

Гибридный суперкомпьютер на базе сети Ангара для задач вычислительного материаловедения*

В.С. Вечер^{1,2}, Н.Д. Кондратюк^{1,2}, Г.С. Смирнов¹, В.В. Стегайлов¹

Объединенный институт высоких температур РАН¹,
Московский физико-технический институт (государственный университет)²

В данной статье представлен суперкомпьютер, состоящий из 32 узлов с графическими ускорителями, соединенных между собой низколатентным интерконнектом с высокой пропускной способностью. Суперкомпьютер нацелен на проведение расчетов методом классической молекулярной динамики, а также может эффективно ускорять расчеты электронной структуры материалов. В статье представлены результаты тестов производительности конкретных прикладных задач на основе пакетов LAMMPS, GROMACS, VASP и CP2K. Проведено сопоставление результатов с другими высокопроизводительными системами. Данный суперкомпьютер основан на сети Ангара, разработанной АО «НИЦЭВТ» и поддерживающей топологии трехмерного и четырехмерного тора. Параллельная масштабируемость задач на суперкомпьютере Десмос иллюстрирует возможности использования сети Ангара для эффективного ускорения MPI приложений на массово-параллельных архитектурах.

Ключевые слова: суперкомпьютер, интерконнект Ангара, тесты производительности, вычислительное материаловедение

1. Введение

Быстрое развитие параллельных вычислительных методов и высокопроизводительных систем предоставляет огромные возможности для методов атомистического моделирования. В настоящее время подобные математические модели и вычислительные коды являются уже не только методами фундаментальных исследований, но и инструментами решения намного более широкого круга проблем [1]. Применение классической молекулярной динамики (МД) в данный момент вышло на масштабы триллионов атомов [2] и миллисекунд моделирования [3] (т.е. 10^9 шагов по времени при типичном для МД шаге в 1 фемтосекунду).

Существуют два основных способа ускорить МД расчет. Первый способ представляет собой использование массово-параллельных систем с распределенной памятью (МРР). Для молекулярно-динамических расчетов наиболее естественным способом распределить вычислительную нагрузку и данные по узлам МРР-системы является разделение по пространству (domain decomposition) [4]. Сопутствующие проблемы балансировки нагрузки могут быть решены при этом очень эффективно [5].

Вторая возможность ускорить расчет состоит в увеличении вычислительных возможностей каждого узла МРР-системы. Использование многопроцессорных и многоядерных узлов с общей памятью обеспечивает существенное ускорение МД-расчета. Тем не менее, масштабируемость систем с общей памятью ограничивается их стоимостью и скоростью доступа в оперативную память для многопроцессорных и/или многоядерных узлов. Производительность систем с общей памятью становится значительно выше с развитием технологии GPGPU.

В этом году исполняется 10 лет с момента появления технологии Nvidia CUDA, которая предоставила удобный способ написания программ общего назначения для графических ускорителей. Большое количество вычислительных алгоритмов было переписано и тщательно оптимизировано для наилучшего использования возможностей GPU. При этом, несмотря на уже длительную историю развития, из-за ограничений по скорости обмена

*Работа поддержана грантом РФФИ № 14-50-00124.

данными большинство алгоритмов использует лишь малую часть теоретической производительности GPU даже после аккуратной настройки (см., например, [8]).

Среди программного обеспечения для вычислительного материаловедения, совместимого с технологией CUDA, можно выделить GROMACS [9], как, возможно, самый эффективный с вычислительной точки зрения инструмент молекулярного моделирования, и LAMMPS [10], как один из самых гибких пакетов для построения МД-моделей. В LAMMPS представлены несколько методов по переносу вычислений на GPU [11–14]. Хорошо оптимизированный GPU-алгоритм также имеется и в GROMACS [15].

Среди других способов увеличить производительность отдельных вычислительных узлов следует отметить использование GPU с технологией OpenCL, применение ускорителей Intel Xeon Phi, и разработку узкоспециализированных процессоров для решения задач МД, таких как MDGRAPE [16] или ANTON [3]. Сейчас, тем не менее, обычные GPU с поддержкой Nvidia CUDA являются наиболее экономичными решениями для выполнения высокопроизводительных МД-расчетов [17].

Современные MPP-системы могут объединять до 10^5 узлов для решения одной вычислительной задачи. Обмен между узлами на программном уровне чаще всего реализуется через MPI. Архитектура отдельных узлов может значительно отличаться и обычно выбирается, исходя из задач, на которые ориентирована MPP-система. Самым главным компонентом MPP-системы является интерконнект, свойства которого определяют масштабируемость любого MPI-алгоритма.

В данной работе рассматривается вычислительный кластер Десмос, состоящий из гибридных узлов (CPU+GPU), соединенных интерконнектом Ангара с тороидальной топологией. В статье представлена информация об интерконнекте, продемонстрирована итоговая производительность суперкомпьютера для МД расчетов моделей Леннард-Джонсовской жидкости, углеводорода и биологических систем в LAMMPS и GROMACS, расчетов электронной структуры полупроводников и воды в VASP и CP2K. Проведено сравнение с другими вычислительными комплексами, предназначенными для решения подобных задач.

2. Предшествующие работы

Тороидальные топологии интерконнекта зачастую обеспечивают лучшее параллельное масштабирование задачи по сравнению с топологиями типа толстое дерево (см., например, [31]). В 1990-е годы развитие MPP-систем достигло своего пика во время успеха систем Cray T3E, базировавшихся на трехмерной тороидальной топологии [18]. Cray T3E был первым суперкомпьютером, который достиг производительности 1 ТФлопс/сек для прикладной задачи. В июне 1998 Cray T3E занимал 4 место в списке Top500. Затем системы с уникальным (тороидальным) интерконнектом были вытеснены из лидеров Top500 системами, созданными из обычных комплектующих. В 2004, после нескольких лет доминирования подобных Beowulf-кластеров, выполненный по самостоятельному проекту тороидальный интерконнект появился в составе суперкомпьютера IBM BlueGene/L [19]. Последующие суперкомпьютеры от Cray и IBM также имели тороидальные топологии (за исключением последней серии XC от Cray). Суперкомпьютер K Computer от Fujitsu основан на собственном интерконнекте Tofu [20]. Суперкомпьютеры AURORA Booster и GREEN ICE Booster базируются на тороидальном интерконнекте EXTOLL [21].

