

Multidimensional Static Block Data Decomposition for Heterogeneous Clusters

Alexey Kalinov and Sergey Klimov

Institute for System Programming of Russian Academy of Sciences,
25, Bolshaya Kommunisticheskaya str., Moscow 1090045, Russia,
{ka,sergey}@ispras.ru

Abstract. We propose general static block and block-cyclic heterogeneous decomposition of multidimensional data over processes of parallel program mapped onto multidimensional process grid. The decomposition is compared with decomposition of two-dimensional data over two-dimensional process grid of Beaumont et al and with natural decomposition of three-dimensional data over three-dimensional process grid.

1 Introduction

A lot of parallel algorithms are based on homogeneous static block or block-cyclic decomposition of multidimensional data over multidimensional process grid. Those algorithms provide perfect load balancing for homogeneous parallel systems. But the load balancing that can be achieved using those algorithms on heterogeneous parallel systems is not good enough. In this paper we address to heterogeneous cluster consisting of processors of different performance interconnected with homogeneous communication equipment. The most common example of such systems is local networks in dedicated mode.

For a lot of algorithms mapping processes into multidimensional grid is superior to one-dimensional grid. In these cases the situation with mapping processes into multidimensional grid and subsequent distributing data over the process grid is much more difficult. It is proved in [1] that in the case of two-dimensional process grid the optimal solution for the problem is NP-complete. So, for multidimensional process grid we are forced to use heuristic solution in any case.

The decomposition proposed in this paper is further development of natural multidimensional data decomposition proposed in [2]. More advanced algorithms of processes mapping into process grid and data distribution over it are proposed.

The rest of the paper is organized as follows. In Section 2 we discuss problem of multidimensional heterogeneous static block-cyclic data decomposition. In section 3 we introduce a heuristic solution of the problem. In Section 4 we compare the proposed multidimensional decomposition with decomposition of 2D data over 2D process grid of Beaumont et al [1] and with natural decomposition of 3D data over 3D process grid [2].

2 Problem of multidimensional block-cyclic decomposition on heterogeneous clusters

We call mDnD data decomposition a pair ξ, ζ where ξ specifies the *mapping* of processes onto m-dimensional (mD) grid and ζ specifies the *distribution* of n-dimensional (nD) data over the process grid. In this paper we consider only the case one process per processor and thus we don't distinguish process and processor.

Let set of processes is logically considered as m -dimensional process grid with sizes of edges e_0, e_1, \dots, e_{m-1} respectively. Mapping ξ assigns to each process its composite index $\rho = \{\rho_0, \dots, \rho_{m-1}\}$, $\rho_k \in [0, e_k)$ in the grid.

1D1D homogeneous block-cyclic distribution partitions 1D data space into blocks of size a and distribute these blocks in a cyclic manner along the e processes. This means, that data element k is stored in process $\lfloor (k-1)/a \rfloor \bmod (e)$. In other terms, 1D1D block-cyclic distribution is partition of 1D data space onto *generalized* blocks of size $s = a \cdot e$, which in its turn is distributed over e processes. So, the task of block-cyclic distribution can be considered as task of block distribution of generalized block. In heterogeneous case 1D1D distribution is parameterized also by set $R = \{r_i\}$, $i \in [0, e)$ of process performances and amount of data of generalized block distributed on a process depend on R .

The m -dimensional block-cyclic distribution can be regarded as combination of m 1D1D block-cyclic distributions applied to dimensions of n -dimensional data space ($n \geq m$). The distribution with block size $a_0 \times a_1 \times \dots \times a_{m-1}$ partitions the data space of size $N_0 \times N_1 \times \dots \times N_{n-1}$ into generalized blocks of size $s_0 \times s_1 \times \dots \times s_{n-1}$, where $s_i = a_k \cdot e_k$ if k -th 1D distribution is applied to i -th dimension of data and $s_i = N_i$ otherwise. Each generalized block in its turn is partitioned into $e_0 \cdot e_1 \cdot \dots \cdot e_{m-1}$ blocks with size depending in common case on process performances. Such definition of m -dimensional block-cyclic distribution introduces "true" grid when every process has border only with one neighbor in all axis directions. This minimizes communication overheads but leads to imbalance in computational load.

