Соотношение между энергопотреблением и производительностью для GPU-алгоритмов молекулярной динамики

В.С. Вечер[1,2],В.П. Никольский[3,2],В.В. Стегайлов[2]
МФТИ ГУ[1], ОИВТ РАН[2], НИУ ВШЭ[3]

Энергопотребление гибридных вычислительных систем — актуальная проблема современных высокопроизводительных вычислений. Вопрос выбора между производительностью и энергопотреблением — весьма непростой. В рамках статьи рассмотрены производительность и энергопотребление двух современных гибридных миникомпьютеров Nvidia Jetson TK1 и TX1 на двух типах приложений — тесте Empirical Roofline Tool, разработанном для пиковой вычислительной загрузки, и на молекулярно-динамическом коде LAMMPS. С использованием точного ваттметра проведены измерения потребления миникомпьютеров в разных частотных режимах памяти и графического ускорителя, рассмотрен режим динамического изменения частот. Представлены результаты измерений для достижения оптимального соотношения частоты GPU и частоты DRAM с точки зрения минимума энергопотребления.

Ключевые слова: Tegra, Kepler, Maxwell, ARM, энергоэффективность, молекулярная динамика

1. Введение

Метод молекулярной динамики — современный и мощный метод компьютерного моделирования, позволяющий рассчитывать поведение систем, состоящих из миллионов атомов. Предсказание поведения систем на молекулярном уровне, а значит — и их свойств, является одним из главных достоинств данного метода. Ввиду своей эффективности, метод часто является востребованным в исследовательских работах по физике, химии и биологии.

В основе метода молекулярной динамики лежит идея представления системы в виде частиц, подчиняющихся законам классической механики. Поэтому, эволюция системы как зависимость скоростей, координат и ускорений частиц от времени легко находится интегрированием уравнений движения Ньютона.

По мере увеличения вычислительных мощностей, доступных исследователям, увеличивались размеры и сложность исследуемых систем — а значит, и повышалась точность получаемых результатов. Однако рост вычислительной нагрузки, связанный с увеличением размера моделируемых систем с целью добиться еще более высокой точности, привел к ситуации, когда единственным выходом стало применение параллельных вычислений.

Кроме того, повсеместное удешевление графических акселераторов и повышение их мощности привело к тому,что для ряда задач по соотношению “цена-производительность” они стали превосходить обычные процессоры. Естественно, данный факт не остался незамеченным, и сейчас для задач молекулярной динамики применение графических ускорителей является повсеместной практикой.

Однако рост энергопотребления и тепловыделения вычислительных платформ также является весьма значимой проблемой, особенно в связи с перспективой экзафлопсных систем. Рост вычислительных возможностей CPU архитектуры ARM, а также их достаточно низкие показатели тепловыделения и энергопотребления открывают один из перспективных путей развития высокопроизводительных вычислительных систем.

Таким образом, идея оценки энергоэффективности гетерогенных вычислительных систем возникает вполне закономерно. На рынке уже имеется ряд энергоэффективных ус-

тройств, реализующих концепцию системы на чипе (SoC), т.е. имеющих интегрированный графический акселератор на том же кристалле, что и CPU. Примерами таких систем может, например, являться семейство кристаллов ARM Mali или Nvidia Tegra, причем последние версии — Tegra K1 и Tegra X1 — обладают встроенным графическим ядром, поддерживающим технологию CUDA, а значит, пригодным для вычислений общего назначения.

Таким образом, цель данной работы — оценить реальную эффективность работы молекулярно-динамических алгоритмов на семействе чипов Tegra с точки зрения энергопотребления.

2. Работы по анализу энергопотребления

Касаясь вопроса энергопотребления, можно отметить работу [1], в которой предлагалось снижать потребление за счет понижения числа переключений в электрических цепях процессорных СБИС. Вопрос расчета энергопотребления и его взаимосвязи с оптимизацией кода для 32-битных встраиваемых RISC-процессоров достаточно исследован в работе Joseph и Martonosi [2]. Более сложная модель оценки потребления в реальном времени обсуждалась в работе Russel и Jacome [3]. Оценка энергопотребления на уровне ОС освещена в работе Li и John [4].

Энергопотребление ARM-устройств было рассмотрено в работе Zhang et al. [5]. Подробно производительность ARM-устройств на примере алгоритмов молекулярной динамики рассматривается в работе Никольского и Стегайлова [6]. В работе Calore et al. [7] раскрываются некоторые аспекты соотношения энергопотребления и производительности для устройства Tegra K1.