Среди развивающихся вариантов российского суперкомпьютерного интерконнекта можно отметить сеть МВС-Экспресс, основанную на шине PCI-Express [22], прототипы на ПЛИС тороидальных интерконнектов СКИФ-Аврора [23] и Паутина [24]. И интерконнект Ангара, который к настоящему времени прошел путь от прототипа на ПЛИС [25, 26] до СБИС [27].

Среди тестов производительности суперкомпьютеров МД-тесты являются одними из самых популярных. К примеру, для довольно широко распространенной модели ApoA1 (белок в воде) имеются результаты для суперкомпьютеров Cray XK6 [28], IBM BlueGene/P

и BlueGene/Q systems [29], для K Computer [30] (во всех случаях вычисления были сделаны с помощью пакета NAMD). Суперкомпьютер AURORA Booster был протестирован с помощью пакета LAMMPS [21].

Также следует упомянуть примеры разработки альтернативных высокооптимизированных МД-алгоритмов для суперкомпьютеров. МД-алгоритм ls1 mardyn был создан для процессоров Intel семейства x86 с поддержкой векторизации и параллелизации с применением общей памяти. Это позволило выполнить расчет модели жидкости Леннард-Джонса с размером в несколько триллионов атомов на 146016 ядрах суперкомпьютера SuperMUC [2].

Считается, что тороидальная топология предпочтительна для сильного масштабирования многих алгоритмов. Тем не менее, точных данных для подтверждения этого утверждения довольно мало. Возможно, наиболее подробное на сегодняшний день сравнение было выполнено Фабиано Корсетти, который сравнил топологии тора и толстого дерева на примере пакета для расчета электронной структуры SIESTA на шести крупных суперкомпьютерах, принадлежащих к Tier-0 вычислительной сети PRACE [31]. Результаты Корсетти свидетельствуют, что вычислительные кластеры с тороидальной топологией демонстрируют лучшую масштабируемость для расчетов электронной структуры, чем кластеры на базе топологии толстого дерева. Сравнение тестовых данных для квантового пакета CP2k показывает схожую тенденцию [7].

3. Суперкомпьютер Десмос

Аппаратное обеспечение суперкомпьютера было выбрано с целью максимизировать (с учетом бюджетного лимита) число узлов и эффективность каждого узла для вычислительных МД-задач. Итоговая конфигурация каждого из 32 узлов показана в таблице 1.



Рис. 1. Суперкомпьютер Десмос.

Графические ускорители Nvidia GeForce в отличие от серверных аналогов Nvidia Tesla, не имеют функции коррекции ошибок памяти (ECC). Поэтому перед сборкой суперкомпьютера необходимо было удостовериться в том, что у использующихся ускорителей нет аппаратных ошибок памяти. Тестирование каждого GPU проведено с использованием MemtestG80 [32] при длительности теста более 4 часов для каждой платы. Все 32 ускорителя прошли данный тест.

Отдельным вопросом является охлаждение ускорителей. Мы используем ASUS GeForce

Таблица 1. Компоненты одного узла суперкомпьютера Десмос (без интерконнекта Ангара). В правой колонке указана цена оборудования в Москве в Октябре 2016.

Шасси	Supermicro SuperServer 1018GR-T	955 USD
CPU	Intel Xeon E5-1650v3 (6 cores, 3.5 GHz)	700 USD
GPU	Nvidia GeForce GTX 1070 (8 GB GDDR5)	550 USD
DRAM	DDR4 16 GB	140 USD
HDD	HDD 500 GB	44 USD
	Итого (без интерконнекта)	2389 USD

GTX 1070 8GB Turbo Edition. Каждый ускоритель был частично разобран перед установкой в стандартное 1U серверное шасси. Удаление пластикового корпуса и вентилятора на двойном шарикоподшипнике позволило обеспечить эффективное охлаждение платы графического ускорителя горизонтальным потоком холодного воздуха внутри шасси.

Узлы соединены с помощью Gigabit Ethernet и интерконнекта Ангара. Ввиду бюджетных ограничений не удалось использовать все возможные порты для построения полной топологии 4D-тора. Текущей топологией суперкомпьютера является 4D-тор $4 \times 2 \times 2 \times 2$ ($X \times Y \times Z \times K$), но каждый узел по осям Y, Z, K соединен с другим лишь одним линком.

Также в составе суперкомпьютера имеется идентичный по конфигурации головной узел, соединенный с остальными узлами с помощью Gigabit Ethernet. Суперкомпьютер работает под управлением ОС LES 11 SP4 с библиотекой Angara MPI, которая базируется на MPICH 3.0.4.

Энергопотребление суперкомпьютера без нагрузки составляет 6.5 кВт. При полной нагрузке суперкомпьютер потребляет 14.4 кВт.

Производительность кластера на тесте HPL составляет 9.12 ТФлопс/сек (85% от пиковой без учета GPU, размер матрицы 230000). Использование скромных возможностей ускорителей по вычислениям в двойной точности дает производительность кластера на тесте HPL-CUDA 12.75 ТФлопс/сек (77% от пиковой с учетом GPU, размер матрицы 220000).

4. Интерконнект Ангара

Интерконнект Ангара — это спроектированная в России коммуникационная сеть тороидальной топологии. СБИС адаптеров был разработан в АО “НИЦЭВТ” и произведен компанией TSMC по технологическому процессу 65 нм.

Архитектура Ангара использует принципы, аналогичные тороидальным интерконнектам IBM Blue Gene L/P и Cray Seastar2/Seastar2+. Тороидальный интерконнект, созданный EXTOLL также имеет общие черты с Ангарой [21]. Чип Ангара поддерживает deadlock-free адаптивную маршрутизацию [33], направленно-упорядоченную маршрутизацию [18,19], а также методы для повышения отказоустойчивости [18].

Каждый узел имеет выделенную область в памяти, доступную для удаленного доступа с других узлов (поддерживаются атомарные и read/write операции), что обеспечивает поддержку языков OpenSHMEM и PGAS. Для MPP-программной модели имеется поддержка MPI и OpenMP.

