Для цитирования: Ким А.К., Перекатов В.И., Фельдман В.М. Центры обработки данных на базе серверов «Эльбрус» // Вопросы радиоэлектроники. 2017. № 3. С. 6–12.
УДК 004.382.7 (075)

А.К. Ким[1,2], В.И. Перекатов[1,2], В.М. Фельдман[1,2]

[1] ПАО «ИНЭУМ им. И.С. Брука», [2] АО «МЦСТ»

ЦЕНТРЫ ОБРАБОТКИ ДАННЫХ НА БАЗЕ СЕРВЕРОВ «ЭЛЬБРУС»

Рассматривается современное состояние проектов, ставящих целью создание систем сверхвысокой производительности на отечественной аппаратно-программной платформе «Эльбрус». Проводится сравнение с аналогичными системами, построенными на зарубежных микропроцессорах. Уточняются детали построения современных центров обработки данных, в частности топология коммуникационной сети, связывающей десятки тысяч вычислительных узлов, организация системы хранения данных, создание высокой плотности элементов в системе с большой потребляемой и рассеиваемой мощностью.

Ключевые слова: *микропроцессор, контроллер периферийных интерфейсов, сервер, вычислительный узел, суперкомпьютер, система хранения данных.*

Введение

В настоящее время, когда в стране намечен к созданию и проектируется ряд высокопроизводительных вычислительных центров обработки данных на базе импортных серверов, возможность построения таких центров на отечественных аппаратно-программных платформах становится особенно актуальной.

На исходе прошлого десятилетия в журнале «Электроника» были опубликованы статьи, рассматривающие вопросы построения суперЭВМ с использованием вычислительных средств на базе микропроцессоров семейства «Эльбрус» [1, 2]. Часть высказанных тогда соображений сейчас можно было бы скорректировать, но наиболее значимые прогнозы сбылись. В частности, это относится к развитию «эльбрусовской» проектной линии. Универсальные микропроцессоры, разработанные сегодня в проектных коллективах АО «МЦСТ» и ПАО «ИНЭУМ им. И.С. Брука», сопоставимы по характеристикам с зарубежными образцами. В качестве примера можно отметить тот факт, что наш проигрыш по производительности «топовым» универсальным микропроцессорам Xeon фирмы Intel укладывается в пределы от 1,5 до 3 раз.

В дальнейшем изложении будет представлен проект создания центра обработки данных на базе отечественных решений, который также может использоваться как суперЭВМ, предназначенная для решения широкого класса задач численного моделирования.

Развитие линии вычислительных средств с архитектурой «Эльбрус»

По сравнению с показателями вычислительных средств, спроектированных ко времени выхода в свет упомянутых выше публикаций, архитектурная линия «Эльбрус» получила существенное развитие, в первую очередь в части наращивания производительности и возможностей ввода-вывода [3]. В табл. 1 приведены характеристики микропроцессоров этой архитектуры, серийно изготавливаемых, находящихся на стадии разработки и планируемых в разработку на ближайшую перспективу.

Повышение производительности микропроцессоров достигается за счет увеличения числа ядер и исполнительных устройств в ядрах, введения поддержки векторных операций и повышения тактовой частоты.

Помимо разработки новых микропроцессоров был завершен проект по проектированию и изготовлению второй модификации контроллера периферийных интерфейсов КПИ-2. Его характеристики в сравнении с предыдущей версией КПИ-1 приведены в табл. 2.

Мировой опыт создания суперЭВМ

За последние годы существенно поменялись характеристики передовых суперкомпьютеров, входящих в список TOP500. Если в 2009 году американский суперкомпьютер Roadrunner, занимающий первое место в классификации TOP500, имел пиковую производительность 1,325 Пфлопс, то в 2016 году ее лидер, китайский суперкомпьютер Sunway TaihuLight System, показывает пиковую производительность 125,4 Пфлопс [4], что почти на два порядка выше.

