Оптимизация вычислений по электропотреблению

И. Захаров[1], О. Панарин[1]

[1] Сколковский институт науки и технологий
Территория Инновационного Центра "Сколково", улица Нобеля, д. 3
Москва 121205 Россия

Аннотация. Оптимизация функции ввода/вывода и обращения к памяти для Высоконагруженных вычислений по скорости и по энергоэффективности. Проведенное исследование показывает, что для существенной части работы вычислительного устройства увеличение тактовой частоты процессора не приводит к большей вычислительной производительности. Периферийные устройства работают заведомо медленнее центрального процессора. Поэтому скорость работы периферийных устройств определяет скорость выполнения операций и тактовая частота процессора может поддерживаться на более низком уровне достаточном для загрузки этих устройств. В работе мы приходим к необходимости получения специфичной метрики характеризующей вычислительный процесс для настройки оптимальной тактовой частоты вычислителя.

Ключевые слова: Высоконагруженные вычисления, Энергоэффективность

Application Energy Efficiency Optimization

I. Zacharov[1], O. Panarin[1]

[1] Skolkovo Institute of Science and Technology
Skolkovo Innovation Center, 3 Nobel Street, Moscow 121205, Russia

Abstract. Efficiency optimization for the input/output functions and memory access for the HPC applications requires simultaneous considerations for the computational speed and the electricity consumption. The proposed study shows that for the substantial part of the calculations the increase of the clock frequency of the CPU does not lead to the increase of the application execution speed. Peripheral units work at a lower rate than the CPU. Therefore the speed of the peripheral devices will determine the speed of the application that uses them and the CPU clock may be adjusted to a lower rate with corresponding reduction of electricity consumption. In this work we arrive at the need for an application specific metrics that describes the computational advance for setting the CPU clock frequency to the optimal rate.

Keywords: High Performance Computing, Energy Efficiency, DVFS

В высоконагруженных вычислениях (ВВ) традиционно внимание уделялось методам решения оптимизированных по времени решения задачи. При этом полностью игнорировался аспект использования электроэнергии для получения этого решения.

В современных центрах обработки данных (ЦОД) и центрах коллективного пользования потребление электроэнергии играет существенную роль в оперативных расходах центров. Так же, потолок поставляемой электроэнергии ограничивает вычислительную мощность установленных устройств.

Энергия, потребляемая вычислительными устройствами, является основным параметром оптимизации вычислительного процесса и предоставляемых услуг в интернете. Исследование ЦОД Google [1] показало, что сервера в комьютерном центре в среднем работают при 10-50% загрузки. Хотя электропотребление уменьшается при уменьшении нагрузки, это уменьшение не прямо пропорционально уменьшению нагрузки, и сервера потребляют до 50% пиковой мощности, не производя никакой вычислительной работы. Разработаны методы борьбы с этим явлением, включающим виртуализацию серверов [2]. При этом множество сервисов консолидируется на одном физическом сервере, увеличивая утилизацию и соответственно энергоэффективность.

Такое направление развития не удовлетворяет требованиям ВВ и даже требованиям обработки баз данных. Для последних, так же характерно условие увеличения энергоэффективности при увеличении производительности [3], что идет вразрез с консолидацией сервисов на одной физической платформе.

Оптимизация приложений по энергопотреблению и по скорости вычислений особенно важны на мобильных платформах. Скорость вычислений критична для обработки данных на автономных автомобилях и в то же время на таких платформах есть серьезное ограничение по потребляемой электрической мощности. В частности, эксперименты с оптимизацией процесса ввода/вывода проведенные в данном исследовании диктовались требованиями мобильных приложений.

Для обеспечения энергоэффективности ВВ предлагается путь увеличения эффективности самих вычислений. Для этого проводятся исследования эффективности вычислений как функции потребляемой энергии. В данной статье приведено отдельное исследование высокопроизводительной записи данных, являющейся компонентом многих вычислительных пакетов и систем анализа Больших данных а так же системы оперативной памяти.

Утилизация процессоров и Энергоэффективность вычислений

Высоконагруженные вычисления имеют повышенные требования к производительности компьютера и при выполнении ВВ компьютер загружен полностью. Это можно охарактеризовать уровнем утилизации (U), который

определяется через отношение времени простоя T_{idle} к времени использования T_{tot}:

$$U = 1 - \frac{T_{idle}}{T_{tot}} \quad (1)$$

при этом в операционной системе (ОС) Linux время простоя и другие характеристики, накопленные с начала запуска системы, регистрируется ОС в метрике `/proc/stat`.

Полная утилизация не означает, что вычисления обладают высокой степенью эффективности, в частности по потреблению электроэнергии. Эффективность потребления электроэнергии (EE) выражается через произведенную работу:

$$EE = \frac{\text{Работа приложения}}{\text{Электрическая Энергия}} = \frac{\text{Работа приложения}}{Power\ x\ Time} = \frac{\text{Скорость приложения}}{Power} \quad (2)$$

при этом *Работа приложения* характеризует колличество произведенных операций и С*корость приложения* - колличество операций в секунду, соответственно. Так для ВВ с плавающей точкой скорость выражается через Flop/s, а мощность измеряется в Ваттах. По этой метрике для бенчмарка Linpack составляется список наиболее экономных систем Green 500 [4], где первое место на ноябрь 2017 года занимает японская система "Shoubu system B" со значением 17 GFlops/W. Для ВВ не ориентированных на операции с плавающей точкой *Скорость приложения* может характеризоваться в других единицах. На пример HPCC RandomAccess измеряет "global updates per second" (GUPs) [5], а GRAPH500 измеряет "traversed edges per second" (TEPS) [6].

