Determination of the optimal shape of matrix elements partitioning on three abstract heterogeneous processors

Ye.G Klyuyeva, V.V. Yavorskij, A.A. Adamov & I.T. Utepbergenov |

Published online: 25 May 2020.

Submit your article to this journal ↗

Article views: 312

View related articles ↗

View Crossmark data ↗

Citing articles: 1 View citing articles ↗

*Corresponding author: I.T. Utepbergenov, Innovation End Smart Technologies, Institute of Information and Computing Technologies MES Republic of Kazakhstan, Pushkin, Almaty 051310, Kazakhstan
E-mail: i.utepbergenov@turan-edu.kz

COMPUTER SCIENCE | RESEARCH ARTICLE

Determination of the optimal shape of matrix elements partitioning on three abstract heterogeneous processors

Ye.G Klyuyeva[1], V.V. Yavorskij[2], A.A. Adamov[2] and I.T. Utepbergenov[3]*

Abstract: The paper presents the results of a study done to find the optimal shapes of matrix element partitioning on three abstract heterogeneous processors when performing multiplication operations. An abstract processor model allows applying the research results in systems with different heterogeneous architectures. To determine the optimal partitioning shape, the work uses non-rectangular candidate shapes identified by Ashley DeFlumere in her work as a result of applying the technology of redistribution of matrix elements between the processors «push»: Square Corner, Rectangle Corner, Square Rectangle, Block Rectangle, L-Rectangle, Traditional 1D Rectangular. The optimality of shapes is determined for four classes of matrix multiplication algorithms: Serial Communication with Barrier (SCB), Parallel Communication with Barrier (PCB), Serial Communication with Bulk Overlap (SCO) and Parallel Communication with Overlap (PCO). The Hockney model was used

ABOUT THE AUTHORS

Ye.G. Klyuyeva Senior Lecturer, Karaganda State Technical University, Kazakhstan. Research interests: parallel computing, databases, computer networks, software development.

V.V. Yavorskij Doctor of Technical Sciences, Professor, Karaganda State Technical University, Kazakhstan. Research interests: mathematical modeling, optimization, intelligent systems.

A.A. Adamov Doctor of Technical Sciences, Professor, Head of the Department of Mathematical and Computer Modeling, L.N. Gumilyov Eurasian National University, Research interests: mathematical modeling, optimization, development of information systems.

I.T. Utepbergenov Doctor of Technical Sciences, Professor, Head of Laboratory of Innovative and Smart Technologies Institute of Information and Computational Technologies CS MES of RK, Kazakhstan. Research interests: ACS, Information technology in education, telecommunications, management, transport.

PUBLIC INTEREST STATEMENT

The paper presents the results of a study done to find the optimal shapes of matrix element partitioning on three abstract heterogeneous processors when performing multiplication operations. An abstract processor model allows applying the research results in systems with different heterogeneous architectures. To determine the optimal partitioning shape, the work uses non-rectangular candidate shapes identified by Ashley DeFlumere as a result of applying the technology of redistribution of matrix elements between the processors «push»: Square Corner, Rectangle Corner, Square Rectangle, Block Rectangle, L-Rectangle, Traditional 1D Rectangular. The optimality of shapes is determined for four classes of matrix multiplication algorithms: Serial Communication with Barrier, Parallel Communication with Barrier, Serial Communication with Bulk Overlap and Parallel Communication with Overlap. The Hockney model was used to evaluate the communication complexity of algorithms. Mathematical models of the algorithm execution time were introduced in the paper for each considered candidate shape in all algorithms.

I.T. Utepbergenov

cogent · engineering

to evaluate the communication complexity of algorithms. Mathematical models of the algorithm execution time were introduced in the paper for each considered candidate shape in all algorithms. Based on the developed mathematical models, software was developed that allows to select the form of elements partitioning between processors, depending on the ratio of their speeds and latency of the transmission medium.

Subjects: Computer Mathematics; Mathematical Modeling; Algorithms & Complexity; Software Engineering & Systems Development; Databases

Keywords: parallel computing; parallel matrix multiplication; data partitioning; heterogeneous parallel systems; hockney model; serial communication with barrier; parallel communication with barrier; serial communication with overlap; parallel communication with overlap

1. Introduction

The amount of processed information and the complexity of computational tasks grow every year. In this regard, the field of parallel data processing research became more and more popular all over the world.

Parallel multiplication of matrices is one of the most popular tasks in various fields of scientific knowledge. In particular, such tasks may include work with connectivity matrices and correspondences in information, engineering and transport networks that control the main information about objects and flows movement continuity and at the same time has significant length and dynamic elements.

According to the state of the Top500 list as of November 2019, the maximum computing speed of 200 petaflops was reached by the Summit supercomputer developed by IBM for Oak Ridge National Laboratory. Summit is used in astrophysics, cancer research and systems biology; it consists of 4,608 servers (crosspoints) and 2 397 824 cores.