Показательным является то, что в текущем списке TOP500 первые два места занимают китайские суперкомпьютеры. На втором месте расположился проект Tianhe-2 с пиковой производительностью 54,9 Пфлопс [5].

Таблица 1. Характеристики микропроцессоров с архитектурой «Эльбрус»

Наименование (условное обозначение)	Число ядер		Тактовая частота, МГц	Производительность, Гфлопс	Год выпуска	Технология, нм
	Архитектура «Эльбрус»	Другие				
Эльбрус (1891ВМ4Я)	1	–	300	4,8	2007	130
Эльбрус-2С+ (1891ВМ7Я)	2	1 DSP	500	16 (без DSP)	2011	90
Эльбрус-4С (1891ВМ8Я)	4	–	800	50	2013	65
Эльбрус-1С+ (1891ВМ11Я)	1	2D- и 3D-графика	1000	24 (без GPU)	2015	40
Эльбрус-8С (1891ВМ10Я)	8	–	1300	250	2015	28
Эльбрус-8СВ (1891ВМ12Я)	8	–	1500	512	2018	28
Эльбрус-16С	16	–	2000	До 1500	2020	16
Эльбрус-32С	32	–	–	До 4000	2022	10

Таблица 2. Характеристики контроллеров периферийных интерфейсов

Наименование (условное обозначение)	Интерфейсы	Тактовая частота, МГц	Год выпуска	Технология, нм
КПИ-1 (1991ВГ1Я)	Состав контроллеров: 2 канала ввода-вывода (связь с системой); PCI Express 1.0 8x-1; PCI-32/64, 33/66; Ethernet 10/00/1000; 2 канала IDE; 4 канала SATA; AC ‘97; 2 канала USB; 2 канала RS-232/485; канал IEEE-1284; интерфейсы I2C, SPI, GPIO; системный и сторожевой таймеры; контроллер прерываний	250	2007	130
КПИ-2 (1991ВГ2Я)	Состав контроллеров: канал ввода-вывода (связь с системой); 3 интерфейса Ethernet 10/00/1000; 20 линий интерфейса PCI Express 2.0; интерфейс PCI; 6 каналов SATA-3; 4 канала USB2.0; интерфейс IDE; функциональный блок приема синхросигналов от систем реального времени или от генераторов меток точного времени с обеспечением микросекундной точности определения прихода этих сигналов относительно друг друга; реализация в контроллерах Ethernet 1 Гбит/с международного стандарта IEEE1588 для гарантированной синхронизации времени сетевых ВК с точностью не хуже 1 мс; канал INTEL HDA; канал IEEE-1284; интерфейсы RS-232/485, I2C, SPI, GPIO; канал внешних прерываний; набор таймеров; контроллер управления питанием и режимами энергосбережения SPMC	500	2015	65

Следует также отметить, что если в суперкомпьютере Tianhe-2 для построения вычислительного узла (compute node), структура которого показана на рис. 1, использовались микропроцессоры Ivy Bridge и Xeon Phi фирмы Intel, то в суперкомпьютере Sunway узел построен на микропроцессорах китайской разработки SW26010. Этот факт еще раз подтверждает правильность стратегии импортозамещения в электронике и вычислительной техники, проводимой в настоящее время в России.

Узел Tianhe-2 представляет собой гетерогенную структуру, где три 57-ядерных микропроцессора Xeon Phi выполняют роль мощного ускорителя, позволяющего совместно с двумя универсальными микропроцессорами Ivy Bridge получить суммарную