Сетевой адаптер представляет собой плату PCI Express, которая может быть использована для соединения с соседними узлами с возможностью подключения 6 кабелей (или 8 кабелей при использовании платы расширения). Поддерживаемыми топологиями в данном случае являются кольцо, двумерный, трехмерный и четырехмерный торы.

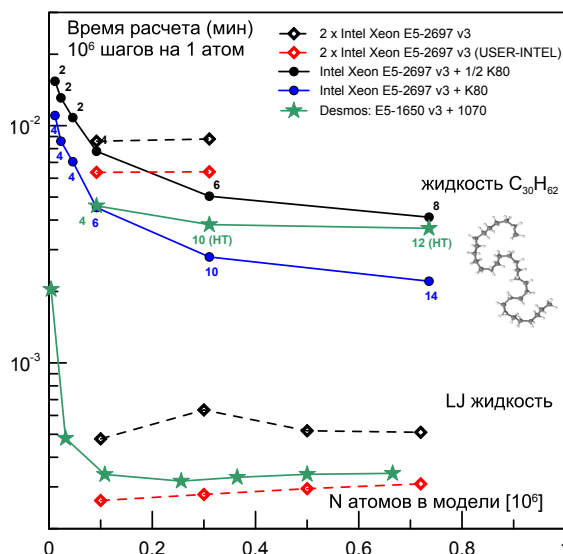


Рис. 2. Время расчета, требуемое на 10^6 шагов интегрирования на атом, для моделей LJ жидкости и масла $C_{30}H_{62}$ в пакете LAMMPS при разных размерах систем, различных методах параллелизации и комбинациях оборудования. Числа около символов для CPU+GPU данных показывают оптимальное число задействованных процессорных ядер равно числу независимых MPI-процессов.

5. Результаты классических МД тестов

На сегодняшний день использование одинарной точности в МД расчетах на GPU является стандартной практикой. Результаты таких проектов, как Folding@Home, подтверждают широкую применимость подобного подхода. Недавно разработанные оптимизированные МД-алгоритмы включают в себя проверку решения в одинарной точности на предмет ошибок (в т.ч. [34]).

Результаты измерения производительности для МД-кода LAMMPS в терминах времени расчета на 10^6 шагов интегрирования на атом и количества атомов в модели показаны на рис. 2. На рисунке продемонстрирована эффективность расчетов запущенных на одном вычислительном узле при разном количестве атомов МД-модели в LAMMPS: с использованием пакета GPU (смешанная точность) и без использования ускорителей. Пакет USER-INTEL в LAMMPS обеспечивает оптимизированные с учетом SIMD-операций векторизованные для процессоров Intel Xeon версии некоторых межатомных потенциалов. Как и ожидалось, пара CPU+GPU демонстрирует максимальную производительность для достаточно больших размеров системы. Результаты для узлов суперкомпьютера Десмос сравниваются на рис. 2 с результатами для двухпроцессорного узла с 14-ядерными процессорами Intel Xeon E5-2697v3 Haswell (суперкомпьютер MBC-10П МСЦ РАН).

Можно видеть, что простой тест леннард-джонсовской (LJ) жидкости на GPU-версии LAMMPS выполняется лишь немногим дольше, чем на оптимизированной версии LAMMPS на двухпроцессорном узле (28 ядер) без ускорителей, начиная с размера системы в 100 тысяч атомов на узел. Однако для более сложного теста $C_{30}H_{62}$ узлы суперкомпьютера Десмос работают быстрее. Для сравнения на рис. 2 представлены результаты такого же теста на ускорителе Nvidia Tesla K80.

Другой популярной биологической МД-моделью для тестирования производительности является апополипротеин A1 (ApoA1) в воде с размером системы в $\sim 10^5$ атомов. На рис. 3 показаны результаты слабой масштабируемости моделей Леннард-Джонсовской жидкости, $C_{30}H_{62}$ и ApoA1. Результаты показывают, что чем больше исходное число атомов в реплицируемой модели, тем более эффективна на суперкомпьютере Десмос масштабируемость в

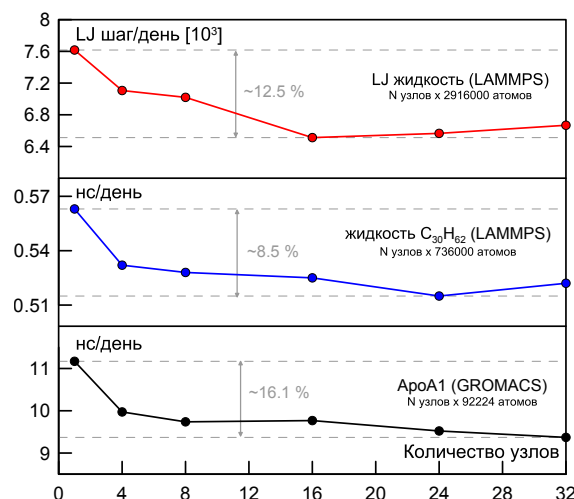


Рис. 3. Результаты тестов слабой масштабируемости трех разных МД моделей на суперкомпьютере Десмос.

слабом смысле.

Также следует отметить, что слабая масштабируемость, вероятно, может быть улучшена после имплементирования в Angara MPI учитывающих топологию сети декартовых MPI-коммуникаторов.

Критерий времени расчета приводит нас к очевидному выбору времени одного МД-шага интегрирования в качестве первого параметра метрики для анализа производительности. Второй параметр должен характеризовать оборудование. Обычно для этого используют число некоторых вычислительных элементов (например, число процессорных ядер). Хотя подобная метрика неплохо работает для анализа результатов тестов *сильной* и *слабой* масштабируемости, она не позволяет сравнивать между собой результаты решения одной и той же задачи на существенно разном вычислительном оборудовании. Для решения данной проблемы, предлагается использовать полную пиковую производительность R_{peak} в качестве второго параметра для метрики, что позволяет на равных рассматривать системы различной архитектуры [7].

Поскольку суперкомпьютер Десмос ориентирован на МД-расчеты в одинарной точности, необходимо принять некоторый обоснованный способ конвертировать значения Флопс/сек для одинарной точности в значения Флопс/сек для двойной точности. В данной работе мы полагаем для одного узла суперкомпьютера Десмос $R_{peakMD}^{node} = R_{peak}^{CPU-DP} + 0.5 * R_{peak}^{GPU-SP} = 336 + 2892 = 3228$ ГФлопс/сек.