Let the 1D1D distribution corresponding to k -th dimension of process grid is applied to η_k dimension of data grid. On process with composite index ρ is distributed block of data with volume $V_\rho = l_{0,\rho_0}^{\eta_0} \cdot \dots \cdot l_{m-1,\rho_{m-1}}^{\eta_{m-1}} \cdot s_{\eta_m} \cdot \dots \cdot s_{\eta_{n-1}}$ where

$\eta = \{\eta_0, \dots, \eta_{n-1}\}$ is permutation of numbers $\{1, \dots, n-1\}$ and $\sum_{\rho_k=0}^{e_k-1} l_{k,\rho_k}^{\eta_k} = s_{\eta_k}$.

So, the distribution ζ is specified by set $\{s_i\}$, $i \in [0, n)$ and sets $\{e_k\}$, $\{\eta_k\}$, and

$\{l_{k,\rho_k}^{\eta_k}\}$, $l_{k,\rho_k}^{\eta_k} \in N$, $\sum_{\rho_k=0}^{e_k-1} l_{k,\rho_k}^{\eta_k} = s_{\eta_k}$, $k \in [0, m)$. Figure 1 presents a 2D3D data distribution with $\eta_0 = 1$, $\eta_1 = 0$.

Let process with composite index ρ has performance p_ρ . Then time of the block processing gets from the formula $t_\rho^{\xi,\zeta} = V_\rho/p_\rho$, time of parallel blocks processing is determined by $\max_\rho(t_\rho^{\xi,\zeta})$, and objective of the task of data decomposition can be formulated as

$$\text{Objective 1} = \min_{\xi, \zeta} [\max_{\rho} (t_{\rho}^{\xi, \zeta})].$$

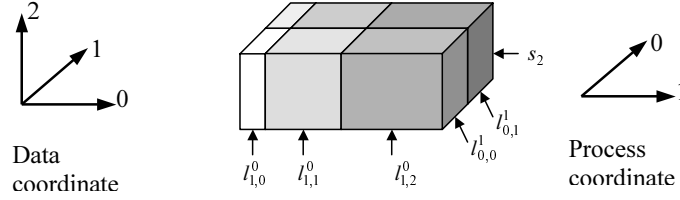


Fig. 1. 2D distribution of 3D generalized block. 1D1D distribution along first dimension of process grid is applied to zero dimension of the data ($\eta_0 = 1$) and 1D1D distribution along zero dimension of process grid is applied to first dimension of data ($\eta_1 = 0$).

3 Heuristic solution

For simplicity we separate determination of mapping ξ and distribution ζ and use the following objective

$$\text{Objective 2} = \min_{\xi} \{ \min_{\zeta} [\max_{\rho} (t_{\rho}^{\xi, \zeta})] \}.$$

3.1 Mapping of processes onto process grid

We propose three heuristics for process mapping. First one is natural heuristics *NAT* introduced in [2]. Let total amount of processes is $E = e_0 \cdot \dots \cdot e_{m-1}$, set of processes $\{p_j\}$ is sorted in ascending order according to process performances and $e_0 \leq e_1 \leq \dots \leq e_{m-1}$. According to natural mapping, processes are mapped onto grid in column-wise order that is j -th process has the following coordinates in the grid:

$$\rho_k = \left\lfloor \frac{j - \sum_{l=k+1}^{m-1} \rho_l \cdot \prod_{i=0}^{l-2} e_i}{\prod_{i=0}^{k-1} e_i} \right\rfloor, \quad k \in [0, m).$$

Natural mapping is good enough for relatively “homogeneous” heterogeneous networks. For heterogeneous networks with heterogeneity essentially shifted to field of weak processes (for example, 1, 10, 11, 12, 13, ...) natural mapping leads to overloading of weak processes and to under loading of powerful ones. For such network we propose modification of natural mapping *NAT1*. Informally this modification can be introduced in the following way. First, we fill according to natural mapping all hyperplanes passing through grid node with coordinates $(0, \dots, 0)$. After that we fill with natural mapping the rest of the process grid.

