ISSN:0975-9646

Development of a New Scalable Parallelization Strategy for GCM Varsha and its Historical Perspective

T. Venkatesh[#] and U.N.Sinha*

[#]*Dept. of Computer Science & Engineering,*
Ghousia College of Engineering, Ramanagaram-562159, India

**Former Distinguished Scientist, NAL /CMMACS,*
Bangalore- 560017, India

Abstract: **In any parallel computing experiment scalability limits the extent to which available computational resources can be used gainfully. In India, though parallel meteorological computing has been in use for over two decades, the issue of scalability has been elusive. The present paper analyses the issues impeding scalability in existing parallelization strategy and proposes alternative having superior features through asymptotic analysis of volume of communication.**

Keywords: Parallel Computing, Scalability, Domain Decomposition, Parallel Algorithm.

1 INTRODUCTION

Parallel Computing in the last four decades has changed the scenario of computational experiments. Problems which were considered intractable or almost impractical to handle are being handled routinely.

Parallel Computing in India, too, has an interesting history. Indian parallel computing started in 1986 with Flosolver programme of CSIR-NAL(Council of Scientific and Industrial Research National Aerospace Laboratory) and got national focus in setting up of C-DAC(Centre for Development of Advanced Computing), and coming up of groups like ANURAG(Advanced Numerical Research and Analysis Group) of DRDO(Defence Research and Development Organization) and ANUPAM of BARC(Bhabha Atomic Research Centre) and many other institutes.

Flosolver Mk1 was the first Indian parallel computer built at CSIR-NAL in 1986 and since then many computationally demanding problem had been solved using Flosolver MK1; Flosolver had the main objective of solving fluid dynamical problems which it did well and a comprehensive account of the initial phase is reported in [1]. The decade of ninety has been the golden era of development of parallel computers and parallel computing in India and has been reflected in [2]. Around the same period at NAL, in the Flosolver series of parallel computers, Flosolver Mk3 had been developed which was instrumental in handling Direct Simulation of Navier-Stokes Equations for axisymmetric jet [3]- a problem of this class was hitherto been considered only within the reach of the western world. It is, therefore, not surprising that meteorological computing which was dependent on Cray computers at NCMRWF (National Centre for Medium Range Weather Forecasting) was selected to assess the emerging power of the parallel computers in India. It was an initiative of DST(Department of Science and Technology) and of the many participating teams Flosolver team of NAL was one of them; the objectives were to first parallelize the operational code and secondly to operationalize on the parallel machines of respective teams.

Thus began the first Indian experiment of parallelizing a large and complex application software – in this case a meteorological code known as GCM (Global Circulation Model). The first report of the parallelization entitled, "Monsoon Forecasting on Parallel Computers", was published in 1994 [4]. Subsequently, NAL also participated and presented in ECMWF (European Centre for Medium-Range Weather Forecasts) organized conferences at Reading in 1996 [5]. What emerged out clearly from these presentations is that

(a) The porting of the application code was scientifically correct i.e. the results from the earlier sequential version and the existing parallel version were within the limits of round off errors, and

(b) The issue of scalability of parallelizing strategy remained unanswered; there was neither experimental evidence nor algorithmic analysis for estimate of efficiency of parallelization in any of these presentations.

B. K. Basu, the coordinator of DST sponsored programme had to state the following on the outcome in [6]- "The Indian machines, however, have not demonstrated scalability clearly and some more effort in this direction is required".

For Flosolver programme of CSIR-NAL a grand support came from NMITLI (New Millennium Indian Technology Leadership Initiative) programme of CSIR in terms of " mesoscale modeling for mansoon related prediction" in 2000 which objective was to build a 1024 processor parallel system and a model software for monsoon forecasting". The project spanned over years during which a communication device, Floswitch, was patented [7] to overcome the communication bottleneck of parallel

computing and a model software VARSHA was developed for monsoon forecasting but the issue of scalability remained unanswered.

It is a truism that if communication bandwidth is infinite, any reasonable parallelizing strategy will be scalable – the limiting factor will be the unparallelized part i.e. sequential component in the application code. But, given the hardware configuration, it is the parallelization strategy which determines what best can be extracted from the hardware.