Производительность разных суперкомпьютеров в тесте АроА1 показана на рис. 4 в метрике числа секунд, нормированных на число атомов и МД-шагов, и пиковой производительности. Пунктирной линией показана идеальная масштабируемость, соответствующая 0.1 МФлопс/сек/атом/шаг.

Таким образом, несмотря на различие архитектур и МД-алгоритмов, мы можем сравнить эффективности применения суперкомпьютеров для биомолекулярных МД расчетов. Новый суперкомпьютер Десмос (пурпурные звездочки) для пакета GROMACS демонстрирует меньшие времена расчета и лучшую масштабируемость, чем суперкомпьютеры Cray XK6 [28] и K Computer для пакета NAMD. Десмос оказался эффективней, чем системы BlueGene [29].

Также на графике показаны времена расчета для увеличенной в 27 раз модели АроА1, полученные на суперкомпьютере Десмос. Результаты демонстрируют практически идеальное масштабирование с увеличением вычислительного поля суперкомпьютера, что указы-

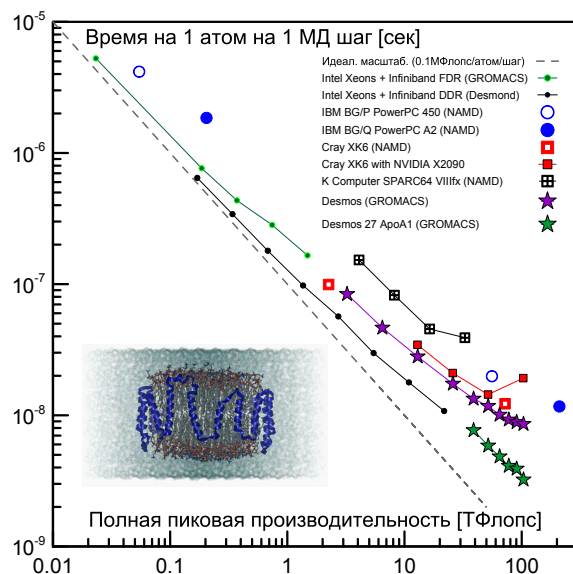


Рис. 4. Результаты теста ApoA1 с использованием GROMACS для суперкомпьютера Десмос (включая увеличенную в $3 \times 3 \times 3 = 27$ раз модель; отмечена зелеными звездочками). Для сравнения приведены данные такого же теста с использованием NAMD для суперкомпьютеров Cray XK6 [28], IBM BlueGene/P и BlueGene/Q [29], K Computer [30]. Черная и зеленая линии с точками показывают производительность суперкомпьютеров с узлами на Intel Xeon [6].

вадет на возможность добавления новых узлов для дальнейшего ускорения счета без существенных потерь в эффективности параллелизации.

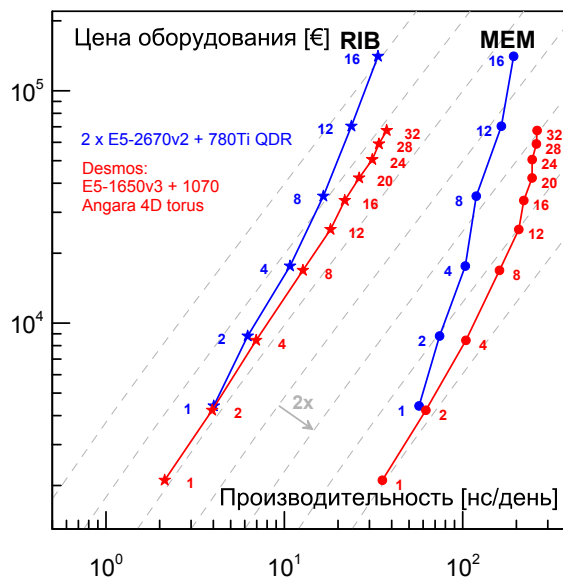


Рис. 5. Стоимость аппаратного обеспечения и достигнутая производительность для тестов MEM (отмечены точками) и RIB (отмечены звездочками). Результаты, полученные на суперкомпьютере Десмос (красный цвет), сравниваются с опубликованными в [17] данными (синий цвет). Серые пунктирные линии соответствуют идеальному масштабированию.

В работе Кутцнера [17] приводятся рекомендации по достижению наилучшего соотношения цена-производительность (по состоянию на 2015 год). Авторы сравнили различные конфигурации гибридных систем, используя два биологических МД-тесте: белок мембран-

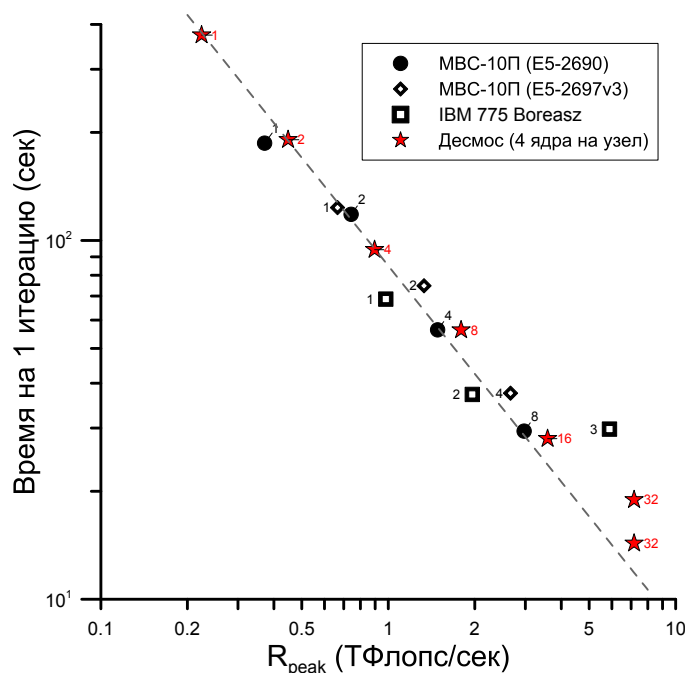


Рис. 6. Масштабируемость расчета модели кристалла GaAs в VASP. На рисунке показано время на 1 итерацию алгоритма поиска самосогласованной электронной плотности в зависимости от полной пиковой мощности вычислительного поля. Приведены данные для различных систем: двух разделов суперкомпьютера MBC-10П в МСЦ РАН и суперкомпьютера Boreasz IBM 775 Варшавского университета. Числа рядом с символами показывают соответствующее каждому случаю число узлов кластера. Нижняя из двух точек, показывающих время счета на 32 узлах суперкомпьютера Десмос, соответствует дополнительному распараллеливанию по k-точкам.