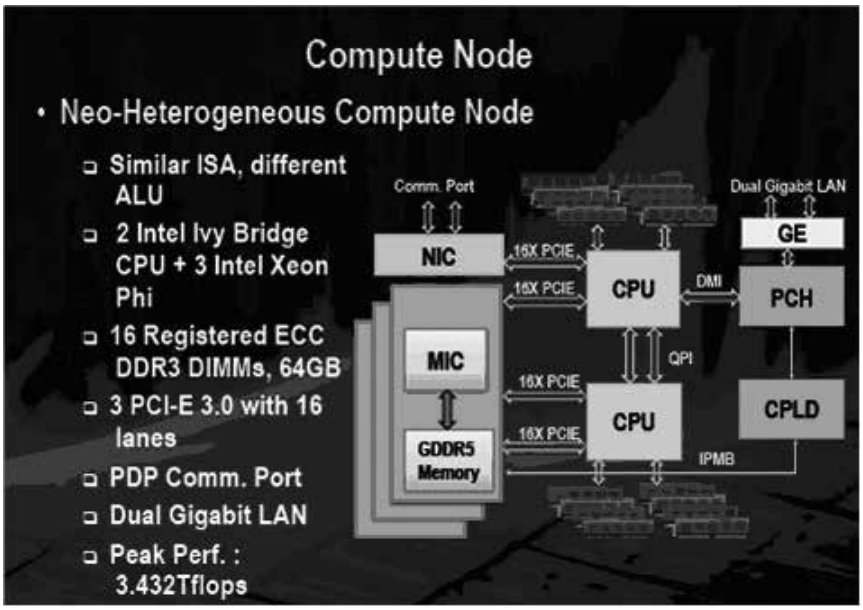

Рисунок 1. Структура вычислительного узла суперкомпьютера Tianhe-2

производительность узла 3,431 Тфлопс. Для создания коммуникационной сети используется интерфейс собственной разработки TH Express-2, на базе которого строится иерархическая структура «толстого дерева». Система охлаждения шкафа, содержащего 128 вычислительных узлов, построена с использованием жидкостного охлаждения. Вся система Tianhe-2, содержащая 125 шкафов, потребляет 17,6 МВт мощности.

Микропроцессор SW26010, являющийся основой суперкомпьютера Sunway, состоит из четырех групп ядер, объединенных внутрикристальным сетевым интерфейсом. Каждая группа ядер, структура которой показана на рис. 2, построена по известному принципу комбинации ядра управляющего процессорного элемента (MPE) и 64 ядер вычислительных процессорных элементов (CPE), представляющих матрицу 8×8 из 64-разрядных RISC-процессоров,

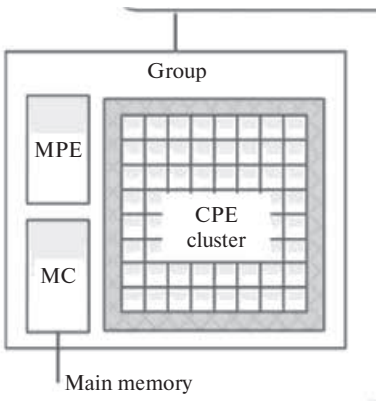

Рисунок 2. Структура группы ядер микропроцессора SW26010

поддерживающих исполнение векторных команд. Кроме того, каждая группа ядер имеет контроллер памяти (MC).

Коммуникационная структура системы Sunway спроектирована под заказ и представляет собой пятиуровневую иерархию, включающую компьютерный узел (два микропроцессора SW26010), компьютерную плату (восемь узлов), суперузел (32 компьютерных платы), шкаф (четыре суперузла), систему (40 шкафов).

Охлаждение в системе Sunway построено на тех же принципах, что и в суперкомпьютере Tianhe-2. Общая потребляемая мощность системы – 15,371 МВт, при этом достигнут рекордный показатель отношения производительности к мощности – 6 Гфлопс/Вт.

Опыт разработки больших информационно-вычислительных систем на микропроцессорах «Эльбрус»

В 2015 году в ПАО «ИНЭУМ им. И. С. Брука» в рамках опытно-конструкторской работы, заданной Министерством промышленности и торговли, была разработана информационно-вычислительная кластерная система на базе четырехъядерного микропроцессора «Эльбрус-4С». Ее основой является четырехпроцессорный сервер, размещаемый в корпусе 19” форм-фактора 1U, внешний вид которого представлен на рис. 3. Производительность сервера – 100 Гфлопс (на операциях двойной точности). 64 сервера размещаются в шкафу (рис. 4) и образуют систему с общей производительностью 6,4 Тфлопс. Мощность, потребляемая шкафом, составляет примерно 20 КВт, что позволяет использовать воздушную систему охлаждения.