В недавнем обзоре [8] был дан обзор ключевых аспектов моделирования производительности систем, основанных на акселераторах. В качестве краеугольного камня в этой области рассматривается модель McPAT (Multicore Power, Area and Timing) [9]. Другой подход GPUWattch [10] направлен на предсказание энергопотребления и его оптимизацию путем тщательной настройки на основе проведения микротестов. Эти подходы дают возможность достаточно точно предсказывать энергопотребление CPU и/или GPU (с точность порядка 5-10 %). Однако применение конкретной модели энергопотребления типа McPAT или GPUWattch к новому типу аппаратного обеспечения и новому типу программно-алгоритмической нагрузки представляет собой очень значительную по объему работу. Поэтому прямые экспериментальные изменения потребляемой мощности и энергии представляют значительный интерес. Кроме того, в настоящее время активно совершенствуется методика определения энергопотребления больших вычислительных систем [11]. Исследования на меньших масштабах, подобные данной работе, призваны способствовать решению задачи предсказания и оптимизации энергопотребления на больших масштабах.

3. Используемое программное обеспечение

3.1. Определение пиковых характеристик: Empirical Roofline Tool

Для оценки производительности гетерогенных систем могут быть применены разные подходы. Следует отметить, что оценка производительности системы, исходя лишь из вычислительной мощности устройства не может считаться полностью верной. Такая оценка справедлива лишь для compute-bound алгоритмов. Для memory-bound алгоритмов существенную роль играет пропускная способность памяти — потому данный аспект должен учитываться при составлении оценки производительности.

Такая мысль привела к появлению модели Roofline и соответствующего программного пакета для тестирования Empirical Roofline Tool (ERT), разработанного в лаборатории университета Беркли. Цель теста — позволить пользователю оценить максимальную производительность различных алгоритмов на доступном оборудовании.

Для этого была введена характеристика арифметической интенсивности, показывающей отношение числа арифметических операций над данными к числу передаваемых данных. Очевидно, что для алгоритмов с большой арифметической интенсивностью ограничителем будет служить максимальная производительность процессора, в то время, как алгоритмы с большой степенью передачи данных ограничиваются пропускной способностью памяти.

Результат применения модели приводит к построению графика ограничений по производительности для алгоритмов с разными арифметическими интенсивностями. Таким образом, зная арифметическую интенсивность алгоритма, можно достаточно точно оценить производительность системы при выполнении искомого алгоритма и тип ограничения. Например, для методов решеточных уравнений Больцмана арифметическая интенсивность составляет менее единицы ФЛОПс/байт, в то время как для методов частиц — порядка десяти ФЛОПс/байт.

Для практической оценки пропускной способности памяти, пиковой вычислительной мощности с учетом особенностей современных сложных гетерогенных распределенных вычислительных систем и используется пакет ERT.

В основе алгоритма ERT лежит идея выполнения в циклах простейших арифметических операций над элементами массива определенной длины. На разных запусках во вложенных циклах варьируется число операций над одним и тем же элементом массива (ERT_FLOPS) и размер массива. Изменение размера массива данных позволяет обнаружить наличие эффекта кэширования. Изменение числа операций над одним элементом массива позволяет выявить эффекты автоматической векторизации.

```
for(int i=0; i<n; ++i){
        if (ERT_FLOPS==1){
                b = a[i] + alpha;
        }
        if (ERT_FLOPS==2){
                b = a[i]*b + alpha;
        }
        if (ERT_FLOPS==4){
                b = a[i]*b + alpha;
                b = a[i]*b + alpha;
        }
...
};
```

Рис. 1: Иллюстрация зависимости тела цикла от параметра ERT_FLOPS

3.2. Классическая молекулярная динамика: LAMMPS

Кроме сугубо тестового инструмента ERT, в рамках данной работы использовался и применяемый на практике молекулярно-динамический код. Выбор пал на МД-пакет с открытым исходным кодом — LAMMPS, разработанный специалистами Сандийской Национальной Лаборатории (США), который применяется для крупных расчетов на атомном и мезоатомном масштабе.

С точки зрения применения графических ускорителей LAMMPS может использовать несколько вариантов гибридных алгоритмов (в том числе, пакеты GPU и USER-CUDA).

Первый из них (GPU) — поддерживает большее количество конфигураций оборудования и способен запускаться не только на платформах, поддерживающих технологию CUDA, но и на вычислительных средствах с поддержкой OpenCL. Однако из-за особенности реализации алгоритма — а именно, перемещения данных с GPU на хост-машину после каждого шага по времени, данная реализация часто проигрывает USER-CUDA по скорости вычисления. Алгоритм USER-CUDA лишен подобного недостатка, хотя и поддерживает только

устройства с технологией CUDA.

Для анализа производительности была выбрана ставшая уже классической для молекулярной динамики модель Леннард-Джонсоновской жидкости (108000 атомов при плотности $0.8442\sigma^{-3}$ и радиусе обрезания потенциала в 2.5σ, NVE-ансамбль, 250 шагов по времени).

4. Используемое аппаратное обеспечение

4.1. Тестовые платформы

Для анализа энергопотребления были выбраны две тестовые платформы из семейства Nvidia Jetson разных поколений, представляющие собой SoC, состоящие из нескольких ARM-ядер, графического ядра и нескольких гигабайт общей памяти, объединенных в рамках одного кристалла. Данные платформы изначально ориентированы на минимальное энергопотребление при достаточно высокой производительности ввиду ориентации на использование в энергоэффективных вычислениях и встраиваемой электронике (робототехника, БПЛА и т.д.).