The present paper aims at critical examination of the parallelization strategy of VARSHA and the alternatives for possible enhancement. To this end, the problem is first stated in mathematical and algorithmic terms and the rationale of evolution of existing parallelization strategy is then outlined. This prepares the background material for exploring the alternative.

2 STATEMENT OF THE PROBLEM

Given the state of the atmosphere at initial instant, using basic principles of physics and its methodology, the problem is to predict its state at future instant. The interval between the initial instant and the next instant is determined by the grid resolution and numerical strategy. The process is repeated till numerical and round off errors swamp the validity of the computations.

As the vertical extent of atmosphere is much smaller than horizontal extent (50 km against 3000 km) hydrostatic approximation is made, i.e. acceleration in the vertical is ignored which has been a standard approximation for GCM.

State of atmosphere is described by the quantities(u, v, w, p, ρ, T, q), where u is the velocity in the horizontal plane along the latitude(East-West); v is the velocity in the horizontal plane along the longitude(South-North); w is the vertical velocity; p is the pressure; ρ is the density; T is the temperature and q is the measure of moisture content and the independent variables are (x, y, z, t); x is measured along latitude, y is measured along longitude, z is the vertical coordinate and t is the time. In view of hydrostatic approximation, z can be replaced by pressure. Instead of z, $\sigma = \frac{p}{p*}$ is introduced, where p* is pressure on the surface of the earth for the simplicity of developing algorithms.

In this set up, the following form of the governing equations (whose details can be found in [8-9]) are used:

2.1 Governing Equations

Law of mass conservation

For air,

$$\frac{\partial}{\partial t}\ln p_* + \overrightarrow{V_H}\,.\,\nabla_H \ln p_* + \nabla_H\,.\,\overrightarrow{V_H} + \frac{\partial \dot{\sigma}}{\partial \sigma} = 0 \quad (1)$$

For moisture,

$$\frac{\partial q}{\partial t} = S \quad (2)$$

Here S represents the sources and sinks. Source comes from evaporation and sink comes from condensing of water vapour into rain etc. The expression for S comes from phenomenological part of physics.

Horizontal Momentum Equations

$$\frac{d\overrightarrow{V_H}}{dt} = -RT\nabla \ln p_* - \nabla\phi - f\,\overrightarrow{k} \times \overrightarrow{V_H} + \overrightarrow{F} \quad (3)$$

Where ϕ is the geopotential, $\overrightarrow{F}$ represents dissipative process coming from phenomenological part of physics, f is the Coriolis component representing contribution from the earth's rotation, $\overrightarrow{V_H}$ is the horizontal velocity vector =(u, v), and **k** is horizontal vector in vertical.

Vertical Momentum Equation

$$\frac{\partial p}{\partial z} = -\rho g$$

Law of Energy Balance

In terms of potential temperature θ ($= \frac{T}{p^{\kappa}}$), κ= ($\frac{c_p - c_v}{c_p}$), the relation is:

$$\frac{\partial}{\partial t}\ln\theta = \frac{H}{C_p T} \quad (4)$$

Where H is the heating rate per unit mass, C_p is the specific heat at constant pressure. H comes from 'physics' and includes solar radiation, heating due to latent heat released during rain or formation of ice etc.

Equation of state It is p=ρRT, where p represents pressure, T is the temperature, R is gas constant; it is taken appropriately for air and water vapour and their mixture as the situation demands.