If at the initial stage of development of high-performance computing, the top lines of the rating were mainly occupied by supercomputers built on identical elements, now we can trace a steady trend towards greater efficiency of heterogeneous cluster solutions.

While the models and methods of distributing computing resources when multiplying matrices on homogeneous systems are studied well enough, the search for optimal solutions for heterogeneous systems is still being studied.

The partitioning of matrix elements among computational elements when performing linear algebra operations is performed for the purpose of reducing the total computation time.

At the first stages of scientific research in this area, rectangular data partitioning shapes were defined and studied in detail (Beaumont et al., 2001; Clarke et al., 2012; Lastovetsky, 2007). Later works (DeFlumere, 2014; DeFlumere & Lastovetsky, 2014) showed that non-rectangular shapes of matrix elements partitioning can also be optimal for certain ratios of the computational power of processors. But these works did not take into account the difference in the latency of data transmission networks among computing elements.

The objective of this study is to determine the optimal shapes for matrix partitioning in heterogeneous systems with different latencies between computational elements. A mathematical model is built for four classes of algorithms for parallel matrix multiplication:

- Serial Communication with Barrier (SCB);

- Parallel Communication with Barrier (PCB);

- Serial Communication with Overlap (SCO);

- Parallel Communication with Overlap (PCO).

To construct a mathematical model of computation, the notion of an "abstract processor", considered in work (Zhong et al., 2011, 2012), is used. Earlier, DeFlumere & Lastovetsky (2014) proved that the model of an abstract processor that is mainly based on the volume of communications and calculations, which accurately forecasts the experimental performance of many processors and even entire clusters for matrix calculations.

The work was done as part of research on the project AP05133699 "Research and development of innovative telecommunication technologies using modern cyber-technical means for the city's intelligent transport system" of the Institute of ICT of the MES RK.

2. Materials and research methods

For constructing model studies, several assumptions were done:

- square source matrices A, B and the resulting matrix Cwith dimension N × N of its elements are used;

- matrix multiplication operation is simulated on three abstract heterogeneous processors with fully connected topology;

- computing power of abstract processors P, R and S is determined by the ratio $P_r:R_r:S_r$, where P is the most powerful processor, and $S_r = 1$. The total processing power of the three-processor system will be equal to $T = P_r + R_r + S_r$;

- matrix elements are divided among the processors in proportion to their computing power;

- transmission network latency is β_1 between the processors P and S, β_2 between the processors P and R, β_3 between processors S and R;

- optimality is evaluated for the following candidate shapes Square Corner (SC), Rectangle Corner (RC), Square Rectangle (SR), Block Rectangle (BR), L-Rectangle (LR), Traditional 1D Rectangular (TR) (DeFlumere, 2014, p. 77) (Figure 1).

The communication complexity of the considered algorithms is estimated using the Hockney model:

$$T_{comm} = \alpha + \beta M, \tag{1}$$

where α—latency or message delay, measured in seconds;

β—bandwidth or time to transmit an element, measured in seconds;

M—message size or number of elements to send.

Figure 1. Candidate shapes defined as potentially optimal for three-processor systems.

1) Square Corner; 2) Rectangle Corner; 3) Square Rectangle; 4) Block Rectangle; 5) L-Rectangle; 6) Traditional 1D Rectangular.

✳ cogent ·· engineering
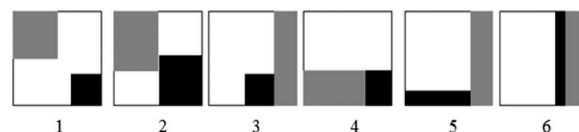
2.1. Finding of the optimal data partitioning for the serial communication with barrier algorithm

The Serial Communication with Barrier algorithm (SCB) is a simple matrix multiplication algorithm in which all data are sent consistently by the processors, and calculations begin only after all processors have completed data transfer and are running in parallel (Figure 2).

The execution time of the algorithm includes the communication time between processors and the time required to perform the calculations:

$$T_{exe} = V\beta + \max\left(c_{P_1}, c_{P_2}, \ldots, c_{P_p}\right), \tag{2}$$

where V—total number of communications;

c_{P_x}- the time required by the processor X to calculate the assigned part of the matrix C.

Because the area of the matrix and the number of elements allocated to each processor are proportional to its computing power, the parameter c_{P_x} will be the same for all candidate forms considered and may not be taken into account when comparing the execution time of the algorithm.

The next step is to define the communication time T_{comm} for each considered form according to the formula (2).