Ввиду ориентации на энергоэффективность обычно данные устройства работают в режиме динамической смены частот графического ядра и контроллера памяти (DVFS). Таким образом, оценивая загрузку устройства и понижая частоты нужных компонентов при их простаивании, управляющий программный модуль DVFS позволяет существенно понизить энергопотребление аппаратного обеспечения при простое. В большинстве измерений (кроме измерений, ориентированных на оценку эффективности DVFS) данный режим был принудительно отключен путем ручного задания частот памяти и графического ядра.

4.1.1. Jetson TK1

Более старым представителем этого семейства является Nvidia Jetson TK1, представляющий собой платформу для разработки на базе 32-битного чипа Tegra K1. Данный чип содержит в себе четыре ядра ARM Cortex-A15 с 2 Мб L2 кэша, способных работать на частотах с 20 МГц до 2.3 ГГц, вместе с дополнительным маломощным ядром, которое используется в специальном режиме пониженной нагрузки (не рассматривалось). Оперативная память представлена 2 Гб Low Power DDR3, способной работать в диапазоне частот от 12 до 930 МГц. Кроме того, на кристалле размещен один модуль потокового мультипроцессора GPU архитектуры Kepler, содержащий 128 ядер CUDA и способный работать в диапазоне частот от 72 до 852 МГц. В качестве операционной системы используется Linux4Tegra r21.4, основанная на Ubuntu 14.01 (ядро 3.10.40 armv7l), вместе с GCC 4.8.4 и CUDA Toolkit 6.5.

4.1.2. Jetson TX1

Новый представитель семейства платформ разработки Jetson — Jetson TX1. Он построен на базе нового 64-битного чипа Nvidia X1. Аналогично старшей модели, чип содержит в себе как процессорную группу более мощных ядер, так и более слабых (big.LITTLE). Первая группа представлена четыремя ядрами ARM Cortex-A57, работающих в диапазоне до 1.9 ГГц с L2 кэшем размером 2 Мб. При необходимости при низкой загрузке может быть задействована более слабая вторая группа ядер — четыре ядра ARM Cortex-A53. Оперативная память представлена 4 Гб Low Power DDR4. Также чип обладает двумя потоковыми мультипроцессорами GPU архитектуры Maxwell с 256 ядрами CUDA и способен работать в диапазоне частот от 76 до 998 МГц. Операционная система — Linux4Tegra r23.1 (Ubuntu 14.01, ядро 3.10.67 aarch64), вместе с GCC 4.8.4 и CUDA Toolkit 7.0.

4.2. Метод измерения энергопотребления

Для измерения энергопотребления тестовых плат использовались цифровые ваттметры Smart Power с интегрированным источником тока. Данные устройства могут обеспечивать напряжение в диапазоне от 3 до 5.25 В, при этом измеряя силу тока и потребляемую мощность каждые 0.2 секунды с номинальной погрешностью не более 0.001 В. Полученные данные отображаются на дисплее в реальном времени, а также могут быть пересланы на компьютер по USB для записи и дальнейшего анализа.

Ввиду того, что тестируемые платформы Jetson имеют номинальное значение напряжения на поставляемых блоках питания выше 5.25 В, было принято решение соединить несколько блоков SmartPower последовательно для достижения необходимого значения напряжения. Для подтверждения точности получаемого напряжения на выходе данной схемы использовался высокоточный осциллограф Tektronix TDS2014C — средняя точность при определении мощности с помощью SmartPower имеет уровень погрешности около 1%.

Номинальное рабочее напряжение для Jetson TK1 составляет 12 В, для Jetson TX1 — 19 В. Было определено, что оба тестируемых устройства способны стабильно функционировать и при много меньших напряжениях — вплоть до 8 В (TX1) и 6 В (TK1).

Измерение энергопотребления в конкретном тесте обеспечивалось путем одновременного запуска программы-логгера на контрольном компьютере и необходимого теста на Jetson через SSH. Таким образом, к тестируемому устройству не была подключена никакая периферия, кроме сетевого кабеля, что увеличивало точность измерения.

Следует отметить, что существуют и иные методы измерения энергопотребления. Например, Intel использует в своих чипах встроенные в кристалл счетчики, которые позволяют весьма точно определять использование вычислительных ресурсов чипа и энергопотребление. Однако отсутствие подобных счетчиков в кристаллах Tegra вынуждает прибегать к методам прямого измерения потребления всей платы.

Внешний вид обеих тестовых плат, а также модулей SmartPower вместе с осциллографом приведен на рис. 2.

