DOI: 10.25586/RNU.V9187.21.03.P.162 УДК 004.75

А.Н. Шульгин, С.С. Зыкова

МОДЕЛЬ ГЕТЕРОГЕННОГО ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА С ГЛОБАЛЬНЫМ РАСПРЕДЕЛЕНИЕМ НАГРУЗКИ НА ОСНОВЕ ЧАСТНЫХ МНОЖЕСТВ

Рассмотрен тип гетерогенной вычислительной системы, топология которой представляет собой совокупность архитектурно идентичных, но различных по производительности вычислительных блоков. Представлена модель вычислительного процесса в гетерогенной вычислительной среде. Сформулирован принцип глобального распределения вычислительной нагрузки на основе метода частных множеств. Представлены результаты расчетов и сравнительный анализ значений ускорения вычислений для гомогенной и гетерогенной организации параллельных вычислений при различном уровне параллелизма. Определена зависимость эффективности вычислений от изменения соотношения вычислительной мощности информационно-вычислительных блоков гетерогенной вычислительной среды при фиксированном общем количестве вычислителей.

Ключевые слова: гетерогенный, параллельный, вычислительная система, вычислительная нагрузка, ускорение вычислительного процесса.

A.N. Shulgin, S.S. Zykova

MODEL OF A HETEROGENEOUS COMPUTATIONAL PROCESS WITH GLOBAL LOAD DISTRIBUTION BASED ON PRIVATE SETS

A type of a heterogeneous computing system is considered, the topology of which is a set of architecturally identical, but different in performance computing units. A model of a computing process in a heterogeneous computing environment is presented. The principle of global distribution of the computational load is formulated based on the method of partial sets. The results of calculations and a comparative analysis of computational acceleration values for homogeneous and heterogeneous organization of parallel computations at different levels of parallelism are presented. The dependence of the efficiency of calculations on the change in the ratio of the computing power of the information and computing units of a heterogeneous computing environment with a fixed total number of computers has been determined.

*Keywords: hetero*geneous, parallel, computing system, computational load, acceleration of computational process.

Введение

Современные тенденции возрастания скорости обработки информации в реальном масштабе времени, внедрения полностью автономных систем с искусственным интеллектом порождают соответствующие требования к производительности, масштабируемости и энергоэффективности вычислительных средств в составе военной и специальной техники [2]. Гомогенные системы, которые до недавнего времени составляли основу парка мобильных вычислительных систем и бортовых вычислительных комплексов, в силу своей однородности имеют пределы эффективности и гибкости вычислений и, как следствие, не в состоянии в полной мере обеспечить оперативное выполнение большого объема сложных вычислений. Этот недостаток ставит разработчиков информационно-вычислительных средств специального назначения перед необходимостью поиска новых подходов.

Шульгин Альберт Николаевич
кандидат технических наук, преподаватель Военно-космической академии имени А.Ф. Можайского, Санкт-Петербург. Сфера научных интересов: информационно-вычислительные системы и технологии. Автор 30 опубликованных научных работ.
E-mail: alex_grid69@mail.ru

Зыкова Светлана Сергеевна
адъюнкт Военно-космической академии имени А.Ф. Можайского, Санкт-Петербург. Сфера научных интересов: информационно-вычислительные системы и технологии, надежность, живучесть. Автор 4 опубликованных научных работ.
E-mail: Swetlanca.zykova@yandex.ru

Одним из путей решения этой проблемы является использование гетерогенных вычислительных систем.

До недавнего времени в качестве основного признака неоднородности гетерогенных вычислительных систем рассматривались архитектурные различия их вычислительных компонентов. Традиционно в фокусе внимания находилось сочетание процессоров общего назначения и специализированных вычислительных модулей, например, сигнальных, графических процессоров и др. [12]. Однако многопроцессорные системы данного типа в настоящее время активно развиваются. Появляются новые решения, определяющие гетерогенность не только по архитектурному признаку, но и по функциональному [6]. В контексте вышесказанного интерес вызывают гетерогенные системы, в основе которых лежит неоднородность вычислительной мощности составляющих их информационно-вычислительных блоков (ИВБ), а также способы эффективного управления вычислительной нагрузкой.