Коммуникационная сеть волоконно-оптических связей построена в виде 2D-тора с использованием специального контроллера, подключаемого к каналу ввода-вывода микропроцессора «Эльбрус-4С» [6].

Проект построения центра обработки данных на многоядерных микропроцессорах «Эльбрус-8С»

В 2016 году на базе восьмиядерных микропроцессоров «Эльбрус-8С» [7] разработаны многопроцессорные серверы, которые можно использовать для построения больших вычислительных систем терафлопсной и петафлопсной производительности. Краткое описание проекта такой вычислительной системы приведено ниже.

Масштабируемая серверная система представляет собой совокупность шкафов с серверами и шкафов с системами хранения данных (СХД), объединенных коммуникационной сетью в массово-параллельную вычислительную структуру с возможностью получения производительности до нескольких десятков Пфлопс. При этом число вычислительных узлов в системе практически не ограничено (более 30 000 узлов).

Серверы и СХД строятся на базе российских микропроцессоров серверного класса «Эльбрус-8С» и контролеров периферийных интерфейсов КПИ-2. Восьмиядерный «Эльбрус-8С» с пиковой производительностью 250 Гфлопс на сегодняшний день является самым производительным российским универсальным микропроцессором, сравнимым с мировыми аналогами.

Для объединения серверов предлагается построить коммуникационную сеть со структурой 4D-тор на базе отечественной микросхемы-маршрутизатора ЕС8430 «Ангара» разработки российского предприятия ПАО «НИЦЭВТ» [8, 9]. Другим вариантом построения сети может быть структура 3D-тор с использованием коммутационного кристалла совместной разработки института ВНИИЭФ (г. Арзамас) и НИИСИ РАН. Отечественные сети сопоставимы по своей функциональности, производительности и надежности с современными разработками мировых лидеров в данной области (Cray, IBM, Mellanox, Intel).

Для отвода тепла прорабатывается вариант использования системы прямого жидкостного охлаждения, позволяющей отводить более 400 кВт от стандартного вычислительного шкафа в широком диапазоне температур (что на сегодняшний день является мировым рекордом в отрасли). Технология такой системы разработана российской фирмой АО «РСК Технологии» – ведущим разработчиком и интегратором суперкомпьютерных решений [10].

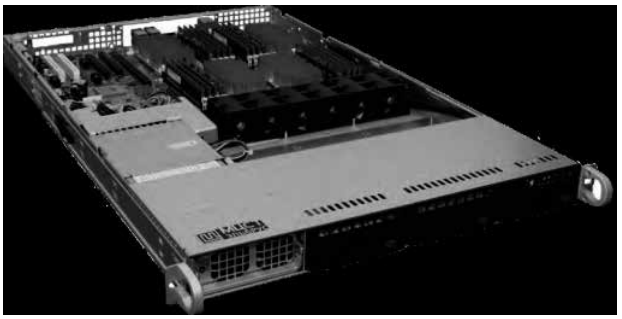

Рисунок 3. Вид четырехпроцессорного сервера на микропроцессоре «Эльбрус-4С»

Рисунок 4. Шкаф вычислительной системы

При построении СХД планируется использование аппаратных и конструктивных решений российской фирмы ООО «Промобит» [11], а также программных решений российской фирмы ООО «Рэйдикс» [12]. Разрабатываемый продукт представляет собой высокопроизводительное горизонтально масштабируемое хранилище данных с распределенными кодами коррекции ошибок, построенное на базе отечественного оборудования и программного обеспечения.