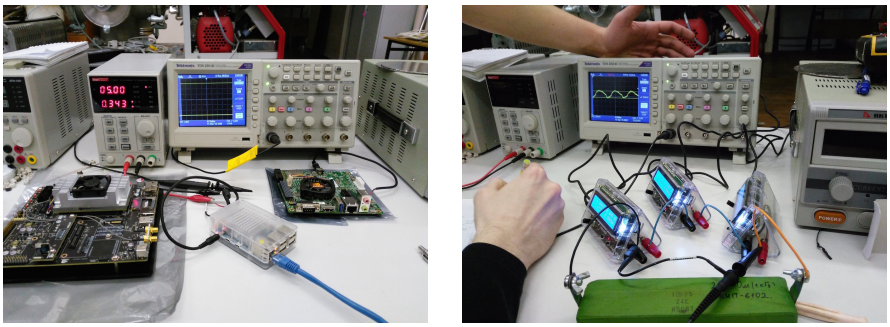

Рис. 2: Тестируемые Jetson TX1 и TK1 (и ODROID-C1 в центре), а также модули SmartPower и осциллограф

5. Результаты измерений

5.1. Энергопотребление на тесте ERT

Результаты запуска ERT использовались для определения отношения пикового значения производительности (ГФЛОПс) к среднему уровню потребления за время теста.

На рис. 3 можно видеть пример графика потребления обоих устройств во время выполнения теста. Исходя из измеренного в ходе экспериментов среднего уровня потребления энергии во время теста, а также из определенной из теста ERT пиковой производительности, можно определить энергоэффективность платформ.

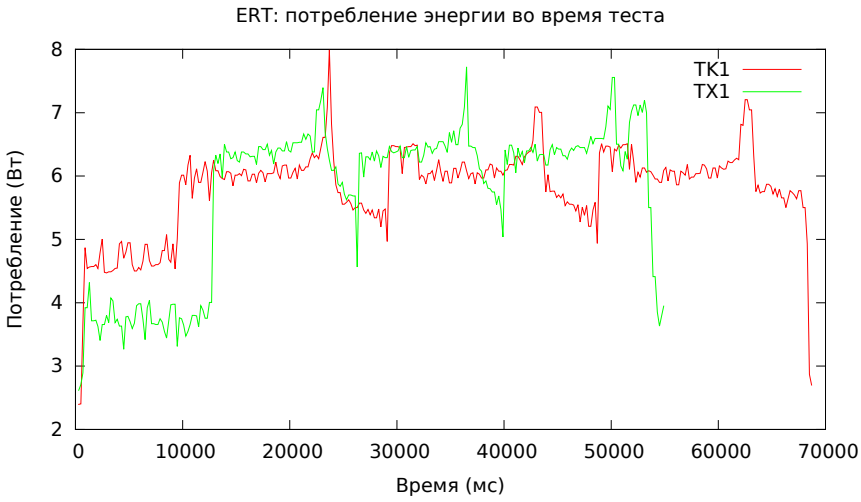

Рис. 3: Пример части графика энергопотребления обеих тестовых платформ при выполнении ERT

Ввиду того, что первые 10 секунд теста уходили на пересборку бинарного файла тестового пакета, в усреднение данная часть лога потребления не входит.

Усредненное значение потребления в ходе теста с использованием графического ускорителя для TK1 на одинарной точности составило 5.92 Вт, при достигнутом максимальном значении производительности в 209.9 ГФЛОПс, что дает соотношение в 35.5 ГФЛОПс/Вт.

Усредненное значение потребления на одинарной точности для TX1 составило 6.28 Вт, что немногим более значения, показанного TK1. Более новое устройство существенно превосходит старое в максимально зафиксированном значении производительности — 485.1 ГФЛОПс, что дает более высокое соотношение 77.2 ГФЛОПс/Вт.

Продемонстрированные устройствами результаты на двойной точности для GPU не так впечатляющи — 2.7 ГФЛОПс/Вт для TX1 и 2.1 ГФЛОПс/Вт для TK1. Причиной этому служит существенно более низкая вычислительная производительность на двойной точности обеих устройств при сопоставимом с одинарной точностью потреблении энергии.

С другой стороны, запуск ERT на CPU TX1 показал, что новые ядра ARM Cortex-A57 имеют меньшие показатели энергопотребления: 0.8 ГФЛОПс/Вт на двойной точности и 4 ГФЛОПс/Вт на одинарной точности.

5.2. Энергопотребление на тесте LAMMPS

5.2.1. Профили потребления LAMMPS

На рис. 4 можно видеть типичные профили энергопотребления при запусках LAMMPS с различными версиями МД-алгоритмов — USER-CUDA, GPU и USER-OMP. Количество израсходованной на расчет энергии представляет собой площадь под графиком мощности, при этом следует вычитать ненулевой уровень потребления в простое.

Вариант ускоренного алгоритма USER-CUDA тратит на свою инициализацию некоторое количество времени и энергии, что можно проследить на рис. 4. При продолжительном расчете вкладом инициализации в энергопотребление следует пренебрегать.

На рис. 5 показаны результаты измерений, представленные в виде общего времени расчета МД-теста как функции от израсходованной энергии и как функции средней потребляемой мощности.