Глобальное управление нагрузкой как основа эффективных гетерогенных вычислений

В настоящей работе рассмотрен тип гетерогенной вычислительной системы, топология которой представляет собой совокупность архитектурно идентичных, но различных по производительности ИВБ. Такое решение нашло свое воплощение в технологии ARMBig-Little [4]. В данном случае интерес вызывает управление вычислительной нагрузкой, основанное на глобальном распределении задач. В отличие от кластерной организации и принципа переключений внутри ИВБ такое решение обладает большей гибкостью и возможностью эффективного задействования вычислительного потенциала всех ИБВ [11; 13].

Структурно такая гетерогенная вычислительная среда содержит ИВБ малой мощности $\alpha_1, \alpha_2, \ldots, \alpha_g$ и ИВБ большой мощности $\beta_1, \beta_2, \ldots, \beta_h$, вычислительная нагрузка каждого из которых непосредственно управляется планировщиком ОС (рис. 1).

Основным свойством рассматриваемого типа гетерогенной системы является то, что на высокопроизводительных ИВБ задачи выполняются быстрее, чем на ИВБ малой вычислительной мощности, при этом длительность вычислений пропорциональна производительности соответствующих вычислительных блоков. Это свойство необходимо учитывать при решении задачи повышения оперативности сложных неоднородных вычислений. В данном случае для расчета и оценивания вычислительного эффекта может быть ис-

пользован показатель (коэффициент) неоднородности φ гетерогенной вычислительной среды, представляющий собой соотношение производительности ИВБ малой и большой мощности:

$$\varphi = \frac{V_\alpha}{V_\beta}, \qquad (1)$$

где V_α – производительность ИВБ малой мощности; V_β – производительность ИВБ большой мощности.

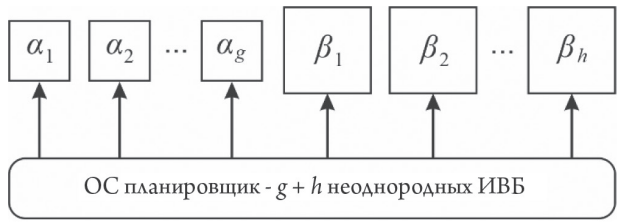

Рис. 1. Гетерогенная мультипроцессорная вычислительная среда с глобальным распределением задач

Формализация гетерогенного вычислительного процесса с глобальным распределением нагрузки на основе частных множеств

Главной характеристикой любого вычислительного процесса, реализующего тот или иной алгоритм, является время его выполнения $T_{алг}$. На временной диаграмме (рис. 2) видно, что $T_{алг}$ представляет собой совокупность n распараллеливаемых участков $L_{пр}$ и m последовательных (нераспараллеливаемых) участков $L_{пс}$. Следует заметить, что каждый участок $L_{пр}$, в свою очередь, состоит из параллельных частей $l_{пр}$ различной вычислительной сложности и, как следствие, различной длительности выполнения.

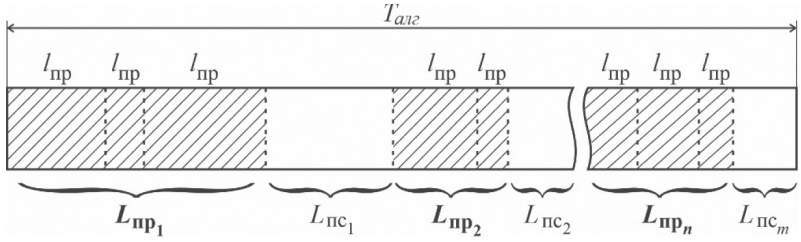

Рис. 2. Временная диаграмма алгоритма с неоднородными параллельными участками

Декомпозиция участков $L_{пр}$ на множество вычислительных модулей позволяет реализовать их параллельное выполнение, обеспечивая тем самым ускорение вычислений [3; 8] на каждом таком участке. Однако в данном случае имеет место неравномерная загрузка ИВБ, а сам распараллеленный участок $L_{пр}$ ограничивается наиболее продолжительной его частью.