Вычислительный шкаф состоит из:

- четырехпроцессорных серверов до 100 штук с системой жидкостного охлаждения;
- матрицы связей интерконнекта 2D-тор 10×10 с оптическими выводами каналов для объединения шкафов системы в 4D-тор;
- коммутаторов Ethernet 1 Гбит/с 1-го уровня для сети технического обслуживания и менеджмента;
- подсистемы жидкостного охлаждения;
- подсистемы электропитания.

Шкаф СХД состоит из:

- двухпроцессорных модулей серверов до 10 штук с системой воздушного охлаждения;
- дисковых массивов – до 36 дисков в дисковой полке, до 360 дисков в стойке;
- матрицы связей интерконнекта 1D-тор 10 с оптическими выводами каналов для объединения шкафов системы;
- коммутаторов Ethernet 1Gb для сети технического обслуживания и менеджмента 2-го и/или 3-го уровней;
- подсистемы электропитания.

Характеристики шкафов приведены в табл. 3.

Возможности коммуникационной сети в предложенной конфигурации позволяют объединять до 300 шкафов в единую вычислительную систему.

Таблица 3. Характеристики вычислительного шкафа и шкафа СХД

Параметр	Вычислительный шкаф	Шкаф СХД
Число микропроцессоров/ядер «Эльбрус-8С»	400/3200	20/160
Тактовая частота микропроцессоров, ГГц	1,3	1,3
Пиковая производительность, Тфлопс	50 (двойная точность)	5 (одинарная точность) 2,5 (двойная точность)
Объем оперативной памяти, Тбайт	25,6	1,28
Суммарный объем дискового пространства, Пбайт	–	До 0,72
Пиковая пропускная способность дуплексного канала «интерконнекта» для одного узла, Гбайт/с	16	16
Ориентировочная потребляемая мощность, кВт	50	10

Таблица 4. Основные конкурирующие решения (с характеристиками – в расчете на шкаф)

Параметры	ORNL Titan (3-е место в TOP500), США	NUDT Tianhe-2 (2-е место в TOP500), Китай	«Т-Платформы», «Ломоносов-2» (41-е место в TOP500), Россия	«РСК Технологии», «Торнадо», Россия	Предлагаемый проект, Россия
Аппаратная платформа	AMD Optron Interlagos (16 cores, 2,2 GHz) plus NVIDIA Tesla K20x (14 cores, 0,732 GHz)	2 – Intel Ivy Bridge (12 cores, 2,2 GHz) plus 3 – Intel Xeon Phi (57 cores, 1,1 GHz)	Xeon E5–2600 v3, NVIDIA Tesla K40 (SXM)	Xeon E5–2600 v3, Xeon Phi 7200	«Эльбрус-8С»
Использование ускорителей	Да	Да	Да	Да	Нет
Число узлов	96	128	256	153	100
Число процессоров/ядер	192/5760	640/44160	256/3584	306/220 32	400/3200
Пиковая производительность, Тфлопс (двойная точность)	139,2	439,2	420	528	50
Объем оперативной памяти, Тбайт	3,65	11,26	8,19	29,4	25,6
Объем дисковой памяти, Тбайт	80	95	30	60	50
Соотношение производительность/мощность, Тфлопс/кВт	~2	~4	~4	~5	~1

Интегральные характеристики предлагаемой вычислительной системы находятся на таком уровне, что могли бы войти в первую сотню суперкомпьютеров списка TOP500 лучших мировых образцов. При этом следует отметить, что в приведенных для сравнения проектах вычислительные узлы строятся с использованием микросхем-ускорителей.

Сравнительные характеристики проекта приведены в табл. 4.