2.2 Computational Strategy

For achieving better accuracy in numerical computation spectral technique is adopted, Moreover velocity variables are recast in terms of divergence and vorticity of horizontal velocity components. The equations for computation are as following:

(i) (a) **The continuity equation** is integrated along σ to yield,

$$\frac{\partial}{\partial t}\ln p_* = -\int_0^1 \left(\nabla\cdot\overrightarrow{V_H} + \overrightarrow{V_H}\,.\nabla \ln p_*\right)\partial\sigma \quad (5)$$

Which is approximated in semi-discretised form as

$$\frac{\partial}{\partial t}\ln p_* = -\sum_{k=1}^{K} C_k\Delta_k - \sum_{k=1}^{K} D_k\Delta_k \quad (6)$$

Where $C_k = \overrightarrow{V_k}\,.\nabla \ln p_*;\ D_k = \nabla\cdot\overrightarrow{V_k}$

(b) **The continuity equation for moisture is**

$$\frac{\partial q}{\partial t} = -\vec{V}\cdot\nabla q - \dot{\sigma}\frac{\partial q}{\partial \sigma} + S \quad (7)$$

(ii) Taking divergence of momentum equation,

$$\frac{\partial D_k}{\partial t} = \frac{1}{a\cos^2\phi}\left(\frac{\partial B_k}{\partial \lambda} - \cos\phi\,\frac{\partial A_k}{\partial \phi}\right) - \nabla^2(E_k + \phi_k + RT_{0\kappa}\ln p_*) \quad (8)$$

(iii) Taking curl of momentum equation,

$$\frac{\partial \eta_k}{\partial t} = \frac{-1}{a\cos^2\phi}\left(\frac{\partial A_k}{\partial \lambda} + \cos\phi\ \frac{\partial B_k}{\partial \phi}\right) \qquad (9)$$

Where

$$A_k = \eta U + \left(\frac{RT}{a}\right)\cos\phi\left(\ \frac{\partial}{\partial t}\ln p_*\right) + \dot{\sigma}\ \frac{\partial V}{\partial \sigma} - \cos\phi\ F_\phi \qquad (10)$$

$$B_k = \eta V + \left(\frac{RT}{a}\right)\left(\ \frac{\partial}{\partial \lambda}\ln p_*\right) + \dot{\sigma}\ \frac{\partial U}{\partial \sigma} - \cos\phi\ F_\lambda \qquad (11)$$

It may be noted that A_k and B_k are nonlinear terms which spectral handling need special consideration because of nonlinearity.

(iv) The thermodynamic equation is

$$\frac{\partial}{\partial t}\ln\ \theta = \frac{H}{C_p T}$$

Where H is the heating rate per unit mass and θ is the potential temperature. This is rewritten to give the following equation for temperature:

$$\frac{\partial T}{\partial t} = -\vec{V}\cdot\nabla\,\mathrm{T} + \kappa\mathrm{T}\left(\frac{\partial}{\partial t} + \vec{V}\cdot\nabla\right)\ln p_* + \frac{H}{C_p} - \pi\,\dot{\sigma}\ \frac{\partial}{\partial \sigma}\left(\frac{T}{\pi}\right) \qquad (12)$$

Where

T= $\pi\,\theta$; $\pi = p^{\kappa}$; $\kappa = \left(\frac{C_p - C_v}{C_p}\right)$; and ∇ is the horizontal gradient in the system.

Nonlinear terms create serious difficulties as their spectral form does not lend naturally to algorithm due to nonlinearity, thus nonlinear terms are evaluated in physical domain which amounts to changing the variables from spectral domain to physical domain, doing nonlinear operations in physical domain and getting them transformed back to spectral domain for numerical implementation. This technique was first used by Orszag [10] and since then it has become standard.

3 RATIONALE FOR EVOLUTION OF EXISTING PARALLELIZATION STRATEGY

For the complete problem , the governing equations are (6) to (9) and (12) and the domain of computation is longitude × latitude × height × time. Thus, for time marching of a single step, calculations are carried out in longitude × latitude × height(level) space. Typical values of longitude is 512, that of latitude is 256 and that of level is 16. The dependent variables are represented both in physical and spectral domains – in the spectral domain the wave number J is typically 120. A typical variable F (λ, φ) in spectral domain is represented as

$$\mathrm{F}(\lambda,\varphi) = \sum_n \sum_l F_n^l\ \ P_n^l\ \ (\sin\phi)\ e^{il\lambda} \qquad (13)$$

where P_n^l is associated Legendre function, one of many special functions in mathematical physics, see for e.g. [11], the number of spectral coefficients grow like J² .