На рисунке 3 видно, что участок $L'_{пр}$ выполняемый на ИВБ α_1, α_2 и α_3 определяется его частью $l_{пр1}$, имеющей максимальную длительность из всего множества параллельных частей:

$$\Lambda = \left\{ l_{пр_1}, l_{пр_2}, l_{пр_3} \right\}.$$

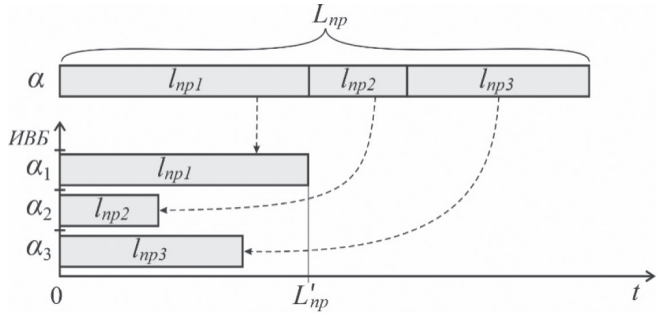

Рис. 3. Параллельный участок алгоритма с неравномерной загрузкой ИВБ

Приведенный выше пример показывает, что в гомогенных вычислительных средах к основному эффекту замедления, обусловленному наличием последовательных частей алгоритма, добавляется наличие параллельных участков большой длительности [1].

В гетерогенной вычислительной среде этот тормозящий фактор может быть преодолен за счет распределения вычислительной нагрузки между ИВБ различной производительности [5; 7]. В данном случае, если основную нагрузку возложить на ИВБ малой вычислительной мощности α, а выполнение последовательных $L_{пс}$ и параллельных частей $l_{пр}$ большой длительности – на β-вычислители, то это позволит получить дополнительный эффект ускорения вычислений как на локальном уровне, так и в масштабе всего алгоритма. Такую организацию вычислений можно реализовать на основе метода формирования частных множеств длительностей задач для планирования их выполнения на соответствующие ИВБ. Далее рассмотрим суть этого метода.

Наличие в параллельном участке $L_{пр}$ нескольких задач большой длительности (рис. 4, *а*) делает необходимым формирование множества для планирования их выполнения на β-вычислители. Однако в этом случае из соображения обеспечения энергоэффективности возникает необходимость определения оптимальных значений длительности соответствующих частей алгоритма и принятия решения планировщиком оцелесообразности их обработки на мощных вычислителях. В качестве основного критерия решения этой задачи можно использовать директивное значение l^d, относительно которого с учетом неоднородности вычислительной системы может быть принято рациональное решение о формировании соответствующего частного множества (рис. 4, *б*). Тогда частное множество Λ^α задач для α-вычислителей и, соответственно, Λ^β – для β-вычислителей определяется следующим образом:

$$\begin{cases} \Lambda^\alpha = \left\{ l_{пр} \in \Lambda \mid l_{пр} \cdot \varphi < l^d \right\} \\ \Lambda^\beta = \left\{ l_{пр} \in \Lambda \mid l_{пр} \cdot \varphi \geq l^d \right\} \end{cases} \qquad (2)$$

где $\Lambda = \left\{ l_{пр_1}, l_{пр_2}, ..., l_{пр_n} \right\}$ – множество длительностей параллельных частей отдельного участка $L_{пр}$; φ – коэффициент неоднородности гетерогенной вычислительной среды (1); l^d – директивное значение длительности параллельной части участка $L_{пр}$.