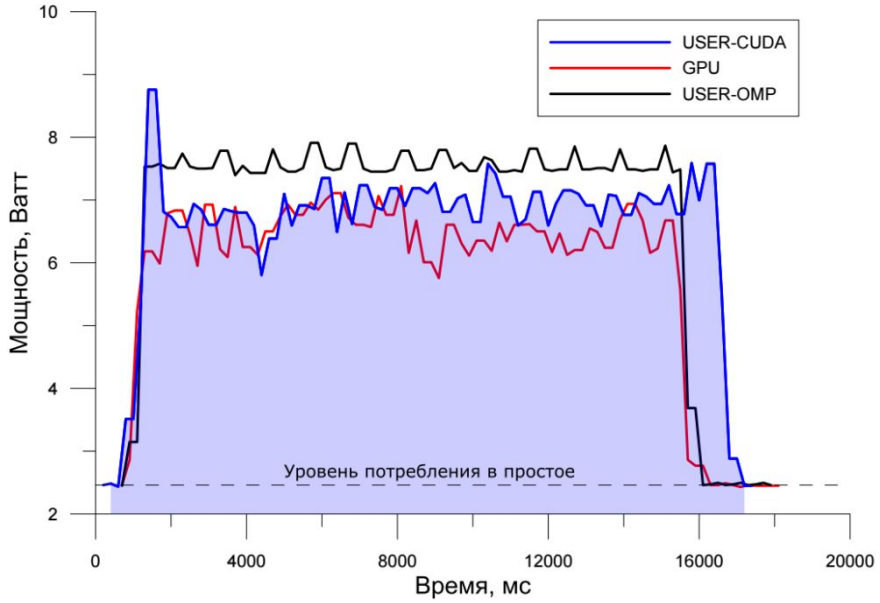

Рис. 4: Пример графика энергопотребления TX1 для отдельных запусков LAMMPS

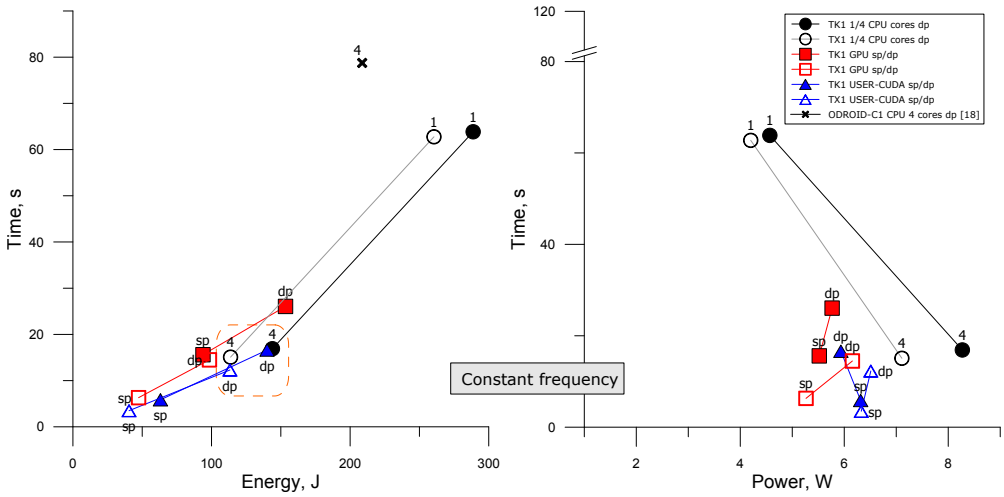

Рис. 5: Время вычисления теста LAMMPS в режиме максимальных фиксированных частот

5.2.2. Использование DVFS

Как уже было отмечено выше, обе платформы поддерживают режим оптимизации энергопотребления DVFS. Ввиду этого было решено проверить, каким образом изменятся результаты (время расчета и потребленная энергия) в этом режиме. Таким образом, для стандартного Леннард-Джонсоновского бенчмарка LAMMPS производилось два запуска — один в режиме принудительно установленных максимальных частот, другой — с включенным DVFS. Запуски производились для обоих реализаций гибридных алгоритмов (GPU и USER-CUDA) в одинарной и двойной точности.

На рис. 6 можно увидеть сравнение энергопотребления LAMMPS в режиме DVFS и для максимальных частот. Видно, что запуск в режиме динамической смены частоты занимает дольше времени, хотя и потребляемая мощность явно ниже.