Square Corner:

$$T_{comm} = 2sN\beta_1 + 2rN\beta_2$$

Square Rectangle:

$$T_{comm} = (2N - r)s\beta_1 + (N^2 - s^2)\beta_2 + +(r + s)s\beta_3$$

where r–side of the square, intended for processor R;
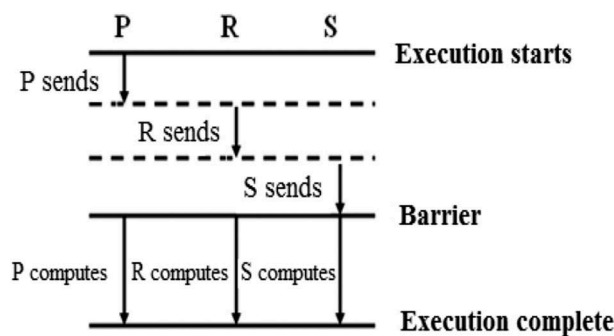
s—side of the square, intended for processor S.

Block Rectangle:

$$T_{comm} = \frac{\beta_1 N^2}{T}\left(P_r - \frac{P_r R_r}{T - P_r} + 1\right) + \frac{\beta_2 N^2}{T}\left(P_r - \frac{P_r}{T - P_r} + R_r\right) + \frac{\beta_3 N^2}{T}(1 + R_r)$$

L-Rectangle:

$$T_{comm} = \frac{\beta_1 N^2}{T}(P_r + 1) + \frac{\beta_2 N^2}{T}\left(P_r + R_r\left(1 - \frac{1}{T - R_r}\right)\right) + \frac{\beta_3 N^2}{T}\left(1 + \frac{R_r}{T - R_r}\right)$$

Figure 2. The serial communication with barrier algorithm for three-processor systems.

Traditional 1D Rectangular:

$$T_{comm} = \frac{\beta_1 N^2}{T}(P_r + 1) + \frac{\beta_2 N^2}{T}(P_r + R_r) + \frac{\beta_3 N^2}{T}(1 + R_r)$$

For the Rectangle Corner shape, the optimal size of R and S will be the combined width of N; that cannot be true, however, based on the classification of candidate forms (DeFlumere & Lastovetsky, 2014, p.78). Alternatively, we can establish that $R_w + S_w = (N - 1)$. Then,

Rectangle Corner:

$$T_{comm} = ((R_h + S_h)(S_w + 1) + N(N - R_h - 1))\beta_1 + ((R_h + S_h)(R_w + 1) + N(N - S_h - 1))\beta_2$$
$$+ ((R_w + S_w)(S_h + R_h - N))\beta_3$$

where R_h, R_w—height and width of the rectangle, intended for processor R;

S_h, S_w—height and width of the rectangle, intended for processor S.

To facilitate the analysis of the mathematical model, we get rid of the coefficient β_3, having obtained the ratio of the throughputs between the computational elements $\frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3}$. Thus, we get four variables.

2.2. Finding of optimal data partitioning for parallel communication with the barrier algorithm

The algorithm of Parallel Communication with Barrier (PCB) assumes that the transfer of data between the processors is carried out in parallel, and that only after the communication is completely completed will the processors begin to perform parallel calculations (Figure 3).

Evaluation of the execution time of the algorithm is performed using the formula:

$$T_{exe} = \max\left(V_{P_1}, V_{P_2}, \ldots, V_{P_p}\right)\beta + \max\left(c_{P_1}, c_{P_2}, \ldots, c_{P_p}\right), \tag{3}$$

where V_{P_x} – the amount of data to be transferred by the processor X.

Similar to the SCB algorithm, the parameter c_{P_x} may not be taken into account in the formulas; the optimality of the forms will be determined based on the communication time T_{comm}.
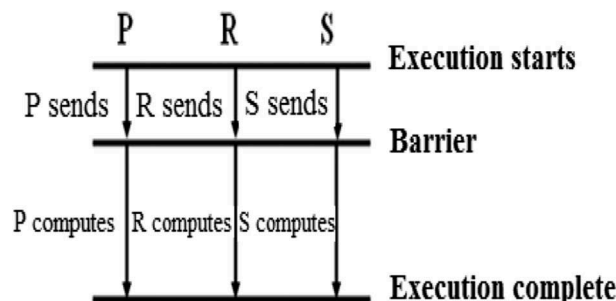
Square Corner:

$$T_{comm} = max\left(2s\beta_1(N - s) + 2r\beta_2(N - r), 2s^2\beta_1, 2r^2\beta_2\right)$$

Square Rectangle:

$$T_{comm} = max\left((2sN - 2s^2 - sr)\beta_1 + (N^2 - Nr - s^2)\beta_2, Nr\beta_2 + sr\beta_3, 2s^2\beta_1 + s^2\beta_3\right)$$

Figure 3. The parallel communication with barrier algorithm for three-processor systems.