Основным конкурентным преимуществом и одновременно уникальностью является использование передовых отечественных технологий, лучших в своем классе, а именно:

- выбор отечественных микропроцессоров и контроллеров, что обеспечивает независимость от зарубежных поставок и информационную безопасность;
- использование жидкостного охлаждения для достижения максимальной вычислительной плотности, надежности, снижения на 40–50% расходов на электроэнергию по сравнению с традиционным воздушным охлаждением, возможность установки в неподготовленных помещениях;
- архитектурные решения по построению системы хранения, передовые в классе горизонтально-масштабируемых распределенных систем, обеспечивающих высокую отказоустойчивость и производительность при умеренной стоимости;
- использование интерконнекта с топологией 4D-тор, которая обеспечит распределенный и максимально быстрый обмен данными между вычислительными узлами и системой хранения данных без «узких мест» в виде общего канала обмена между всеми вычислительными узлами и всеми узлами хранения, а также удешевление системы благодаря отсутствию необходимости в дорогостоящих коммутаторах верхнего уровня.

СПИСОК ЛИТЕРАТУРЫ

1. Микропроцессоры и вычислительные комплексы российской компании МЦСТ / В. Волконский, А. Ким, Л. Назаров, В. Перекатов, В. Фельдман // Электроника. 2008. № 8. С. 62–69.
2. Ким А.К., Фельдман В.М. СуперЭВМ на основе архитектурной платформы «Эльбрус» // Электроника. 2009. № 2. С. 74–81.
3. Фельдман В.М. Состояние и перспективы развития отечественной микропроцессорной аппаратно-программной платформы «Эльбрус» // Международная конференция «Микроэлектроника-2015»: Тез. докл. М.: Техносфера, 2015. С. 127–128.
4. Dongarra J. Report on the Sunway TaihuLight System. University of Tennessee Department of Electrical Engineering and Computer Science Tech. Report UT-EECS-16–742. June 20, 2016.
5. Dongarra J. Visit to the National University for Defense Technology Changsha, China. University of Tennessee Department of Electrical Engineering and Computer Science Tech. Report. June 3, 2013.
6. Белянин И.В., Петраков П.Ю., Фельдман В.М. Функциональная организация и аппаратура сетевого взаимодействия модулей в вычислительном кластере на базе микропроцессоров с архитектурой «Эльбрус» // Вопросы радиоэлектроники. 2015. № 3 (3), С. 7–21.
7. Kostenko V., Kozhin A., Polyakov N., Slesarev M., Tikhorskiy V., Sakhin Yu. Elbrus-8C: the Latest Yield from MCST and MIPT Collaboration. 2015 IEEE DOI 10.1109/EnT.2015.24, 2015 International Conference on Engineering and Telecommunication, pp. 67–68.
8. Высокоскоростная сеть «Ангара» для суперкомпьютеров и кластеров – сделано в России [Электронный ресурс]. URL: https://servernews.ru/931123
9. Ежегодный научно-технический семинар GraphHPC / В.В. Воеводин, А.С. Симонов, А.С. Фролов, А.С. Семенов // Computational nanotechnology. 2015. № 1. С. 5–8.
10. РСК – ведущий в России и СНГ разработчик и интегратор суперкомпьютерных решений [Электронный ресурс]. URL: http://www.rscgroup.ru/
11. Сайт компании «Промобит» [Электронный ресурс]. URL: http://bitblaze.ru/
12. Программное обеспечение Raidix [Электронный ресурс]. URL: http://www.raidix.ru/

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ким Александр Киирович, к.т.н., генеральный директор, ПАО «ИНЭУМ им. И.С. Брука», АО «МЦСТ», 119334, Москва, ул. Вавилова, д. 24, 8 (499) 135-33-21, e-mail: kim_a@ineum.ru.

Перекатов Валерий Иванович, д.т.н., профессор, зам. генерального директора, ПАО «ИНЭУМ им. И.С. Брука», АО «МЦСТ», 119334, Москва, ул. Вавилова, д. 24, 8 (499) 135-05-28, e-mail: perekatov_v@ineum.ru.

Фельдман Владимир Маркович, д.т.н., старший научный сотрудник, зам. генерального директора, ПАО «ИНЭУМ им. И.С. Брука», АО «МЦСТ», 119334, Москва, ул. Вавилова, д. 24, 8 (499) 135-61-56, e-mail: feldman_v@ineum.ru.