The strategy of parallelization is dictated by nonlinear term A_k and B_k in equations (8) and (9) and the summation term is equation (6).

Nonlinearity in A_k and B_k necessitate that for every step of evolution of spectral coefficient, one has to go through the physical domain, thereby, all of the latitudes have to be spanned. This makes looping in latitude inevitable. On the other hand the vertical layers which when computed in a sequential environment suggest that for each latitude, data from vertical layers are to be considered are linked together as a basic unit for further processing. This suggested the following domain decomposition:

longitude × latitude subgroup × vertical levels

and this is what is used in the existing VARSHA software. There is no decomposition in longitude in VARSHA, so, for all practical purposes, the domain of computation is *latitude × levels* and domain decomposition takes place in latitude domain.

A schematic diagram of the domain decomposition is shown in Fig. 1.

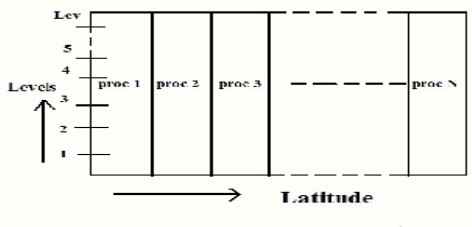

Fig 1: Domain Decomposition in Existing Strategy

For each timestep size of calculation in the existing strategy the spectral coefficients of divergence, vorticity, temperature and moisture for all the vertical levels need to be communicated . This amounts to having communication size of $4J^2 levs$ for each processor.

Though, the above mentioned scheme appears natural, it permits no overlap of computation and communication. When a particular latitude group computation is over, the spectral data need to be globally communicated and unless this is done no further computation can proceed.

Volume of communication in view of global communication is best estimated as

$$\approx 4J^2 \times levs \times Nprc \times \ln_2 Nprc \qquad (14)$$

assuming connectivity like infiniband which is in use in most of the operational HPC platforms.

Here 4 stands for number of spectral variables to be communicated (in this case T, q, D and η; $\ln p_*$ is disregarded for ease of estimate as its size is relatively small). J^2 denotes the size of each spectral variable. The factor *levs* accounts for vertical coupling-all levels are connected. The factor $Nprc \times \ln_2 Nprc$ needs

special mention. It symbolizes tight coupling and global communication. If data of size $4J^2levs$ need to be only sent to $Nprc$ processors, the volume of communication will be $4J^2levs \times Nprc$. But the problem demands that intermediate result accruing from this message passing be again globally communicated. Thus, simple minded strategy will give $Nprc \times Nprc$ which is prohibitive when the number of processors is large. Strategy based on tree structure reduces the factor Nprc to $\ln_2 Nprc$. Use of Floswitch eliminates $\ln_2 Nprc$ factor as it is absorbed in the message processing power of Floswitch. The fundamental essence of scalability is contained in equation (14).

The communication volume is listed below for typical value of Nprc(Number of processors in the parallel computation).

	Existing Strategy	
N	Infiniband Connectivity ($Nprc \times \ln_2 Nprc$)	FloSwitch Connectivity Nprc
2	2	2
4	8	4
8	24	8
16	64	16
64	384	64
128	896	128
256	2048	256
512	4608	512
1024	10240	1024

Table 1: Communication volume in units of $4J^2 \times levs$

It is clear from the above table that the strategy will be made ineffective by the rapid rise of communication demand and it is, therefore not surprising that as the number of processors grew, the large scale parallelization strategy did not succeed in the earlier experiment of VARSHA.

4 PROPOSED NEW STRATEGY

The existing strategy does not make use of the following additional features available in the scheme of computation:

(a) The vertical extent is small, therefore, per horizontal grid point communication in the vertical direction is one order of magnitude smaller.

(b) The coupling among the layers is further weak, it basically occurs only through equation(5), and

(c) Nonlinear part of communication for each latitude is independent of each other.