На рисунке 4, *б* показан принцип вышеизложенного подхода. Здесь задачи, длительности которых обеспечивают выполнение условия $l_{пр} \geq l^d$, идут на формирование частного множества Λ^β, в противном случае формируется множество Λ^α. В результате имеют место две группы задач для параллельного выполнения ИВБ различной производительности.

На основе (2) можно определить предельную длительность каждой параллельной группы, выбрав аргумент с максимальным значением из соответствующего частного множества:

$$L_{пр}^{\alpha} = \max \Lambda^{\alpha};\ L_{пр}^{\beta} = \max \Lambda^{\beta}.$$

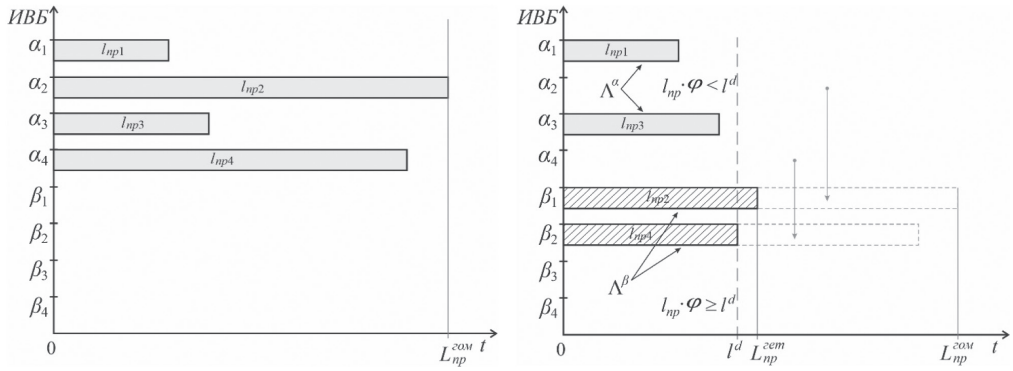

Рис. 4. Параллельный участок алгоритма:
а– с множеством частей большой длительности в гомогенной вычислительной системе;
б – в гетерогенной вычислительной системе с распределением нагрузки на ИВБ различного типа

Очевидно, что при определении величины *i*-го параллельного участка $L_{пр}$ алгоритма определяющим будет значение $L^{\beta}_{пр}$. В случае, если частное множество Λ^{β} пустое, то $L_{пр}$будет ограничен значением $L^{\alpha}_{пр}$:

$$L_{пр_i} = \begin{cases} L_{пр}^{\alpha}, & \text{при } \Lambda^{\beta} = 0 \\ L_{пр}^{\beta} \cdot \varphi, & \text{при } \Lambda^{\beta} \neq 0 \end{cases}$$

Исходя из вышеизложенного длительность $T_{алг}$ выполнения алгоритма в гетерогенной вычислительной среде можно определить как

$$T_{алг} = \sum_{i=1}^{n} L_{пр_i} + \sum_{j=1}^{m} L_{пс_j} \cdot \varphi,$$

где $L_{пр}$ – длительность распараллеленных участков алгоритма; $L_{пс}$ – длительность последовательных участков; φ – коэффициент неоднородности гетерогенной вычислительной среды (1); *n* – количество распараллеленных участков алгоритма; *m* – количество последовательных участков алгоритма.

Оценивание вычислительного эффекта в гетерогенной вычислительной системе с глобаотным распределением нагрузки

Одним из показателей эффективности параллельных вычислений является ускорение выполнения алгоритма. Закономерность роста ускорения при распараллеливании вычислений сформулирована в [9]. Однако данная закономерность предполагает изначальную фиксированность рабочей нагрузки, реализация которой рассматривалась применительно к ИВБ одинакового типа и вычислительной мощности [10].

