theory behind this model is to make

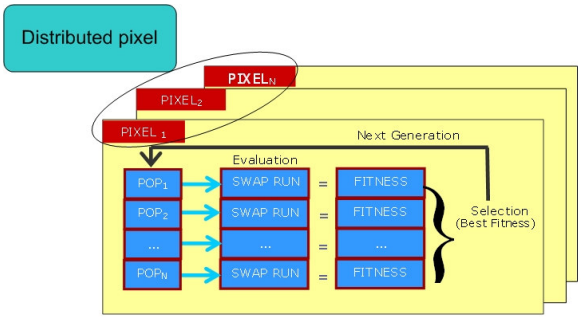

Fig. 4: Domain space of distributed pixel methodology

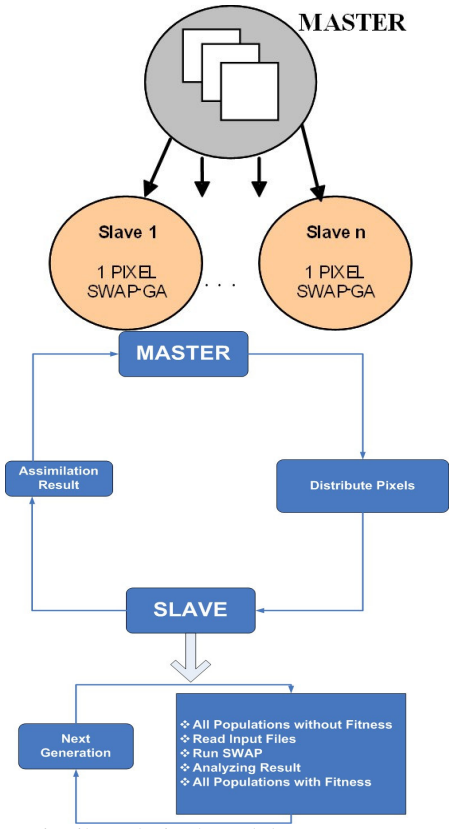

Fig. 5: Distributed pixel model structure

partitions inside the Cluster according to processor no (including master-rank "0" processor). The working methodology is to divide the Cluster (Equation 3) in such way that the Cluster contains more than one master (local master, including rank "0" processor) and each local master contains equal number of slaves (Equation 4). Each local master will take one pixel at a time and distribute all populations to its corresponding slaves. After evaluating the populations, slaves return

the population with fitness to its own master and master again sends the population to slaves for the next generation. This process will continue to satisfy the threshold value and each pixel assimilation will then be saved in a file. The examples of Hybrid model with 6 and 8 available processors are in Fig. 7.

$$\text{No of Masters (M)} = \text{sqrt}(\text{No of available processors}) \quad (3)$$

$$\text{No of Slaves (S)} = \frac{(\text{No of available processors} - \text{No of Masters})}{\text{No of Masters}} \quad (4)$$

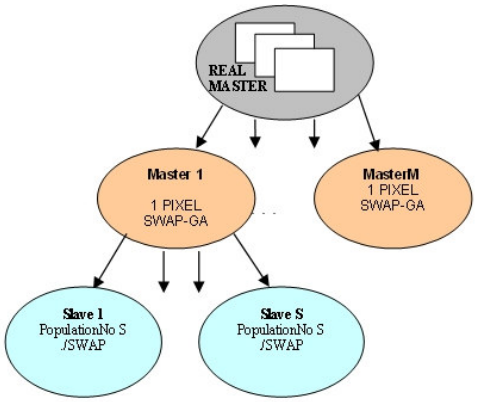

Fig. 6: Hybrid model conceptual structure diagram

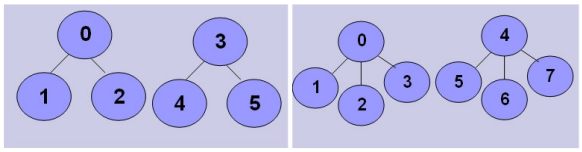

Fig. 7: Cluster partitioning with 6 and 8 available processors using hybrid model

DISCUSSION AND CONCLUSION

Different methodologies are developed for considering different types of environment. When the number of pixels to be processed is smaller than the number of CPUs, distributed population method can be used to utilize whole computing power together. Otherwise, distributed pixel may be useful. Moreover, it was clearly realized that the whole SWAP-GA coding can be separated into two different independent parts. One is the population evaluation procedure, where a large amount of population is evaluated for generating their fitness value and each population's fitness evaluation takes more than two (2) minutes. So, a methodology can be proposed to make the population's evaluation running in distributed manner. That is the core concept of distributed population method. Where as, another part is to assimilate the satellite ETa data for pixel by pixel basis and it takes approximately 30 minutes for generating one pixel's assimilation. Distributed pixel method provides an opportunity to assimilate each pixel's ETa data in parallel. Implementation of two of these methodologies, namely the population and pixel based distribution methods are in preparation at the time of the writing of this paper[8].

In the near future, a procedure will be implemented in Cluster computing where a Remote Sensing dataset may be analyzed by one preferred methodology examined here.

Further research may extend to Grid computing, where SWAP-GA may be applied with the hybrid methodology. In a Grid environment, the Grid master node will distribute one pixel into each cluster and each cluster will execute SWAP-GA locally through Distributed population method. Successful implementation of these methodologies will bring an opportunity to run SWAP-GA for larger Remote Sensing Image within an acceptable time limit and include new research opportunities not only in Grid-Cluster computing but also in agriculture monitoring. Furthermore, these applications will be the first time going to implement with crop model identification method using remote sensing on parallel computing environment. The output of this research will contribute greatly to advance the agricultural monitoring, which will contribute to country level food and economy security.

REFERENCES

1. Ines, A.V.M., 2002, Improved crop production integrating GIS and genetic algorithms. Ph. D. Thesis. AIT. Bangkok, Thailand. AIT Diss no. WM-02-01.
2. Van Dam, J.C., J. Huygen, J.G. Wesseling, R.A. Feddes, P. Kabat, P.E.V. van Waslum, P. Groenendjik and C.A. Van Diepen, 1997. Theory of SWAP version 2.0: Simulation of water flow and plant growth in the Soil-Water-Atmosphere-Plant ebvironment. Technical Document 45. Wageningen Agricultural University and DLO Winand Staring Centre. The Netherlands.
3. Goldberg, D.E. and K. Deb, 1989. A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. Foundations of Genetic Algorithms. Morgan Kaufman, San Mateo, Calif., pp: 69-93.
4. Michalewicz, Z., 1996. Genetic Algorithms + Data Structures= Evolution Programs. 3rd Ed. Rev. and extended Ed. Springer. USA.
5. Chemin, Y., K. Honda and A.V.M. Ines, 2004. Genetic algorithm for assimilating remotely sensed evapotranspiration data using a soil-water-atmosphere-plant model. Implementation Issues. Proc. FROSS/GRASS Users Conference, Bangkok, Thailand.
6. PACS Training Group, 2004. http://pacont.ncsa.uiuc.edu:8900/public/MPI/
7. Hoffman, F.M. and W.W. Hargrove, 2000. High Performance Computing: An Introduction to Parallel Programming with Beowulf, http://climate.ornl.gov/~forrest/osdj-2000-11/.
8. Akhter, S., 2005. Implementing the SWAP-GA Model in Cluster Computers. M.Sc. Thesis. Asian Institute of Technology, Khlong Luang, Thailand.