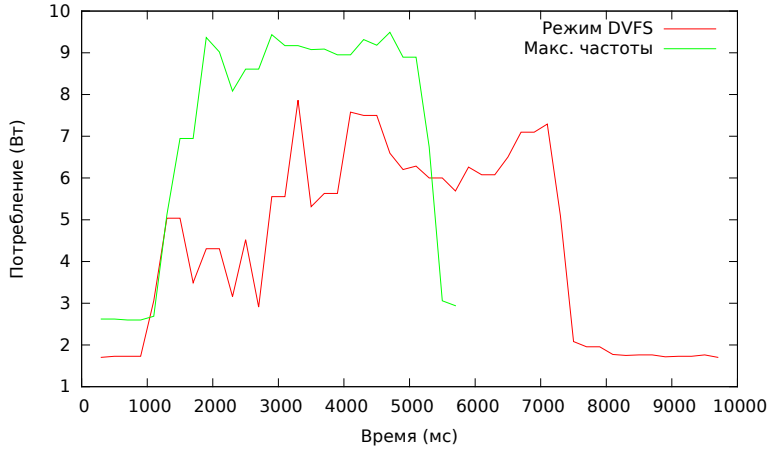

Рис. 6: Сравнение энергопотребления TX1 в тесте LAMMPS с USER-CUDA двойной точности: случай DVFS и случай максимальной фиксированной частоты

Для ответа на вопрос, оправдан ли такой подход, описанным выше методом был произведен подсчет затраченной энергии. На рис. 7 представлены результаты измерений.

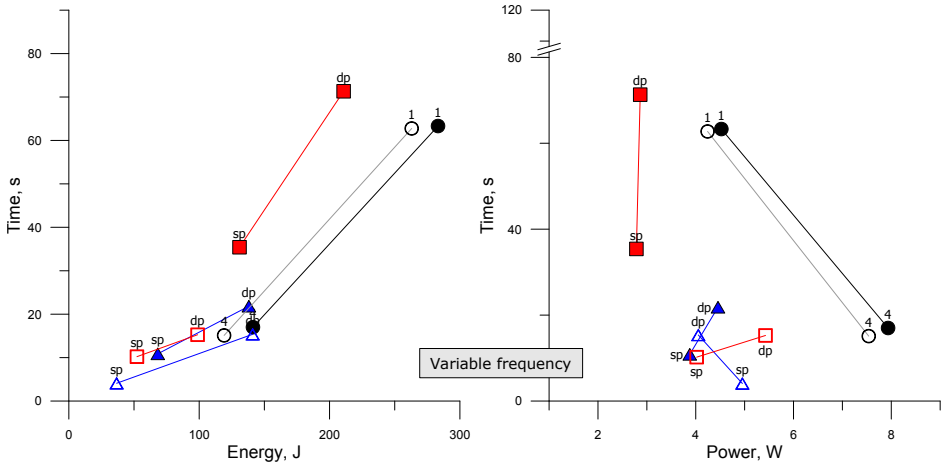

Рис. 7: Время вычисления теста LAMMPS в режиме DVFS

Анализ результатов показал, что итоговое потребление энергии за весь расчет в режиме DVFS примерно равно, либо выше аналогичного значения для режима с максимальной фиксированной частотой при стабильно большем времени на расчет.

6. Переход между compute-bound и memory-bound режимами и энергопотребление

Для определения режима вычисления на TK1 были проведены расчеты МД теста с помощью пакета USER-CUDA в LAMMPS. При этом частота графического ядра менялась от расчета к расчету, а частота памяти фиксировалась для всей группы запусков. Для Jetson TK1 в каждой группе экспериментов частота GPU последовательно устанавливалась на следующие значения (в МГц): 72, 108, 180, 252, 324, 396, 468, 540, 612, 648, 684, 708, 756, 804 и 852. Запуск производился на частотах памяти в 924, 396, 204 и 102 МГц. Одновременно с МД расчетом проводился замер энергопотребления.

Результаты измерений показаны на рис. 8 и рис. 9. Видно, что, хотя повышение частоты GPU сопровождается понижением времени расчета, с точки зрения энергопотребления ситуация несколько иная — потребление чипа понижается при повышении частоты графического ядра до некоторого предела, потом достигает минимума, а затем, дальнейшее повышение частоты акселератора приводит к некоторому повышению энергопотребления.

Достижение минимума связано с переходом алгоритма LAMMPS из compute в memory-bound режим. В режиме малых частот GPU пониженная производительность графического ядра лимитирует вычисления, что отвечает compute-bound режиму. Наоборот, в режиме высоких частот графического ядра лимитирующим фактором является пропускная способность памяти, что соответствует memory-bound режиму. Повышение потребления в этом режиме не является желательным эффектом, т.к. ускорение от повышения частоты GPU выше оптимальной почти отсутствует, а энергопотребление становится выше. Также следует отметить, что понижение частоты памяти приводит к сдвигу оптимальной точки в область более низких частот GPU, что и следовало ожидать.