ного канала, помещенный в липидный бислой, окруженный водой (MEM, $\sim 10^5$ атомов) и бактериальная рибосома в воде с ионами (RIB, $\sim 2 \cdot 10^6$ атомов). Тесты проводились с использованием пакета GROMACS. В тестах суперкомпьютера Десмос мы использовали идентичные модели и протоколы расчета, но в целях получения лучшей точности мы использовали более длительные МД-расчеты, чем в работе [17].

Проведено сравнение результатов, полученных на суперкомпьютере Десмос с лучшим вариантом, выбранным в работе [17] (система с двумя процессорами Intel Xeon E5-2670v2 и двумя ускорителями Nvidia GPU 780Ti на узел, с интерконнектом Infiniband QDR). На рис. 5 по оси ординат откладывается стоимость оборудования в Евро (для пересчета стоимости суперкомпьютера Десмос принято соотношение Евро/USD=1.1, соответствующее курсу Октября 2016 года), а по оси абсцисс — производительность в нс/день. Числами у символов на рис. 5 обозначено количество узлов. Идеальное масштабирование показано серыми пунктирными линиями.

Суперкомпьютер Десмос показывает лучшую сильную масштабируемость на тесте MEM, чем наилучший вариант, выбранный авторами работы [17]. Основной причиной, очевидно, является новая архитектура GPU Pascal. Насыщение достигается при числе узлов выше 20, что соответствует небольшому числу атомов на узел, что ниже порога эффективности для GPU. Наклон зависимостей, полученных на суперкомпьютере Десмос, наглядно демонстрирует существенно более высокие скоростные характеристики сети Ангара по сравнению с QDR Infiniband. В случае с тестом RIB Десмос демонстрирует близкое к идеальному масштабирование при числе узлов выше 16, что указывает на возможность увеличения скорости соответствующих расчетов при дальнейшем увеличении числа узлов в системе.

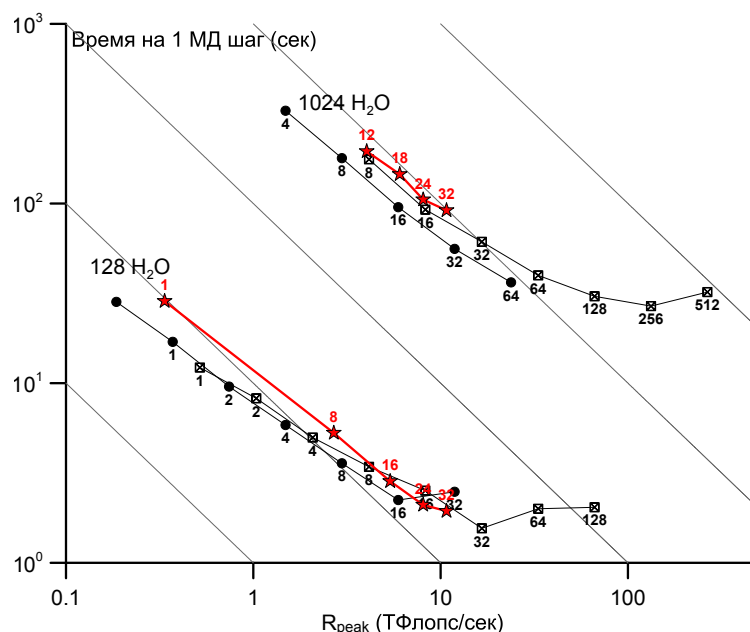


Рис. 7. Масштабируемость расчета модели воды (128 и 1024 молекул H_2O) в CP2K. Круги соответствуют результатам, полученным на MBC-10П, а квадраты — суперкомпьютеру Cray XC30. Числа рядом с символами показывают соответствующее каждому случаю число узлов кластера.

6. Результаты тестовых расчетов электронной структуры

Основная вычислительная мощность суперкомпьютера Десмос сосредоточена в ускорителях и ориентирована на классические МД расчеты. На каждом узле имеется только один 6-ти ядерный процессор Intel Xeon E5-1650v3, работающий на тактовой частоте 3.5 ГГц. Однако в методических целях представляет интерес провести тесты масштабируемости расчетов электронной структуры. Электронно-структурные расчеты в рамках теории функционала электронной плотности предъявляют более высокие требования к скорости коллективных операций типа all-to-all, чем модели классической МД. Для тестов мы используем два очень популярных пакета VASP и CP2K. Согласно существующим оценкам [35, 36] расчеты с использованием VASP потребляют до 20 процентов всего вычислительного времени суперкомпьютеров в мире.

Для теста в VASP была использована модель кристалла GaAs, состоящая из 80 атомов [37]. Для теста CP2K была использована стандартная модель воды (при варьировании числа молекул от 32 до 1024 — больший размер системы невозможно использовать на суперкомпьютере Десмос из-за ограничения по оперативной памяти).

Результаты тестов показаны на рис. 6 и рис. 7. Полученные данные свидетельствуют, что сеть Ангара очень эффективно объединяет отдельные узлы суперкомпьютера в единое вычислительное поле. Суперкомпьютеры, данные по которым используются для сравнения, имеют двухпроцессорные узлы (IBM 775 даже четырехпроцессорные). Тем не менее, MPI-обмены по сети Ангара — в смысле итоговой производительности — обеспечивают аналогичный результат, как и MPI-обмены в общей памяти.

7. Выводы

В данной статье приведено описание суперкомпьютера Десмос, ориентированного на расчеты методами классической молекулярной динамики. Суперкомпьютер Десмос является первым практическим применением отечественного интерконнекта Ангара для гибри-

ной массово-параллельной системы. Результаты тестов и их сравнение с литературными данными по производительности суперкомпьютеров экстра-класса показывают высокую эффективность данной системы для классических МД расчетов. Проверка масштабируемости расчетов электронной структуры также показала высокую эффективность параллельного счета. Рассмотренный пример суперкомпьютера Десмос свидетельствует о том, что сеть Ангара обеспечивают высокий уровень эффективности МРР-систем, и создание на ее основе суперкомпьютеров на 1-2 порядка большего размера даст эффективный вычислительный инструмент для решения задач вычислительного материаловедения.