Block Rectangle:

$$T_{comm} = \max\left(\frac{\beta_1 N^2 P_r}{T}\left(1 - \frac{R_r}{T - P_r}\right) + \frac{\beta_2 N^2 P_r}{T}\left(1 - \frac{1}{T - P_r}\right), \frac{N^2}{T}\beta_1 + \frac{N^2}{T}\beta_3, \frac{R_r N^2}{T}\beta_2 + \frac{R_r N^2}{T}\beta_3\right)$$

L-Rectangle:

$$T_{comm} = max\left(\frac{N^2 P_r}{T}\beta_1 + \frac{N^2 P_r}{T}\beta_2, \frac{N^2 R_r}{T}\beta_2\left(1 - \frac{1}{T - R_r}\right) + \frac{N^2 R_r}{T(T - R_r)}\beta_3, \frac{N^2}{T}\beta_1 + \frac{N^2}{T}\beta_3\right)$$

Traditional 1D Rectangular:

$$T_{comm} = max\left(\frac{N^2 P_r}{T}\beta_1 + \frac{N^2 P_r}{T}\beta_2, \frac{N^2 R_r}{T}\beta_2 + \frac{N^2 R_r}{T}\beta_3, \frac{N^2}{T}\beta_1 + \frac{N^2}{T}\beta_3\right)$$

Rectangle Corner:

$$T_{comm} = max(2S_w S_h \beta_1 + S_w(S_h + R_h - N)\beta_3, 2R_w R_h \beta_2 + R_w(S_h + R_h - N)\beta_3, (N(N - R_h - 1)$$
$$- S_w S_h + S_w R_h + S_h + R_h)\beta_1 + (N(N - S_h - 1) - R_w R_h + R_w S_h + S_h + R_h)\beta_2)$$

2.3. Finding of the optimal data partitioning for the serial communication with overlap algorithm

In the algorithm of the Serial Communication with Overlap (SCO), all data are transmitted by processor serially, and in parallelelementsnot engaged in the exchange are calculated.

The rest part of the operations is done only after completion of communication and overlap of computations (Figure 4).

The algorithm execution time is calculated by the formula:

$$T_{exe} = \max\left(\max\left(V\beta, O_{P_1}\right) + c_{P_1}, \max\left(V\beta, O_{P_2}\right) + c_{P_2}, \ldots, \max\left(V\beta, O_{P_p}\right) + c_{P_p}\right), \tag{4}$$

where O_{p_x} is time spent by processor X for computing elements that do not require communication;
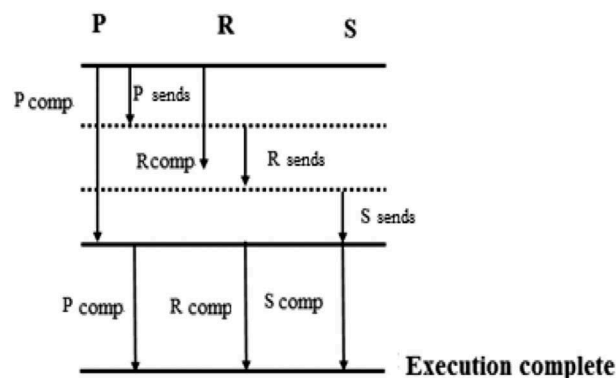
C_{p_x} is time spent by processor X for computing the rest elements.

Let us determine the execution time of the algorithm T_{exe} for each shape considered according to formula (4).

Algorithm execution time for the shape of Square Corner is determined as follows:

$$T_{com} = 2SN + 2$$

Figure 4. The serial communication with overlap algorithm for three-processor systems.

$$O_p = \frac{N \cdot (N - r - s)^2}{S_p}$$

$$Texe_1 = T_{comm} + C_p = 2SN\beta_1 + 2rN\beta_2 + \frac{2N}{S_p}\left(Nr - r^2 - \frac{r^2}{\sqrt{R_r}} + \frac{N_r}{\sqrt{R_r}} - \frac{r^2}{R_r}\right)$$

$$Texe_2 = O_p + C_p = \frac{N}{S_p}\left(N^2 - \frac{r^2}{R_r} - r^2\right)$$

$$Texe_3 = T_{comm} + C_R = 2SN\beta_1 + 2rN\beta_2 + r^2 \cdot \frac{N}{S_r}$$

Algorithm execution time for the shape of Square Rectangle equals to:

$$Texe = T_{com} + \max(C_P, C_R, C_S)$$

$$Texe = (2N - r)S\beta_1 + (N^2 - S^2)\beta_2 + (r + S)S\beta_3 + \max\left(\frac{P_r N^3}{TS_p}, \frac{R_r N^3}{TS_r}, \frac{N^3}{TS_s}\right)$$

Since matrix elements are distributed among the processors proportionally to their computational power, $\frac{P_r N^3}{TS_p}, \frac{R_r N^3}{TS_r}$ and $\frac{N^3}{TS_s}$ values will have the same value for all partitioning shapes, and in the rest formulas, it will be indicated as C_P, C_R, C_s, respectively.