The above mentioned three features can be exploited to arrive at parallelization strategies for various components viz dynamical, physical and radiation components. The present paper focuses on the dynamical component which permeates most in the governing equations. It may be remarked that physical process and radiation effect enter only through two empirical terms viz F in equation (3) and H in equation (4). Thus it will suffice if dynamical component is presented in detail. The strategy for 'dynamical component of the software' exploiting the above mentioned features has its root in [12]. In the new scheme the number of processors has to be multiple of number of levels. The simplest configuration is when number of processors equals the number of levels as shown in Fig. 2.

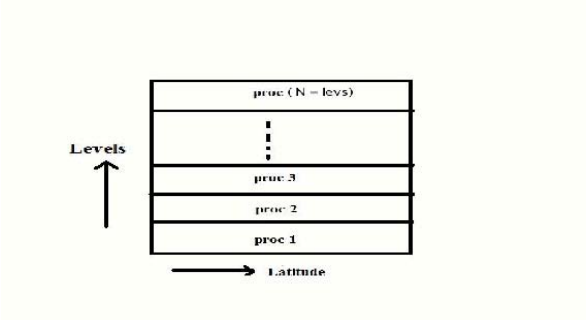

Fig 2: Domain Decomposition in Proposed Strategy (simple case)

In this case , a processor is associated with every level for computing its state for which it has all the necessary data except A_k and B_k as given in (8) and (9). Every processor needs a small chunk of data in form of A_k and B_k from all the remaining processors for computing the interlevel interaction term. To compute such interaction terms a separate processor which will be termed Vertical Integrator is associated. Each processor sends the necessary data to the vertical integrator and gets the processed interaction term from vertical integrator. In view of remarks(c) calculation over each latitude is independent so while vertical Integrator computes the interaction term the calculation of following latitudes can proceed in a normal way.

The above strategy critically depends on the overlapping and communication in the way described above. To assess this, a large number of computational experiment was made on CSIR-4PI 360TF HPC platform for the above mentioned configuration of 120 mode, 256 latitudes, 512 longitudes and 16 levels and the following conclusion emerged: On computational platform of CSIR-4PI 4 latitudes of computing has sufficient computational volume to overlap computation and communication without any overhead. This limits the number of processors to be employed gainfully to 64× 16 = 1024 which is a significant improvement over the previous attempt of limiting number of processors at 152[12]. Processor configuration for this case is shown in Fig. 3.

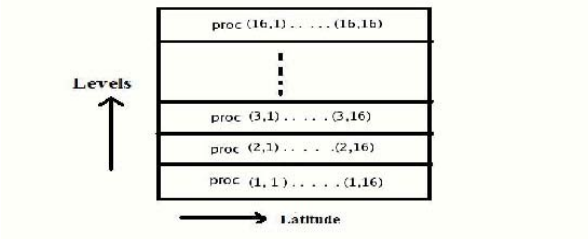

Fig 2: Domain Decomposition in Proposed Strategy (General case)

Communication overhead
$=4J^2\left(\frac{Nprc}{Nlevs}\right)\times \ln_2\left(\frac{Nprc}{Nlevs}\right)$

The effectiveness of new strategy can be easily seen from the expression (14) and (15).
Analysis of communication overhead is central theme of the paper and at the risk of being repetitive, it is shown both in tabular and graphical form. Table 2 presents the various overhead in tabular form. The smaller numbers appearing in column of proposed strategy presents a striking contrast. One may arguably enquire if the contrast is so dominant why was it not discovered earlier. This has to do with the complexities of the application code, too many details clouding the central theme of computation and the built in biasing of the earlier operational sequential code. Fig. 4. Shows the same scenario where the rate of growth of overhead with the number of processor is easily visualized.

	Existing strategy	Existing strategy	Proposed Strategy
Nprc	Infiniband Connectivity $Nprc \times \ln_2 Nprc$	FloSwitch Connectivity $Nprc$	Infiniband Connectivity $\left(\frac{Nprc}{Nlevs^2}\right) \times \ln_2\left(\frac{Nprc}{Nlevs}\right)$
16	64	16	0
32	160	32	0.125
64	384	64	0.5
128	896	128	1.50
256	2048	256	4
512	4608	512	10
1024	10240	1024	24