Авторы статьи рады выразить признательность коллегам из АО “НИЦЭВТ” за плодотворное сотрудничество при создании суперкомпьютера Десмос. Авторы благодарят МСЦ РАН (<http://www.jssc.ru>) за доступ к суперкомпьютеру МВС-10П и выражают благодарность компании “Форсайт” за предоставление доступа к тестовому серверу с Nvidia Tesla K80. Отдельную благодарность хотелось бы выразить Dr. Maciej Cytowski и Dr. Jacek Reichota (Варшавский университет) за предоставление данных по тесту GaAs [37].

Литература

1. Heinecke A., Eckhardt W., Horsch M., Bungartz H.-J. Supercomputing for Molecular Dynamics Simulations. Springer, 2015. 76 p.
2. Eckhardt W., Heinecke A., Bader R., Brehm M., Hammer N., Huber H., Kleinhenz H.-G., Vrabec J., Hasse H., Horsch M., Bernreuther M., Glass C. W., Niethammer C., Bode A., Bungartz H.-J. 591 TFLOPS multi-trillion particles simulation on SuperMUC // Lecture Notes in Computer Science. 2013. V. 7905. P. 1–12.
3. Piana S., Klepeis J. L., Shaw D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations // Current Opinion in Structural Biology. 2014. V. 24. P. 98–105.
4. Sutmann G. Classical molecular dynamics // Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms (NIC Series): John von Neumann Institute for Computing, Julich. 2002. V. 10. P. 211–254.
5. Begau Ch., Sutmann G. Adaptive dynamic load-balancing with irregular domain decomposition for particle simulations // Computer Physics Communications. 2015. V. 190. P. 51–61.
6. Смирнов Г. С., Стегайлов В. В. Эффективность алгоритмов классической молекулярной динамики на суперкомпьютерном аппаратном обеспечении // Матем. моделирование. 2016. Т. 28. С. 95–108.
7. Stegailov V. V., Orekhov N. D., Smirnov G. S. HPC Hardware Efficiency for Quantum and Classical Molecular Dynamics // Parallel Computing Technologies: 13th International Conference, PaCT 2015 (Petrozavodsk, Russia, August 31-September 4, 2015): Springer International Publishing, 2015. P. 469–473.
8. Rojek K., Wyrzykowski R., Kuczynski L. Systematic adaptation of stencil-based 3D MPDATA to GPU architectures // Concurrency and Computation: Practice and Experience. 2017. V. 29. DOI: <http://dx.doi.org/10.1002/cpe.3970>.
9. Berendsen H.J.C., van der Spoel D., van Drunen R. Gromacs: A message-passing parallel molecular dynamics implementation // Computer Physics Communications. 1995. V. 91. P. 43–56.

10. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics // J. Comput. Phys. 1995. V. 117. P. 1–19.
11. Trott C. R., Winterfeld L., Crozier P. S. General-purpose molecular dynamics simulations on GPU-based clusters // ArXiv e-prints. 2010.
12. Brown W. M., Wang P., Plimpton S. J., Tharrington A. N. Implementing molecular dynamics on hybrid high performance computers — short range forces // Computer Physics Communications. 2011. V. 182. P. 898–911.
13. Brown W. M., Kohlmeyer A., Plimpton S. J., Tharrington A. N. Implementing molecular dynamics on hybrid high performance computers — Particle-particle particle-mesh // Computer Physics Communications. 2012. V. 183. P. 449–459.
14. Edwards H. C., Trott C. R., Sunderland D. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns // Journal of Parallel and Distributed Computing. 2014. V. 74. P. 3202–3216.
15. Abraham M. J., Murtola T., Schulz R., Páll S., Smith J. C., Hess B., Lindahl E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers // SoftwareX. 2015. V. 1–2. P. 19–25.
16. Ohmura I., Morimoto G., Ohno Y., Hasegawa A., Taiji M. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations // Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 2014. V. 372. P. 2021.
17. Kutzner C., Páll S., Fechner M., Esztermann A., de Groot B. L., Grubmüller H. Best bang for your buck: Gpu nodes for gromacs biomolecular simulations // Journal of Computational Chemistry. 2015. V. 36. P. 1990–2008.
18. Scott S. L., Thorson G. M. The Cray T3E network: Adaptive routing in a high performance 3D torus // HOT Interconnects IV, Stanford University, August 15-16, 1996.
19. Adiga N. R., Blumrich M. A., Chen D., Coteus P., Gara A., Giampapa M. E., Heidelberger P., Singh S., Steinmacher-Burow B. D., Takken T., Tsao M., Vranas P. Blue Gene/L torus interconnection network // IBM J. Res. Dev. 2005. V. 49. P. 265–276.
20. Ajima Y., Inoue T., Hiramoto S., Takagi Y., Shimizu T. The Tofu interconnect // IEEE Micro. 2012. V. 32. P. 21–31.
21. Neuwirth S., Frey D., Nuessle M., Bruening U. Scalable communication architecture for network-attached accelerators // 2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA). 2015. P. 627–638.
22. Елизаров Г. С., Горбунов В. С., Левин В. К., Лацис А. О., Корнеев В. В., Соколов А. А., Андрюшин Д. В., Климов Ю. А. Коммуникационная сеть МВС-Экспресс // Выч. мет. программирование. 2012. Т. 13. С. 103–109.
23. Адамович И. А., Климов А. В., Климов Ю. А., Орлов А. Ю., Шворин А. Б. Опыт разработки коммуникационной сети суперкомпьютера «СКИФ-Аврора» // Программные системы: теория и приложения. 2010. Т. 1. С. 107–123.
24. Климов Ю. А., Шворин А. Б., Хренов А. Ю., Адамович И. А., Орлов А. Ю., Абрамов С. М., Шевчук Ю. В., Пономарев А. Ю. Паутина: высокоскоростная коммуникационная сеть // Программные системы: теория и приложения. 2015. Т. 6. С. 109–120.