Algorithm execution time for the shape of Block Rectangle equals to:

$$Texe = T_{com} + \max(C_P, C_R, C_s)$$

$$Texe = \frac{\beta_1 N^2}{T}\left(P_r - \frac{P_r R_r}{T - P_r} + 1\right) + \frac{\beta_2 N^2}{T}\left(P_r - \frac{P_r}{T - P_r}\right) + \frac{\beta_3 N^2}{T}(1 + R_2) + \max(C_P, C_R, C_s)$$

Algorithm execution time for the shape of L-Rectangle is determined as:

$$Texe = T_{com} + \max(C_P, C_R, C_S)$$

$$Texe = \frac{\beta_1 N^2}{T}(P_r + 1) + \frac{\beta_2 N^2}{T}\left(R_r + R_r \cdot \left(1 - \frac{1}{T - R_r}\right)\right) + \frac{\beta_3 N^2}{T}\left(1 + \frac{R_r}{(T - R_r)}\right) + \max(C_P, C_R, C_S)$$

Algorithm execution time for the shape of Traditional 1D Rectangular is presented as follows:

$$Texe = T_{comm} + \max(C_{P_1}, C_{P_2}, C_{P_3})$$

$$Texe = \frac{\beta_1 N^2}{T}(P_r + 1) + \frac{\beta_2 N^2}{T}(P_r + R_2) + \frac{\beta_3 N^2}{T}(1 + R_r) + \max(C_{P_1}, C_{P_2}, C_{P_3})$$

For the shape of Rectangle Corner, R and S optimal size is a combined width N, which cannot be true based on the classification of candidate shapes (DeFlumere, 2014, p.78). As alternative, let us state that $R_w + S_w = (N - 1)$. Then,

$$Texe = ((R_h + S_h)(S_w + 1) + N(N - R_h - 1)) \cdot \beta_1 + ((R_h + S_h) \cdot (R_w + 1) + N(N - S_h - 1)) \cdot \beta_2$$
$$+ ((S_h + R_W)(S_h + R_h - N)) \cdot \beta_3 + \max(C_{P_1}, C_{P_2}, C_{P_3})$$

To simplify the analysis of obtained mathematical model, let us get rid of β_3 coefficient having received the ratio of bandwidths between computational elements $\frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3}$. Thus, we get four variables.

2.4. Finding of the optimal data partitioning for the parallel communication with barrier algorithm

Algorithm of Parallel Communication with Overlap (PCO) completes all communications in parallel and simultaneously calculates any sections of the C matrix that do not require interprocessor communication. As soon as the operation data are completed, the rest computations are done (Figure 5).

Evaluation of the execution time of the algorithm is performed using the formula:

$$T_{exe} = \max\left(\max\left(T_{comm}, O_{P_1}\right) + c_{P_1}, \max\left(T_{comm}, O_{P_2}\right) + c_{P_2}, \ldots, \max\left(T_{comm}, O_{P_p}\right) + c_{P_p}\right), \tag{5}$$

where T_{comm} is equal to the communication time for the algorithm of the parallel communication with barrier PCB;

O_{P_x} is time spent by the processor X for computing the elements that do not require communication;

c_{P_x} is time spent by the processor X for computing the rest elements.

Let us construct computation models for each candidate shape according to the formula 5.

Square Corner:

$$Texe_1(T_{comm} + C_P) = \max[2r(N - r) \cdot \beta_2 + (2S(N - S)) \cdot \beta_1], 2r^2 \cdot \beta_2, 2S^2 \cdot \beta_1)$$
$$+ \frac{2r(N - r) + 2S(N - r - S)}{\frac{S_p}{N}}$$

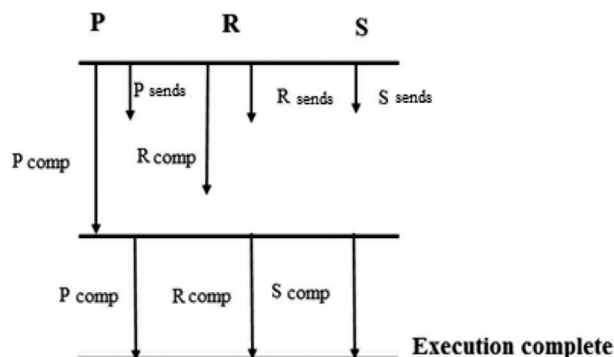$$Texe_2 = (O_P + C_P) = \frac{N^2 - r^2 - s^2}{\frac{S_p}{N}}$$

$$Texe_3(T_{comm} + C_R) = \max\left([2r(N - r) \cdot \beta_2 + 2S(N \cdot S) \cdot \beta_1], 2r^2\beta_2, 2S^2\beta_1\right) + \frac{r^2}{\frac{S_r}{N}}$$