More formal description is following. On i -th $i \in [1, m]$ step of mapping we select $(m - 1)$ -dimensional process grid of size $\{\tilde{e}_0, \dots, \tilde{e}_{m-2}\}$: $\{\tilde{e}_0 = e_0, \dots, \tilde{e}_{m-i-1} = e_{m-i-1}, \tilde{e}_{m-i} = (e_{m-i+1} - 1), \dots, \tilde{e}_{m-2} = (e_{m-1} - 1)\}$ such that, $\rho_0 \in [0, e_0), \dots, \rho_{m-i-1} \in [0, e_{m-i-1}), \rho_{m-i} = 0, \rho_{m-i+1} \in [1, e_{m-i+1}), \dots, \rho_{m-1} \in [1, e_{m-1})$. Processes are mapped onto this process grid according to natural mapping starting from process with lowest performance that was not mapped on the previous steps. On $m + 1$ step of algorithm processes are mapped according to natural mapping onto reminder - m -dimensional process grid of size $\tilde{e}_0 = (e_0 - 1), \dots, \tilde{e}_{m-1} = (e_{m-1} - 1)$ such that $\rho_0 \in [1, e_0), \rho_1 \in [1, e_1), \dots, \rho_{m-1} \in [1, e_{m-1})$. Figure 2 presents sequence of steps for 3D *NAT1* mapping. Numbers on I-IV refers to process subgrids onto which processes are mapped on the steps.

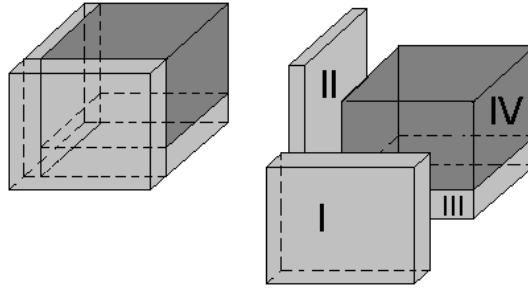


Fig. 2. Sequence of steps for 3D *NAT1* mapping. Numbers on I-IV refers to process subgrids onto which processes are mapped on the steps.

Modification *NAT2* for case of heterogeneity shifted to the field of powerful processes is symmetric to *NAT1*. The only difference is that process set is sorted in descending order and mapping is started with the most powerful process.

Table 1 presents mappings according to mentioned heuristics of processes with performance 1, 2, 3, 4, 5, 6, 7, 8, and 9 onto 2D grid 33.

We propose to solve optimization problem with all three mappings and choose the best one. *Table 1. Mapping of processes with performance 1, 2, 3, 4, 5, 6, 7,*

NAT1	NAT	NAT2
3 7 9	3 6 9	7 3 1
2 6 8	2 5 8	8 4 2
1 4 5	1 4 7	9 6 5

8, and 9 onto 2D grid 33 according to NAT1, NAT, and NAT2.

3.2 Distribution of data over process grid

On this stage we have fixed mapping ξ and we are looking for optimal ζ .

Let $Sl_{k,i}$ is set of processes with composite index $\rho : \rho_0 \in [0, e_0), \dots, \rho_{k-1} \in [0, e_{k-1}), \rho_k = i, \rho_{k+1} \in [0, e_{k+1}), \dots, \rho_{m-1} \in [0, e_{m-1})$ and $f_{k,i}^{\xi,\zeta} = \max_{\rho \in Sl_{k,i}} (t_{\rho}^{\xi,\zeta})$, $k \in [0, m), i \in [0, e_k)$. Suppose $l_{k,i}^{\eta_k} \in R$. The requirement of local minimum of function $\max_{\rho} (t_{\rho}^{\xi,\zeta})$ at ζ_0 in this case is

$$\forall \{k \in [0, m), i \in [0, e_k)\} \Rightarrow f_{k,i}^{\xi,\zeta_0} = \text{const}(\zeta_0)$$

We propose this requirement as objective for solving optimization problem of data distribution and propose to solve this optimization problem for every k independently. So, we try to reach equality of all $f_{k,i}^{\xi,\zeta}$ for every $k \in [0, m)$ independently.