25. Симонов А. С., Макагон Д. В., Жабин И. А., Щербак А. Н., Сыромятников Е. Л., Поляков Д. А. Первое поколение высокоскоростной коммуникационной сети «Ангара» // Научные технологии. 2014. Т. 15. С. 21–28.
26. Мукосей А. В., Семенов А. С., Симонов А. С. Имитационное моделирование подсети коллективных операций сети «Ангара» // Вестник Южно-Уральского государственного университета. Серия: Вычислительная математика и информатика. 2015. Т. 4. С. 40–55.
27. Агарков А. А., Исмаилов Т. Ф., Макагон Д. В., Семенов А. С., Симонов А. С. Результаты оценочного тестирования отечественной высокоскоростной коммуникационной сети Ангара // Сборник трудов международной конференции «Суперкомпьютерные дни в России — 2016». 2016. С. 626–639.
28. Sun Y., Zheng G., Mei C., Bohm E. J., Phillips J. C., Kalé L. V., Jones Th. R. Optimizing fine-grained communication in a biomolecular simulation application on Cray XK6 // Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, SC '12, Los Alamitos, CA, USA: IEEE Computer Society Press, 2012. P. 55:1–55:11.
29. Kumar S., Sun Y., Kalé L. V. Acceleration of an asynchronous message driven programming paradigm on IBM Blue Gene/Q // 2013 IEEE 27th International Symposium on Parallel and Distributed Processing. 2013. P. 689–699.
30. Yonezawa A. Concurrent objects, agents and HPC: AICS RIKEN, 2012. URL: <http://www.slideshare.net/AkinoriYonezawa/ageretalk>.
31. Corsetti F. Performance analysis of electronic structure codes on HPC systems: A case study of SIESTA // PLOS ONE. 2014. V. 9. P. 1–8.
32. Haque I. S., Pande V. S. Hard data on soft errors: A large-scale assessment of real-world error rates in GPGPU // Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGRID '10, Washington, DC, USA: IEEE Computer Society, 2010. P. 691–696.
33. Puente V., Beivide R., Gregorio J. A., Prellezo J. M., Duato J., Izu C. Adaptive bubble router: a design to improve performance in torus networks // Proceedings of the 1999 International Conference on Parallel Processing. 1999. P. 58–67.
34. Höhnerbach M., Ismail A. E., Bientinesi P. The vectorization of the Tersoff multi-body potential: An exercise in performance portability // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '16: IEEE Press, Piscataway, NJ, USA, 2016. P. 7:1–7:13.
35. Bethune I. “Ab initio molecular dynamics,” Introduction to Molecular Dynamics on ARCHER, 2015. URL: https://www.archer.ac.uk/training/course-material/2015/06/MolDy_Strath/AbInitioMD.pdf.
36. Hutchinson M. “VASP on GPUs. When and how,” GPU technology theater, SC15, 2015. URL: <http://images.nvidia.com/events/sc15/pdfs/SC5107-vasp-gpus.pdf>.
37. Cytowski M. “Best Practice Guide — IBM Power 775,” PRACE, November 2013. URL: <http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-IBM-Power-775.pdf>.

Angara-based hybrid supercomputer for efficient acceleration of computational materials science studies

V.S. Vecher^{1,2}, N.D. Kondratyuk^{1,2}, G.S. Smirnov¹, V.V. Stegailov¹

Joint Institute for High Temperatures of Russian Academy of Sciences¹,
Moscow Institute of Physics and Technology²

In this paper, we describe the supercomputer Desmos consisting of 32 hybrid nodes connected by a low-latency high-bandwidth torus interconnect. This supercomputer is aimed at classical molecular dynamics calculations and is able to speed-up efficiently electronic structure calculations as well. In this report, we present scalability benchmarks for a representative set of models in LAMMPS, GROMACS, VASP and CP2K. We compare the benchmark results with other HPC systems. This supercomputer is based on the Angara interconnect developed by JSC NICEVT that supports 3D and 4D torus topologies. The observed parallel scaling of the benchmarks on the supercomputer Desmos illustrates the capability of the Angara interconnect to speed-up efficiently MPI-based applications on MPP systems.

Keywords: supercomputer, Angara interconnect, performance benchmarks, computational materials science

References

1. Heinecke A., Eckhardt W., Horsch M., Bungartz H.-J. Supercomputing for Molecular Dynamics Simulations. Springer, 2015. 76 p.
2. Eckhardt W., Heinecke A., Bader R., Brehm M., Hammer N., Huber H., Kleinhenz H.-G., Vrabec J., Hasse H., Horsch M., Bernreuther M., Glass C. W., Niethammer C., Bode A., Bungartz H.-J. 591 TFLOPS multi-trillion particles simulation on SuperMUC // Lecture Notes in Computer Science. 2013. V. 7905. P. 1–12.
3. Piana S., Klepeis J. L., Shaw D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations // Current Opinion in Structural Biology. 2014. V. 24. P. 98–105.
4. Sutmann G. Classical molecular dynamics // Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms (NIC Series): John von Neumann Institute for Computing, Julich. 2002. V. 10. P. 211–254.
5. Begau Ch., Sutmann G. Adaptive dynamic load-balancing with irregular domain decomposition for particle simulations // Computer Physics Communications. 2015. V. 190. P. 51–61.
6. Smirnov G. S., Stegailov V. V. Efficiency of classical molecular dynamics algorithms on supercomputers // Mathematical Models and Computer Simulations. 2016. V. 8. P. 734–743.
7. Stegailov V. V., Orekhov N. D., Smirnov G. S. HPC Hardware Efficiency for Quantum and Classical Molecular Dynamics // Parallel Computing Technologies: 13th International Conference, PaCT 2015, Petrozavodsk, Russia, August 31-September 4: Springer International Publishing, 2015. P. 469–473.
8. Rojek K., Wyrzykowski R., Kuczynski L. Systematic adaptation of stencil-based 3D MPDATA to GPU architectures // Concurrency and Computation: Practice and Experience. 2017. V. 29. DOI: <http://dx.doi.org/10.1002/cpe.3970>.