Square Rectangle:

$$Texe_1 = (N^2 - Nr - S^2)\beta_2 + (2SN - 2S^2 - Sr) \cdot \beta_1 + C_P$$

$$Texe_2 = Nr \cdot \beta_2 + Sr \cdot \beta_3 + C_r$$

Figure 5. The parallel communication with overlap algorithm for three-processor systems.

cogent··engineering

$$Texe_3 = S^2 \cdot \beta_3 + 2S^2 \cdot \beta_1 + C_c$$

Block Rectangle:

$$Texe_1 = \frac{P_r N^2 \cdot \beta_1}{T}\left(1 - \frac{R_r}{T - P_r}\right) + \frac{P_r N^2 \beta_2}{T}\left(1 - \frac{1}{T - P_r}\right) + C_P$$

$$Texe_2 = \frac{N^2}{T} \cdot \beta_1 + \frac{N^2}{T} \cdot \beta_3 + C_S$$

$$Texe_3 = \frac{R_r N^2}{T} \cdot \beta_2 + \frac{R_r N^2}{T} \cdot \beta_3 + C_R$$

L-Rectangle:

$$Texe_1 = \frac{P_r N^2}{T} \cdot \beta_1 + \frac{P_r N^2}{T} \cdot \beta_2 + C_P$$

$$Texe_2 = \frac{R_r N^2}{T} \cdot \beta_2 \cdot \left(1 - \frac{1}{T - R_r}\right) + \frac{R_r N^2}{T(T - R_r)} \cdot \beta_3 + C_R$$

$$Texe_3 = \frac{N^2}{T} \cdot \beta_1 + \frac{N^2}{T} \cdot \beta_3 + C_S$$

For the shapes of Traditional 1D Rectangular and Rectangle Corner, to evaluate the total duration of execution, only time of communication between the elements is needed.

Traditional 1D Rectangular:

$$T_{comm1} = \frac{P_r N^2}{T} \cdot \beta_1 + \frac{P_r N^2}{T} \cdot \beta_2$$

$$T_{comm2} = \frac{R_r N^2}{T} \cdot \beta_2 + \frac{R_r N^2}{T} \cdot \beta_3$$

$$T_{comm3} = \frac{N^2}{T} \cdot \beta_1 + \frac{N^2}{T} \cdot \beta_3$$

Rectangle Corner:

$$T_{comm2} = 2S_w S_h \cdot \beta_1 + S_w(S_h + R_h - N) \cdot \beta_3$$

$$T_{comm3} = 2R_h R_w \cdot \beta_2 + R_w(S_h + R_h - N) \cdot \beta_3$$

$$T_{comm1} = (N(N - S_h - 1) - R_w R_h + S_h R_w + S_h + R_h) \cdot \beta_2 + (N(N - R_h - 1) - S_w S_h + R_h S_w + S_h + R_h) \\ \cdot \beta_1$$

By analogy with the SCO algorithm let us get rid of the β_3 coefficient, having received ratio of bandwidths between computational elements $\frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3}$.

3. Research results

Based on the constructed mathematical model, software was developed in the Java programming language that in text form or via the web interface allows entering the initial parameters (P, R, $\frac{\beta_1}{\beta_3}$, $\frac{\beta_2}{\beta_3}$) and determining the theoretical optimal data partitioning shape (Figures 6–8).

Figure 6. Class diagram of "parallel matrix computation on heterogeneous processors" web-application.
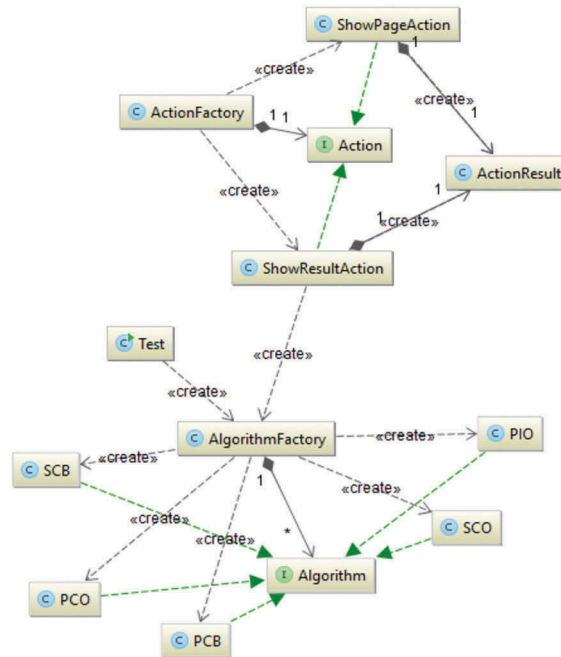