В контексте данной работы для адекватной оценки ускорения вычислений необходимо учитывать организацию вычислений сложных алгоритмов и гетерогенность вычислительной среды.

В данном случае, если в качестве основного принципа организации вычислений рассматривать выполнение на ИВБ большой мощности всех последовательных операций алгоритма и параллельных операций значительной длительности (определяемому по принципу, изложенному выше), ускорение вычислений $S_{гет}$ алгоритма применительно к гетерогенной параллельной вычислительной среде будет выглядеть следующим образом:

$$S_{гет} = \frac{1}{(\tau_{пс} + \tau_{пр}) \cdot \varphi + \frac{1 - (\tau_{пс} + \tau_{пр}) \cdot \varphi}{k}}, \quad (3)$$

где $\tau_{пс}$ – доля последовательных участковалгоритма; $\tau_{пр}$– доля параллельных участков большой длительности; k – количество ИВБ основной параллельной обработки; φ – коэффициент неоднородности гетерогенной вычислительной среды (1).

Для проверки состоятельности изложенного в данной работе принципа распределения вычислительной нагрузки были проведены расчеты (3) и сравнительный анализ значений производительности для гомогенной и гетерогенной организации параллельных вычислений при различном уровне параллелизма. Результаты расчетов ускорения для гомогенной и гетерогенной вычислительной среды представлены в таблице 1.

Таблица 1

Ускорение вычислительного процесса для гомогенной и гетерогенной вычислительной среды при q = 0,5

	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8
$S_{гом}$	1,00	1,33	1,52	1,60	1,66	1,72	1,75	1,78
$S_{гет}$	1,00	1,60	2,01	2,29	2,50	2,66	2,81	2,92

На рисунке 5 показаны графики зависимостей ускорения вычислительного процесса в гомогенной и гетерогенной вычислительной среде.Очевидно, что предложенный подход позволяет существенно поднять верхнюю границу ограничения на рост производительности при распараллеливании вычислений.

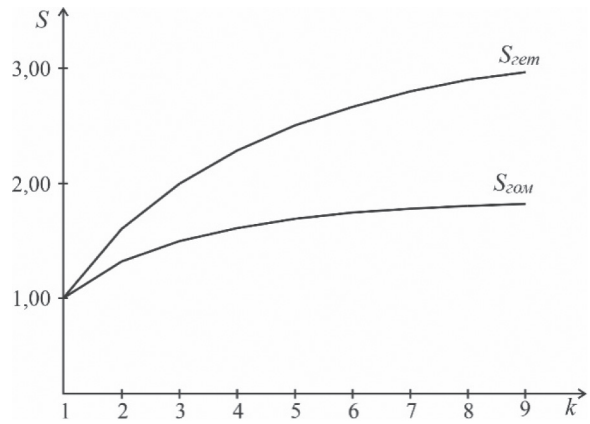

Рис. 5. Ускорения вычислительного процесса в гомогенной $S_{ГОм}$ и гетерогенной $S_{гет}$ вычислительной среде при q = 0,5

В результате исследования выявлена зависимость эффективности вычислений от изменений соотношения вычислительной мощности ИВБ гетерогенной вычислительной среды при фиксированном общем количестве вычислителей. Результаты расчетов представлены в таблице 2.

Таблица 2

Ускорение вычислительного процесса в гетерогенной вычислительной среде при фиксированном количестве ИВБ ($k = 4$) и различных значениях φ

	$q = \tau_{пс} + \tau_{пр}$					
	$q = 0,1$	**$q = 0,2$**	**$q = 0,3$**	**$q = 0,4$**	**$q = 0,5$**	**$q = 0,6$**
$S_{\varphi=1,0}$	3,07	2,50	2,11	1,82	1,61	1,43
$S_{\varphi=0,5}$	3,48	3,07	2,76	2,50	2,28	2,10
$S_{\varphi=0,25}$	3,73	3,48	3,33	3,07	2,94	2,77

На рисунке 6 показаны графики ускорения вычислений при различных значениях коэффициента неоднородности в условиях ограниченной масштабируемости.