9. Berendsen H.J.C., van der Spoel D., van Drunen R. Gromacs: A message-passing parallel molecular dynamics implementation // *Computer Physics Communications*. 1995. V. 91. P. 43–56.
10. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics // *J. Comput. Phys.* 1995. V. 117. P. 1–19.
11. Trott C. R., Winterfeld L., Crozier P. S. General-purpose molecular dynamics simulations on GPU-based clusters // *ArXiv e-prints*. 2010.
12. Brown W. M., Wang P., Plimpton S. J., Tharrington A. N. Implementing molecular dynamics on hybrid high performance computers — short range forces // *Computer Physics Communications*. 2011. V. 182. P. 898–911.
13. Brown W. M., Kohlmeyer A., Plimpton S. J., Tharrington A. N. Implementing molecular dynamics on hybrid high performance computers — Particle-particle particle-mesh // *Computer Physics Communications*. 2012. V. 183. P. 449–459.
14. Edwards H. C., Trott C. R., Sunderland D. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns // *Journal of Parallel and Distributed Computing*. 2014. V. 74. P. 3202–3216.
15. Abraham M. J., Murtola T., Schulz R., Páll S., Smith J. C., Hess B., Lindahl E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers // *SoftwareX*. 2015. V. 1–2. P. 19–25.
16. Ohmura I., Morimoto G., Ohno Y., Hasegawa A., Taiji M. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations // *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. 2014. V. 372. P. 2021.
17. Kutzner C., Páll S., Fechner M., Esztermann A., de Groot B. L., Grubmüller H. Best bang for your buck: Gpu nodes for gromacs biomolecular simulations // *Journal of Computational Chemistry*. 2015. V. 36. P. 1990–2008.
18. Scott S. L., Thorson G. M. The Cray T3E network: Adaptive routing in a high performance 3D torus // *HOT Interconnects IV*, Stanford University, August 15–16, 1996.
19. Adiga N. R., Blumrich M. A., Chen D., Coteus P., Gara A., Giampapa M. E., Heidelberger P., Singh S., Steinmacher-Burow B. D., Takken T., Tsao M., Vranas P. Blue Gene/L torus interconnection network // *IBM J. Res. Dev.* 2005. V. 49. P. 265–276.
20. Ajima Y., Inoue T., Hiramoto S., Takagi Y., Shimizu T. The Tofu interconnect // *IEEE Micro*. 2012. V. 32. P. 21–31.
21. Neuwirth S., Frey D., Nuessle M., Bruening U. Scalable communication architecture for network-attached accelerators // *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*. 2015. P. 627–638.
22. Elizarov G. S., Gorbunov V. S., Levin V. K., Latsis A. O., Korneev V. V., Sokolov A. A., Andryushin D. V., Klimov Yu. A. Communication fabric MVS-Express // *Vychisl. Metody Programm.* 2012. V. 13. P. 103–109.
23. Adamovich I. A., Klimov A. V., Klimov Yu. A., Orlov A. Yu., Shvorin A. B. Thoughts on the development of SKIF-Aurora supercomputer interconnect // *Programmnye Sistemy: Teoriya i Prilozheniya*. 2010. V. 1. P. 107–123.

24. Klimov Yu. A., Shvorin A. B., Khrenov A. Yu., Adamovich I. A., Orlov A. Yu., Abramov S. M., Shevchuk Yu. V., Ponomarev A. Yu. Pautina: the high performance interconnect // *Programmnye Sistemy: Teoriya i Prilozheniya*. 2015. V. 6. P. 109–120.
25. Simonov A.S., Makagon D. V., Zhabin I. A., Scherbak A. N., Syromyatnikov E. L., Polyakov D. A. The first generation of the Angara interconnect // *Naukoemkieologii*. 2014. V. 15. P. 21–28.
26. Mukosey A. V., Semenov A. S., Simonov A. S. Simulation of collective operations hardware support for Angara interconnect // *Vestn. YuUrGU. Ser. Vych. Matem. Inform.* 2015. V. 4. P. 40–55.
27. Agarkov A. A., Ismagilov T. F., Makagon D. V., Semenov A. S., Simonov A. S. Performance evaluation of the Angara interconnect. // *Proceedings of the International Conference “Russian Supercomputing Days” — 2016*. 2016. P. 626–639.
28. Sun Y., Zheng G., Mei C., Bohm E. J., Phillips J. C., Kalé L. V., Jones Th. R. Optimizing fine-grained communication in a biomolecular simulation application on Cray XK6 // *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, SC '12, Los Alamitos, CA, USA: IEEE Computer Society Press*, 2012. P. 55:1–55:11.
29. Kumar S., Sun Y., Kalé L. V. Acceleration of an asynchronous message driven programming paradigm on IBM Blue Gene/Q // *2013 IEEE 27th International Symposium on Parallel and Distributed Processing*. 2013. P. 689–699.
30. Yonezawa A. Concurrent objects, agents and HPC: AICS RIKEN, 2012. URL: <http://www.slideshare.net/AkinoriYonezawa/ageretalk>.
31. Corsetti F. Performance analysis of electronic structure codes on HPC systems: A case study of SIESTA // *PLOS ONE*. 2014. V. 9. P. 1–8.
32. Haque I. S., Pande V. S. Hard data on soft errors: A large-scale assessment of real-world error rates in GPGPU // *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, CCGRID '10, Washington, DC, USA: IEEE Computer Society*, 2010. P. 691–696.
33. Puente V., Beivide R., Gregorio J. A., Prellezo J. M., Duato J., Izu C. Adaptive bubble router: a design to improve performance in torus networks // *Proceedings of the 1999 International Conference on Parallel Processing*. 1999. P. 58–67.
34. Höhnerbach M., Ismail A. E., Bientinesi P. The vectorization of the Tersoff multi-body potential: An exercise in performance portability // *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '16: IEEE Press, Piscataway, NJ, USA, 2016*. P. 7:1–7:13.
35. Bethune I. “Ab initio molecular dynamics,” *Introduction to Molecular Dynamics on ARCHER*, 2015. URL: https://www.archer.ac.uk/training/course-material/2015/06/MolDy_Strath/AbInitioMD.pdf.
36. Hutchinson M. “VASP on GPUs. When and how,” *GPU technology theater, SC15*, 2015. URL: <http://images.nvidia.com/events/sc15/pdfs/SC5107-vasp-gpus.pdf>.
37. Cytowski M. “Best Practice Guide — IBM Power 775,” *PRACE*, November 2013. URL: <http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-IBM-Power-775.pdf>.