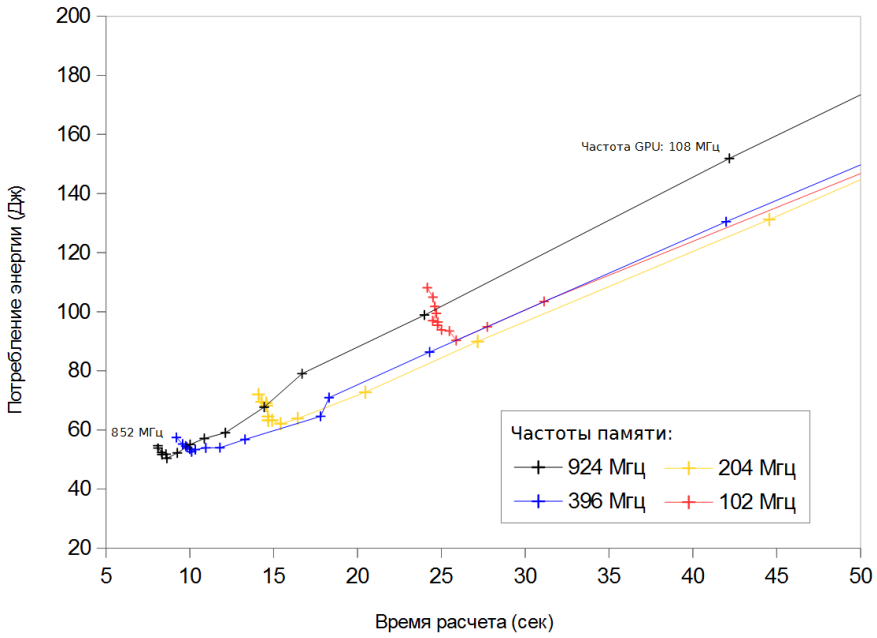

Рис. 8: Потребление LAMMPS при разных частотах памяти (TK1) с пакетом USER-CUDA

7. Заключение

Была произведена оценка энергопотребления миникомпьютеров Nvidia Jetson TK1 и TX1, основанных на гибридных системах-на-чипе Tegra K1 и X1, соответственно.

Пиковые режимы нагрузки были реализованы в рамках тестовых запусков Empirical

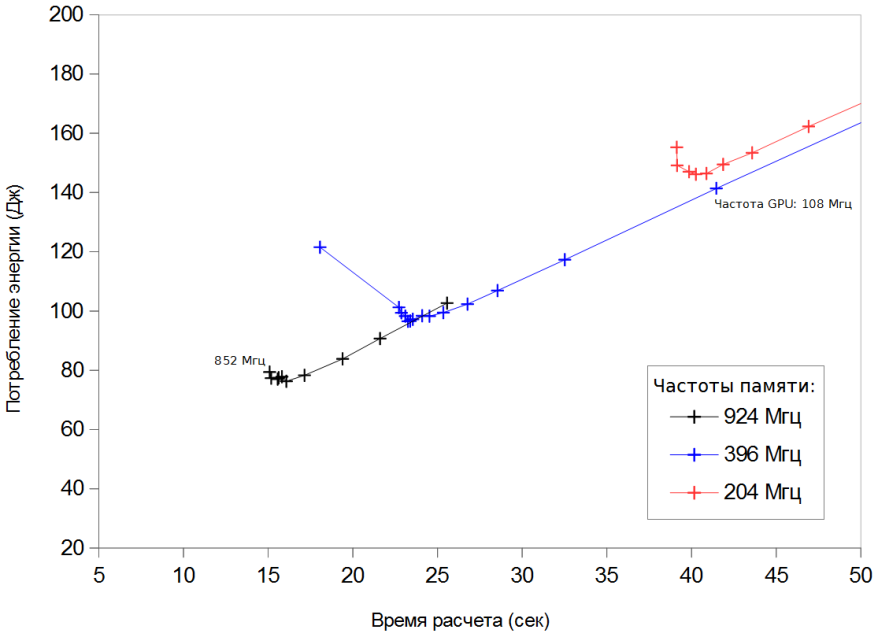

Рис. 9: Потребление LAMMPS при разных частотах памяти (TK1) с пакетом GPU

Roofline Tool в CPU и GPU версиях. В CPU версии на 1 ядре для Jetson TX1 были получены значения 0.8 ГФЛОПс/Вт на двойной точности и 4 ГФЛОПс/Вт на одинарной. В GPU версии для Kepler на TK1 и Maxwell на TX1 были получены значения 2.1 и 2.7 ГФЛОПс/Вт в двойной точности и 35.5 и 77.2 ГФЛОПс/Вт в одинарной точности.

Было сопоставлено энергопотребление для трех вариантов молекулярно-динамических алгоритмов пакета LAMMPS (один на основе OpenMP и два варианта на основе CUDA). Были сопоставлены режимы одинарной и двойной точности. Путем варьирования частоты GPU и частоты оперативной памяти был показан переход обоих рассматриваемых гибридных МД-алгоритмов из memory bound в compute bound режим. Было показано существование минимального значения энергопотребления.

В будущем планируется провести аналогичный анализ для систем с обычным графическим ускорителем (GeForce или Tesla). Кроме того, планируется рассмотреть случай более сложных молекулярно-динамических моделей, например, с кулоновским взаимодействием.

Аппаратное обеспечение, использованное в данной работе, было приобретено при поддержке МФТИ и ВШЭ. Работа поддержана грантом РНФ 14-05-00124.