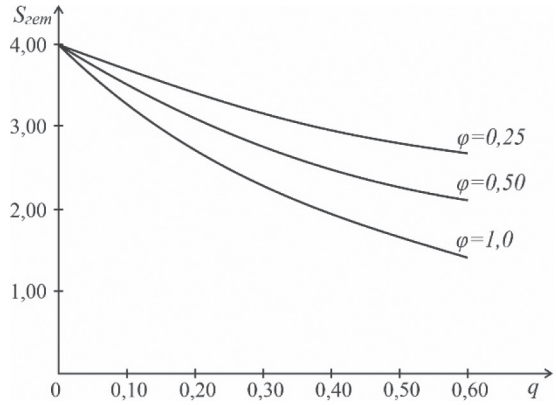

Рис. 6. Соотношение зависимостей *S* от *q* при различных значениях коэффициента неоднородности в гетерогенной вычислительной среде

Очевидно, что при увеличении совокупной доли последовательных участков реализуемого алгоритма и параллельных частей большой длительности прирост ускорения вычислений зависит от неоднородности гетерогенной вычислительной среды и может быть получен за счетувеличения разницы в вычислительной мощности ИВБ.

Выводы

Представленная модель реализации алгоритма позволяет осуществить расчет длительности его выполнения и оценить оперативность вычислений различной сложности в гетерогенной вычислительной среде.

Принцип распределения вычислительной нагрузки в гетерогенной вычислительной системе с глобальным распределением задач, изложенный в данной работе, дает возможность:

– получить дополнительный эффект ускорения параллельного вычислительного процесса при наличии значительной доли последовательных участков и параллельных частей большой длительности;

– выбрать оптимальную топологию гетерогенной вычислительной среды для обеспечения высокой эффективности вычислений в соответствии с объемом и сложностью задач в условиях ограничений на масштабируемость и энергоемкость функционирования.

Литература

1. *Борисов А.А., Краснов С.А., Нечай А.А.* Технология блокчейн и проблемы ее применения в различных информационных системах // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2018. № 2. С. 63–67.

2. *Буренок В.М., Гладышевский В.Л.* Информатика и вычислительная техника: перспективы развития и применения в военном деле / Вооружение и экономика. 2015. № 3 (32). С.17–32.

3. *Воеводин В.В., Воеводин Вл.В.* Параллельные вычисления. СПб.: БХВ- Петербург, 2002. 608 с.

4. *Гринхальг П.* Секреты архитектуры Big-Little // Электронные компоненты. 2012. № 1. С. 104–106.

5. *Нечай А.А., Котиков П.Е.* Методика повышения надежности функционирования систем, организованных на перепрограммируемых элементах // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2016. № 1-2. С. 87–89.

6. *Светличный А.Н.* Краткий обзор достижений в области гетерогенных вычислений // Молодой ученый. 2016. № 1 (105). С. 213–216.

7. *Свинарчук А.А., Нечай А.А.* Использование квантовых вычислений при выборе управленческогорешения // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2018. № 2. С. 31–36.

8. *Шульгин. А.Н., Шушаков А.О.* Исследование влияния методов управления вычислительной нагрузкой мобильных многопроцессорных вычислительных комплексов на их автономность // Труды Военно-Космической академии им. А.Ф. Можайского / под ред. Ю.В. Кулешова. 2019. Вып. 668. С. 64–70.

9. *Amdahl G.M.* (1967) Validity of the single-processor approach to achieving large scale computing capabilities. In AFIPS Conference Proceedings (Atlantic City, N.J., Apr. 18–20). AFIPS Press, Reston, vol. 30, pp. 483–485.

10. *Gustafson J.* (1988) Reevaluating Amdahl's Law. *Communications of the ACM*, vol. 31, no. 5, pp. 532–533.