For citation: Kim A.K., Perekatov V.I., Feldman V.M. Data centers based on «Elbrus» servers. Voprosy radioelektroniki, 2017, no. 3, pp. 6–12.

A.K. Kim, V.I. Perekatov, V.M. Feldman

DATA CENTERS BASED ON «ELBRUS» SERVERS

The current state is considered in the article projects aimed at establishing systems of ultra-high performance on domestic hardware and software «Elbrus» platform. Descripted a comparison with similar systems built on foreign microprocessors. Clarifies details of building modern data centers, in particular, the topology of a communication network linking tens of thousands of computational nodes, storage system, create a high density elements in the system with a large and wasted power consumption.

Keywords: microprocessor, controller peripheral interfaces, server, compute node, the supercomputer, a data-storage system.

REFERENCES

1. Volkonskiy V., Kim A., Nazarov L., Perekatov V., Feldman V. Microprocessors and Computer Complexes of MZST Company. *Elektronika,* 2008, no. 8, pp. 62–69.
2. Kim A.K., Feldman V.M. Supercomputers on the Basis of Architectural Platform «Elbrus». *Elektronika,* 2009, no. 2, pp. 74–81.
3. Feldman V.M. State and development prospects of microprocessor hardware and software platform «Elbrus». *Mezhdunarodnaya konferentsiya «Mikroelektronika-2015»: Tez. dokl.* Moscow, Tekhnosfera, 2015, pp. 127–128.
4. Dongarra J. Report on the Sunway TaihuLight System. *University of Tennessee Department of Electrical Engineering and Computer Science Tech. Report UT-EECS-16–742.* June 20, 2016.
5. Dongarra J. Visit to the National University for Defense Technology Changsha, China. *University of Tennessee Department of Electrical Engineering and Computer Science Tech. Report.* June 3, 2013.
6. Belyanin I.V., Petrakov P. Yu., Feldman V.M. Functional organization and hardware means of network interconnetion of modules in computer cluster based on «Elbrus» microprocessors. *Voprosy radioelektroniki,* 2015, no. 3 (3), pp. 7–21.
7. Kostenko V., Kozhin A., Polyakov N., Slesarev M., Tikhorskiy V., Sakhin Yu. Elbrus-8C: the Latest Yield from MCST and MIPT Collaboration. *2015 IEEE DOI 10.1109/EnT.2015.24, 2015 International Conference on Engineering and Telecommunication,* pp. 67–68.
8. High-speed network «Angara» for supercomputers and clusters-made in Russia. Available at: https://servernews.ru/931123
9. Voevodin V.V., Simonov A.S., Frolov A.S., Semenov A.S. The GraphHPC workshop. *Computational nanotechnology,* 2015, no. 1, pp. 5–8.
10. [RSC – Russia and CIS leading developer and integrator of supercomputer solutions] (In Russ.). Available at: http://www.rscgroup.ru/
11. [Promobit company website] (In Russ.). Available at: http://bitblaze.ru/
12. [Software RAIDIX] (In Russ.). Available at: http://www.raidix.ru/

AUTHORS

Kim Aleksandr, PhD, general director, PJSC «Brook INEUM», JSC «MCST», Moscow, 119334, Russian Federation, Vavilova st., 24, tel.: +7 (499) 135-33-21, e-mail: kim_a@ineum.ru.
Perekatov Valeriy, Dr., professor, deputy general director, PJSC «Brook INEUM», JSC «MCST», Moscow, 119334, Russian Federation, Vavilova st., 24, tel.: +7 (499) 135-05-28, e-mail: perekatov_v@ineum.ru.
Feldman Vladimir, Dr., senior researcher, deputy general director, PJSC «Brook INEUM», JSC «MCST», 119334, Moscow, Vavilova st., 24, tel.: +7 (499) 135-61-56, e-mail: feldman_v@ineum.ru.