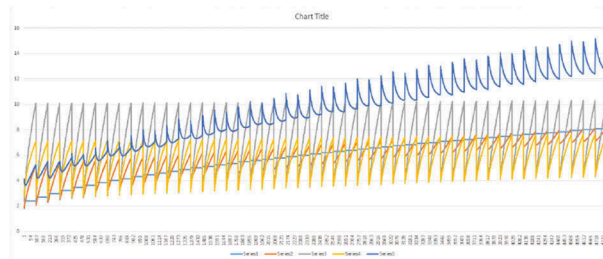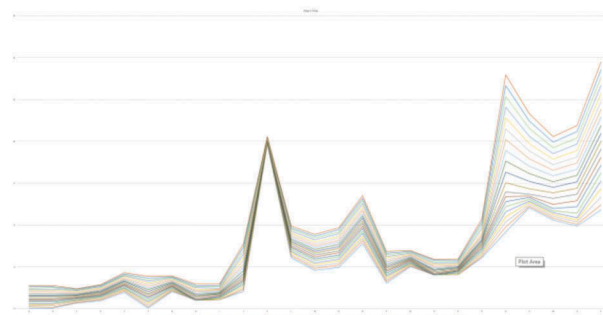**Figure 7. Example of Program Computations for SCB algorithm with different input values of Pr, Rr, $\beta_1, \beta_2, \beta_3$.**

Figure 8. Example of program computations for SCB and PCB algorithms with different input values of *Pr, Rr,* $\beta_1, \beta_2, \beta_3$.

Y axis—execution time; X axis—number of experiments. Series 1—Square Corner; Series 2—Rectangle Corner; Series 3—L Rectangular; Series 4—Block 2D Rectangular; Series 5—Square Rectangle.

Y axis—execution time; X axis—number of experiments. Series shows different types of element partition.

Based on the analysis performed on the basis of calculations obtained according to the constructed mathematical model, the following conclusions were made.

BR, LR, SC and SR partition shapes are optimal for the Serial Communication with Barrier Algorithm. The LR partition shape is optimal at a value of $\frac{\beta_1}{\beta_3}<1$ and roughly equal powers of the processors P and R, while significantly exceeding the power of the processor S. The SR form is optimal in cases of small values of the coefficients $\frac{\beta_1}{\beta_3},\frac{\beta_2}{\beta_3}$ and the powers of the processors P and R, significantly exceeding the power of the processor S. In all other cases, the BR and SC forms are optimal.

As in the case of using the SCB algorithm, only BR, LR, SC and SR forms can be optimal for Parallel Communication with the Barrier Algorithm. LR is optimal only for values of the coefficients $\frac{\beta_1}{\beta_3} = 0,1$, $\frac{\beta_2}{\beta_3}<1$ and roughly equal the powers of processors P and R while being much larger than S. SR is optimal for $\frac{\beta_1}{\beta_3} \leq 2$ and the computing power values of processors P and R are larger than those in the processor S. In other cases, depending on the selected parameters, the forms BR and SC are optimal.

For the Serial Communication with Overlap Algorithm in the same way as in previous algorithms, optimal matrix partitioning shapes may be BR, LR, SC and SR. SR and SC shapes may beoptimal-withthe P processor power values that significantlyexceedpower of the processors R and S. At the same time, for the SR shape, this statement is true only with the coefficient value of $\frac{\beta_1}{\beta_3}\geq2$. The LR shape is optimal with the coefficient value of $\frac{\beta_1}{\beta_3}<1$ and approximately equal power of the processors P and R that significantly exceeds the power of the processor S. In all other cases, optimal shapes are BR.

For the Parallel Communication with Barrier Algorithm, optimal partitioning shapes may be also only BR, LR, SC and SR. The LR shape is only with the coefficient value of $\frac{\beta_1}{\beta_3}< 0.6$, $\frac{\beta_2}{\beta_3} = <1$ and approximately equal power of the processors P and R. The SR shape may be optimalwith $\frac{\beta_1}{\beta_3} \leq 2$ andpower values of processors P and R that significantly exceed the power of the processor S. The SC shape is optimal with the power of the processor P that significantly exceeds the power of the processors R and S and values $\frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3} \geq 1$. In other cases, depending on the chosen parameters, optimal shapes are BR.

4. Conclusion

Based on the results of obtained calculation, it can be concluded that, as in the case of three-processor systems with the same bandwidth between computational elements, the Rectangle Corner and Traditional 1D Rectangular data partitioning shapes are optimal for any set of parameters.