To reach equality of $f_{k,i}^{\xi,\zeta}$ we solve task of moving of borders between blocks of data distributed on $Sl_{k,i}$ on the force of difference $f_{k,i}^{\xi,\zeta} - f_{k,i+1}^{\xi,\zeta}$ similarly to moving partition between two volumes of gas on the force of pressure difference in them. It is the physical analogy that is original for the proposed distribution.

As first approximation for optimization problem we use natural distribution introduced in [2]. According to this distribution $l_{k,i}^{\eta_k}$ are computed using the formula:

$$l_{k,i}^{\eta_k} = \frac{\sum_{\rho \in Sl_{k,i}} p_{\rho}}{\sum_{\rho} p_{\rho}} \cdot s_{\eta_k}, \quad k \in [0, m), \quad i \in [0, e_k).$$

After optimization $l_{k,i}^{\eta_k}$ are rounded. The sum $\sum_{i=0}^{e_k-1} l_{k,i}^{\eta_k} = s_{\eta_k}$ may be less than s_{η_k} . In that case $l_{k,i}^{\eta_k}$ with greater difference ($s_{\eta_k} \cdot \sum_{\rho \in Sl_{k,i}} p_{\rho} - l_{k,i}^{\eta_k} \cdot \sum_{\rho} p_{\rho}$) are iteratively incremented to achieve equality.

4 Experimental results

Proposed data decomposition was compared with 2D2D data decomposition of Beaumont et al [1] and with 3D3D natural decomposition introduced in [2].

4.1 Comparison with 2D2D decomposition of Beaumont et al

For 2D2D case we have conducted two computational experiments. As factor of comparison we use the ratio of time of computation with proposed decomposition to time of computation with decomposition of Beaumont et al. The size of generalized block is 1000x1000. A factor characterizing heterogeneity of the network we use *heterogeneity level* computed as the ratio of maximal to minimal process performance values ($\max_j p_j / \min_j p_j$). Figure 3 (a) presents plot of this factor against heterogeneity level (axis X) and size of square process grid (axis Y). Every point of the plot is computed as average value of the factor computed

from 32000 random variants of the network with the heterogeneity level and the size. Presented results shows that data decomposition of Beaumont et al is better then proposed decomposition (ratio greater then 1) only in limited region. Of course, it is better for size equal to 2 when it is proved optimal solution (the maximal benefit is 1,01762 in case 2x2 grid and heterogeneity 16). Figure 3 (b) presents plot of the average times ratio against the heterogeneity level (axis X) and different process grid: 1x36, 2x18, 3x12, 4x9, 6x6 consisting of 36 processes (axis Y). One can see that for essentially different sizes of 2D process grid and low heterogeneity results are practically the same but in the remainder proposed decomposition is a bit better. We for purpose choose the case of 36 processes because for 6x6 grid the both decompositions have advantage over other. It is interesting to see what are the results provided by the both decomposition for different variants. The heterogeneity level $\max_j p_j / \min_j p_j$ does not fully characterize performance heterogeneity. For the estimation of “distribution of heterogeneity” of heterogeneous network we introduce two functions:

$$Fhet_min = -\frac{1}{(E-1)} \ln \left(\frac{\left(\min_j p_j \right)^E}{p_0 \cdot \dots \cdot p_{E-1}} \right),$$

$$Fhet_max = -\frac{1}{(E-1)} \ln \left(\frac{\left(\max_j p_j \right)^E}{p_0 \cdot \dots \cdot p_{E-1}} \right).$$

Former characterizes shift of heterogeneity to the field of lower performances and the latter characterizes shift of heterogeneity to the field of higher performances. We call *internal heterogeneity* of the network the value $\max(Fhet_min, Fhet_max)$.

Let examine variant – grid 6x6 and heterogeneity level equal to 2 with a bit better average value for decomposition of Beaumont et al. Figure 4 presents ratio of time of computation with the both decompositions to time of computation with ideal decomposition against internal heterogeneity of the network for that variant. Time of ideal decomposition is computed as $\sum_\rho V_\rho / \sum_\rho p_\rho$. One can see that proposed decomposition has is less spread in results than decomposition of Beaumont et al. We did not inspect time of computation of data decompositions. But for all cases except 2x2 process grid proposed decomposition is computed faster then the decomposition of Beaumont et al and for not squire grids it is several orders faster.