Table 2: Communication Overhead

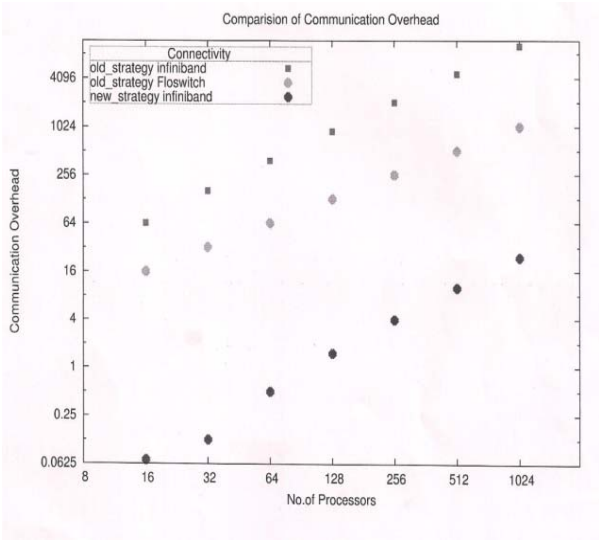

Fig. 4. Communication Overhead for different strategies

5 Conclusion

For a massive parallel computing superiority of hardware both in computing power and communication power are critical, but for a given application the parallelization strategy limits the extent to which the power the power can be exploited. The paper has focused on one such large application code namely GCM code of meteorological computing, Varsha, though developed in-house did not provide scalable parallel computing. The asymptotic analysis suggests a different parallelization strategy which will scale up to a thousand processors instead of existing smaller bunch of processors such as 16.

Acknowledgement

It is our pleasure to thank Director, NAL and Scientist in-charge CSIR-4PI for their approval, support and encouragement.

References

[1] U. N. Sinha, M. D. Deshpande and V. R. Sarasamma, "FLOSOLVER: A parallel computer for fluid Dynamics", *Current Science,* Vol 57, No 23 1277-1285, Dec 5, 1988.
[2] Vijay p . Bhatkar, Ashok V. Joshi, Anirban Basu, Ashok K.Sharma, "Advanced Computing", Tata MCGraw-Hill, New Delhi, pp over 750, 1991.
[3] A.J.Basu, R. Narasimha and U. N. Sinha, "Direct numerical simulation of the initial evolution of a turbulent axissymmetric wake", *Current Science,* Vol 63, No 12, 734-740, Dec 1992.
[4] U. N. Sinha, V. R. Sarasamma, S. Rajalakshmy, K. R. Subramanian,P. V. R. Bharadwaj, C. S. Chandrashekar, T. N. Venkatesh, R. Sunder, B. K. Basu, Sulochana Gadgil and A. Raju, "Monsoon Forecasting on Parallel Computers", *Current Science*, Vol 67, No 3, 178-184, August 1994.
[5] U. N. Sinha, R. S.Nanjundiah, "A decade of parallel meteorological computing on the Flosolver", *Proceedings of the seventh ECMWF workshop on the use of parallel processors in meteorological*, pp 449-460, Nov 2-6, 1996.
[6] B.K. Basu, "Usability of parallel processing computers in numerical weather prediction", Current Science, Vol 74, No 6, 508-516, March 1998.
[7] Sinha U N, Sarasamma V R , Rajalakshmy S and Venkatesh T N, "A Device for Scalable Inter-nodal Communication in a Parallel Computing System", Indian Patent No 2088242007.
[8] U. N. Sinha et al., NMITLI project on mesoscale modeling for monsoon related Predictions, High level document for the new model, prepared by NAL,IISC,TIFR, 2003.
[9] Holton J, An Introduction to dynamic meteorological, Academic Press, 1972.
[10] Orszag S.A., Transform method for calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. Journal of Atmospheric Sciences, Vol 27, 890-895, 1970.
[11] George E. Andrews, Richard Askey , Ranjan Roy, "Special Functions" , Encyclopedia of Mathematics and its Applications 71, Cambridge University Press, 2010.
[12] U. N. Sinha, Md Mahfooz Sheikh, "Efficient Parallelization Strategies ", CSIR, pp 55, Annual Report 2011-12.