In contrast to the results obtained in matrix (DeFlumere, 2014, p. 85, 88), the L-Rectangle form is optimal with the powers of the processors P and R greatly exceeding the power of the processor S, and approximately equalling each other in the case of $\frac{\beta_1}{\beta_3}<1$ for the SCB algorithm and $\frac{\beta_1}{\beta_3} = 0,1$ for the PCB algorithm. However, in these cases, there is no sense in using a low-power processor S, and the task can be solved on two processors.

The Square Rectangle shape is optimal for small values of the coefficients $\frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3}$ with the powers of processors P and R significantly exceeding the power of the processor S. With such initial data, the problem can also be solved using two powerful processors.

In other cases, the Square Corner and Block Rectangle forms are optimal in the case of using both algorithms, depending on the values of the coefficients $P_r, R_r, \frac{\beta_1}{\beta_3}, \frac{\beta_2}{\beta_3}$.

L-Rectangle shape may be optimal with approximately equal powers of the processors P and R that significantly exceed the power of the processor S and $\frac{\beta_1}{\beta_3}<1$ for the SCO algorithm, and $\frac{\beta_1}{\beta_3}<$

0.6, $\frac{\beta_2}{\beta_3} = <1$ for the PCO algorithm. But in this case, the problem may be solved by two processors, without engaging the processor S.

Square Rectangle shape may be optimal if both caseswithpower values of the processors P and R that significantly exceed the power of the processor S. With such input data, the problem also may be solved using two powerful processors.

In most cases for both algorithm classes, it is preferable to use the Block Rectangle shape.

The developed software can be used to determine the optimal form of partitioning the matrix elements for certain parameters of a real computing system. The results can be used to implement a specific subject area algorithm, for example, for the analysis of large volumes of data presented in the form of matrices, as well as for the analysis of urban structures in modern metropolitan areas, in particular for determining the reachability between the city districts.

Author details
Ye.G Klyuyeva[1]
E-mail: e.klyueva@kstu.kz
ORCID ID: http://orcid.org/0000-0002-3409-264X
V.V. Yavorskij[2]
E-mail: yavorskiy-v-v@mail.ru
ORCID ID: http://orcid.org/0000-0001-6508-1954
A.A. Adamov[2]
E-mail: adam1955@mail.ru
ORCID ID: http://orcid.org/0000-0001-9515-1263
I.T. Utepbergenov[3]
E-mail: i.utepbergenov@turan-edu.kz
ORCID ID: http://orcid.org/0000-0003-0758-4849
[1] Information Technologies and Security, Karaganda State Technical University, Karaganda, 100027, Kazakhstan.
[2] Mathematical and Computer ModelingI, Eurasian National University Named after L.N. Gumilyov, N. Nazarbaev Ave.56, 010008, Kazakhstan.
[3] Innovation and Smart Technologies, Institute of Information and Computational Technologies CS MES of RK, Pushkin, Almaty 051310, Kazakhstan.

Citation information
Cite this article as: Determination of the optimal shape of matrix elements partitioning on three abstract heterogeneous processors, Ye.G Klyuyeva, V.V. Yavorskij, A.A. Adamov & I.T. Utepbergenov, *Cogent Engineering* (2020), 7: 1769948.

References
Beaumont, O., Boudet, V., Legrand, A., Rastello, F., & Robert, Y. (2001). Heterogeneous matrix-matrix multiplication or partitioning a square into rectangles: Np-completeness and approximation algorithms// Parallel and Distributed processing, 2001. *Proceedings Ninth EuromicroWorkshop* (pp. 298–305). IEEE.
Clarke, D., Lastovetsky, A., & Rychkov, V. (2012). Column-based matrix partitioning for parallel matrix multiplication on heterogeneous processors based on functional performance models. *Euro-par 2011: parallel processing Workshops* (pp. 450–459). Springer.
DeFlumere, A. (2014). *Optimal partitioning for parallel matrix computation on a small number of abstract heterogeneous processors* [PhD thesis]. University College Dublin. P. 161.
DeFlumere, A., & Lastovetsky, A. (2014). Optimal data partitioning shape for matrix multiplication on three fully connected heterogeneous processors, Euro-Par 2014WS, HeteroPar 2014. *Twelfth International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms* (pp. 201–214.
Lastovetsky, A. L. (2007). *On grid-based matrix partitioning for heterogeneous processors//ISPDC*.
Zhong, Z., Rychkov, V., & Lastovetsky, A. (2011). Data partitioning on heterogeneous multicore platforms// Cluster Computing (CLUSTER). *2011 IEEE International Conference* (pp. 580–584). IEEE.
Zhong, Z., Rychkov, V., & Lastovetsky, A. (2012). Data partitioning on heterogeneous multicore and multi-GPU systems using functional performance models of data-parallel applications. *Cluster Computing (CLUSTER), 2012 IEEE International Conference* (pp. 191–199). IEEE.

cogent •• engineering

Cogent Engineering (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com