4.2 3D3D case

For 3D3D we examine efficiency of proposed decomposition relative to ideal and natural ones. Figure 5 presents plots of average ratio of time of computation with

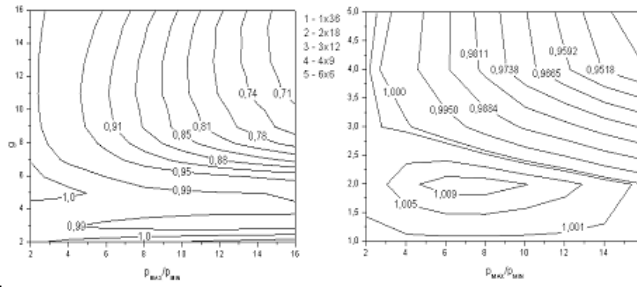


Fig. 3. The average ratio of time of computation with proposed decomposition to time of computation with decomposition of Beaumont et al against the ratio of maximal and minimal process performance values and (a) – size of square process grid, (b) – different variants of process grid consisting of 36 processes (1 – 1x36, 2 – 2x18, 3 – 3x12, 4 - 4x9, 5 – 6x6)

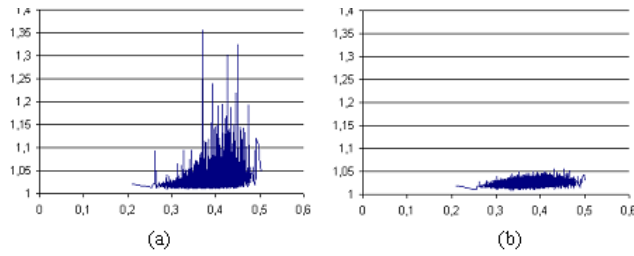


Fig. 4. Ratio of time of computation with (a) – decomposition of Beaumont et al and (b) – proposed decomposition to time of computation with ideal decomposition against internal heterogeneity of the network for grid 6x6 and heterogeneity level equal to 2

proposed decomposition to time of computation with (a) ideal decomposition and (b) natural decomposition against heterogeneity level (axis X) and size of square process grid (axis Y). Every point of the plot is computed as average value computed from 1000 random variants of the network with the heterogeneity level and the size. The figure 5(a) shows that proposed decomposition is essentially worse than ideal one in region of small networks. The figure 5(b) shows that for high heterogeneity level proposed decomposition essentially better than natural one.

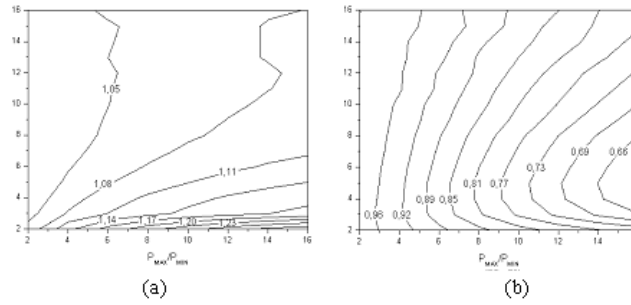


Fig. 5. The average ratio of time of computation with proposed decomposition to time of computation with (a) – ideal decomposition and (b) – natural decomposition against the ratio of maximal and minimal process performance values and size of square process grid

5 Conclusion

We proposed general heterogeneous block data decomposition of multidimensional data over multidimensional process grid that is further development of the simplest general heterogeneous multidimensional decomposition - natural block data decomposition. We showed that proposed decomposition in general is better than specialized decomposition of two-dimensional data over two-dimensional process grid of Beaumont et al. We also showed for three-dimensional case that proposed decomposition in most cases is close to ideal one and that it is much better than natural one in the case of high heterogeneity level.

References

- [1] Olivier Beaumont, Vincent Boudet, Antoine Petit, Fabrice Rastello, and Yves Robert: A Proposal for a Heterogeneous Cluster ScaLAPACK (Dense Linear Solvers). IEEE Trans. Computers. Vol.50, 10 (2001) 1052-1070

- [2] Y.Dovolnov, A.Kalinov, and S.Klimov: Natural Block Data Decomposition for Heterogeneous Clusters. Proceedings of HCW'03, IEEE CS Press, Nice, France, 22 April 2003