Литература

1. Su C.L., Tsui C. Y., Despain A. M. Low power architecture design and compilation techniques for high-performance processors // Compcon Spring'94, Digest of Papers, IEEE, 1994, 489-498.

2. Joseph R., Martonosi M. Run-time power estimation in high performance microprocessors // Proceedings of the 2001 international symposium on Low power electronics and design, ACM, 2001, 135-140.

3. Russell J. T., Jacome M. F. Software power estimation and optimization for high performance, 32-bit embedded processors // Computer Design: VLSI in Computers and Processors, ICCD'98, 1998, 328-333.

4. Li T., John L. K. Run-time modeling and estimation of operating system power consumption // ACM SIGMETRICS Performance Evaluation Review, 2003, T. 31, 1,

160-171.

5. Zhang L. et al. Accurate online power estimation and automatic battery behavior based power model generation for smartphones // Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, ACM, 2010, 105-114.

6. Никольский В., Стегайлов В. Эффективность процессоров архитектуры ARM для расчетов классической молекулярной динамики // Вычислительные методы и программирование, 2015, Т. 16. С. 578–585.

7. Calore E., Schifano S. F., Tripiccione R. Energy-Performance Tradeoffs for HPC Applications on Low Power Processors // Euro-Par 2015: Parallel Processing Workshops, Springer International Publishing, 2015, 737-748.

8. Lopez-Novoa U., Mendiburu A., Miguel-Alonso J. A survey of performance modeling and simulation techniques for accelerator-based computing // Parallel and Distributed Systems, IEEE Transactions on. – 2015. – Т. 26. – №. 1. – С. 272-281.

9. Li S. et al. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures // Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture. – ACM, 2009. – С. 469-480.

10. Leng J. et al. GPUWattch: enabling energy optimizations in GPGPUs // ACM SIGARCH Computer Architecture News. – 2013. – Т. 41. – №. 3. – С. 487-498.

11. Scogland T. et al. Node variability in large-scale power measurements: perspectives from the Green500, Top500 and EEHPCWG // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. – ACM, 2015. – С. 74.

GPU-accelerated molecular dynamics: energy consumption and performance

V.S. Vecher[1,2],V.P. Nikolsky[3,2], V.V. Stegailov[2]

MIPT[1], JIHT RAS[2], NRU HSE[3]

Energy consumption of hybrid systems is an actual problem of modern high-performance computing. The trade-off between the power consumption and performance becomes more and more prominent. In this paper we discuss the energy and power efficiency of two modern hybrid minicomputers Jetson TK1 and TX1. We use the Empirical Roofline Tool to obtain peak performance data and the molecular dynamics simulator LAMMPS as a real application example. Using precise wattmeter we measure Jetsons power consumption profiles at different frequency modes of DRAM and GPU. The effectiveness of DVFS has been examined as well. We determine the optimal GPU and DRAM frequencies from the point of minimum energy use.

Keywords: Tegra, Kepler, Maxwell, ARM, energy efficiency, molecular dynamics

References

1. Su C.L., Tsui C. Y., Despain A. M. Low power architecture design and compilation techniques for high-performance processors// Compcon Spring'94, Digest of Papers, IEEE, 1994, 489-498.

2. Joseph R., Martonosi M. Run-time power estimation in high performance microprocessors// Proceedings of the 2001 international symposium on Low power electronics and design, ACM, 2001, 135-140.

3. Russell J. T., Jacome M. F. Software power estimation and optimization for high performance, 32-bit embedded processors// Computer Design: VLSI in Computers and Processors, ICCD'98, 1998, 328-333.

4. Li T., John L. K. Run-time modeling and estimation of operating system power consumption// ACM SIGMETRICS Performance Evaluation Review, 2003, T. 31, 1, 160-171.

5. Zhang L. et al. Accurate online power estimation and automatic battery behavior based power model generation for smartphones// Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis, ACM, 2010, 105-114.

6. Nikolsky V., Stegailov V. Efficiency of ARM processors for classical molecular dynamics calculations // Book of Abstracts: International Conference on Computer Simulation in Physics and beyond. 2015.

7. Calore E., Schifano S. F., Tripiccione R. Energy-Performance Tradeoffs for HPC Applications on Low Power Processors// Euro-Par 2015: Parallel Processing Workshops, Springer International Publishing, 2015, 737-748.

8. Lopez-Novoa U., Mendiburu A., Miguel-Alonso J. A survey of performance modeling and simulation techniques for accelerator-based computing //Parallel and Distributed Systems, IEEE Transactions on. – 2015. – Т. 26. – №. 1. – С. 272-281.

9. Li S. et al. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures //Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture. – ACM, 2009. – С. 469-480.

10. Leng J. et al. GPUWattch: enabling energy optimizations in GPGPUs //ACM SIGARCH Computer Architecture News. – 2013. – Т. 41. – №. 3. – С. 487-498.

11. Scogland T. et al. Node variability in large-scale power measurements: perspectives from the Green500, Top500 and EEHPCWG //Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. – ACM, 2015. – С. 74.