11. *Hexus (2013)* Tech explained – ARM big.LITTLE Processing, 24 October. Available at: http://hexus.net/tech/techexplained/cpu/48693-tech-explained-arm-biglittleprocessing (Date of the Application: 28.06.2021).

12. *John Shalf* (2007) The New Landscape of Parallel Computer Architecture Journal of Physics: Conference Series, 78.

13. *Peter Clarke* (2013) BenchmarkingARM'sbig-littlearchitecture. Embedded Staff, 6 August.

References

1. Borisov A.A., Krasnov S.A., Nechay A.A. (2018) *Tekhnologiya blokcheyn i problemy ee primeneniya v razlichnykh informatsionnykh sistemakh* [Blockchain technology and problems of its application in various information systems]. *Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie*, no. 2, pp. 63–67 (in Russian).
2. Burenok V.M., Gladyshevsky V.L. (2015) *Informatika i vychislitel'naya tekhnika: perspektivy razvitiya i primeneniya v voennom dele* [Informatics and Computer Engineering: Prospects for Development and Application in Military Affairs]. *Vooruzhenie i ekonomika*, no. 3 (32), pp. 17–32 (in Russian).
3. Voevodin VV, Voevodin Vl.V. (2002) *Parallel'nye vychisleniya* [Parallel Computing]. St. Petersburg, BKhV- Peterburg, 608 c. (in Russian).
4. Greenhalg P. (2012) *Sekrety arkhitektury Big-Little* [Secrets of Big-Little Architecture]. *Elektronnye komponenty*, no. 1, pp. 104–106 (in Russian).
5. Nechay A.A., Kotikov P.E. (2016) *Metodika povysheniya nadezhnosti funktsionirovaniya sistem, organizovannykh na pereprogrammiruemykh elementakh* [Methods for improving the reliability of the functioning of systems organized on reprogrammable elements]. *Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie*, no. 1-2, pp. 87–89 (in Russian).
6. Svetlichny A.N. (2016) *Kratkiy obzor dostizheniy v oblasti geterogennykh vychisleniy* [A brief overview of advances in heterogeneous computing]. *Molodoy uchenyy*, no. 1 (105), pp. 213–216 (in Russian).
7. Svinarchuk A.A., Nechay A.A. (2018) *Ispol'zovanie kvantovykh vychisleniy pri vybore upravlencheskogoresheniya* [Using quantum computing when choosing a management decision]. *Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie*, no. 2, pp. 31–36 (in Russian).
8. Shulgin. A.N., Shushakov A.O. (2019) *Issledovanie vliyaniya metodov upravleniya vychislitel'noy nagruzkoy mobil'nykh mnogoprotsessornykh vychislitel'nykh kompleksov na ikh avtonomnost'* [Investigation of the influence of methods for managing the computational load of mobile multiprocessor computing systems on their autonomy]. *Trudy Voenno-Kosmicheskoy akademii im. A.F. Mozhayskogo*, vol. 668, pp. 64–70 (in Russian).
9. Amdahl G.M. (1967) Validity of the single-processor approach to achieving large scale computing capabilities. In AFIPS Conference Proceedings (Atlantic City, N.J., Apr. 18–20). AFIPS Press, Reston, vol. 30, pp. 483–485.
10. Gustafson J. (1988) Reevaluating Amdahl's Law. *Communications of the ACM*, vol. 31, no. 5, pp. 532–533.
11. Hexus (2013) Tech explained – ARM big.LITTLE Processing, 24 October. Available at: http://hexus.net/tech/techexplained/cpu/48693-tech-explained-arm-biglittleprocessing (Date of the Application: 28.06.2021).
12. John Shalf (2007) The New Landscape of Parallel Computer Architecture Journal of Physics: Conference Series, 78.
13. Peter Clarke (2013) BenchmarkingARM'sbig-littlearchitecture. Embedded Staff, 6 August.