Управление энергопотреблением осуществляется через установку рабочей частоты процессоров с механизмом динамической зависимости DVFS. Это возможно на всех типах процессоров Intel (автоматически и командой `cpufreq-set`) и графических процессорах Nvidia (автоматически и командой `nvidia-smi`). Так система настройки потребляемой мощности процессора Intel RAPL допускает регулировку в широких пределах [5]. Подстройка тактовой частоты из операционной системы осуществляется соответствующим планировщиком по показанию загруженности системы через череду значений таблицы (P-states) связывающей частоту и рабочее напряжение по данным производителя. Однако эта регулировка не охватывает другие подсистемы, такие как оперативная память, системы хранения, сети и т.д., которые могут потреблять до 50% всей поставляемой мощности. Управление этими ресурсами требует отдельного исследования.

Для представленой работы использовалась рабочая станция с одним процессорм Intel E3-1281 и графической картой Nvidia GX1050. Это

оборудование было выбрано из-за удобства использования интегрированного датчика энергопотребления и для отработки методов измерения эффективности вычислений на всех устройствах, как то процессор, графическая карта и системы хранения. Основные параметры потребления рабочей станции приведены в таблице:

Устройство / Потребление	**min[W]**	**max[W]**
1 x Intel E3-1281 (4 cores), 0.7 - 3.7 GHz DVFS	9	92
Nvidia GTX 1050	35	75
2 x NVMe SSD storage	18	50
Mellanox ConnectX-4 card	10	10
Board и преобразователи энергии	25	50
Total	**87**	**277**

Колонки min и max показывают разброс параметров потребления без нагрузки и при полной нагрузке соответственно. В данной работе не исследовалось управление потреблением периферийных устройств, так как эта работа запланирована на более поздний срок, и внимание уделено только управлению энергопотреблением процессора изменением тактовой частоты. .

Зависимость полной потребляемой мощности от частоты процессора исследовалась с использованием регистров процессора RAPL и интегрированного датчика потребления всего компьютера. Результаты показаны на рис. 1.

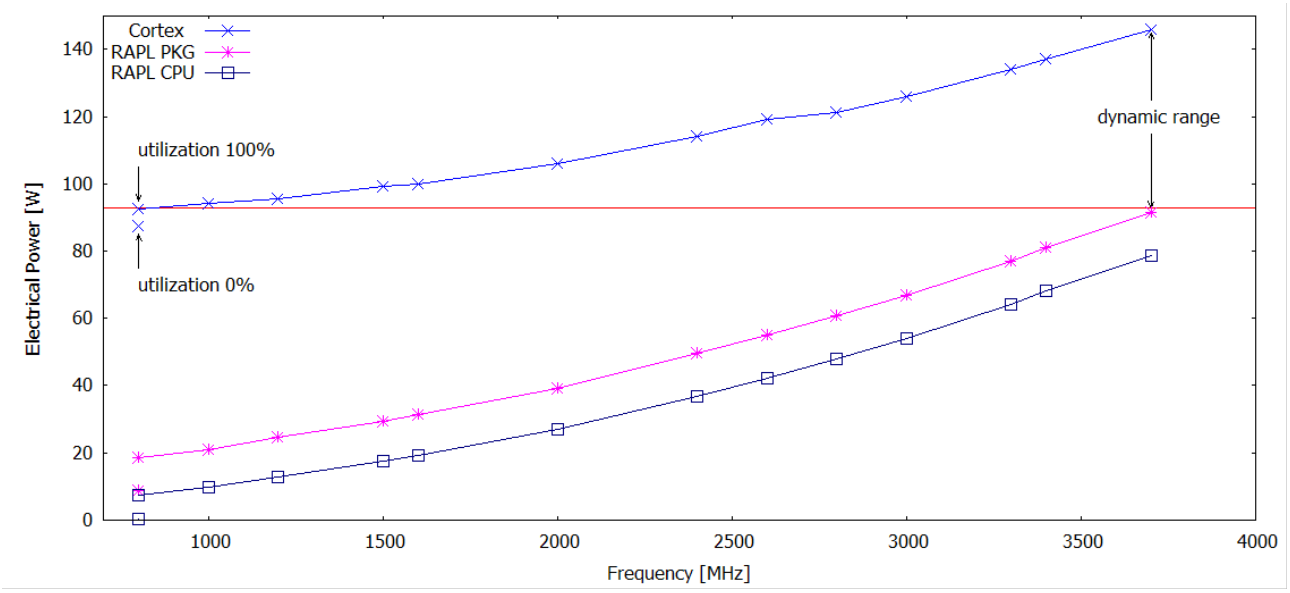

Рис. 1. Зависимость потребляемой мощности от рабочей частоты процессора. Синий график (Cortex) показывает общее потребление всего компьютера.

Когда компьютер не нагружен (утилизация равна нулю) процессор имеет дополнительные ступени ожидания, опускающие потребление вычислительных ядер почти до нуля. Потребление пакета, состоящего из различных

преобразователей энергии, процессора и памяти, спускается до 9 W (из них память потребляет 4.5 W), независимо от выставленной частоты. Этот режим обозначен на рис. 1. При частоте 700 МГц и утилизации 100% процессорные ядра потребляют 7 W, весь пакет потребляет 19 W. С этого режима потребление процессорных ядер плавно увеличивается с увеличением частоты, что видно на рис. 1.

Хотя процессорная мощность варьируется средствами DVFC (тактовой частотой процессора) в пределах 9-92 W, для потребления всей рабочей станции эта вариация составляет только 54 W (см. "Dynamic range" на рис. 1) из за того, что потребление всего компьютера не опускается ниже определенного порога задаваемого нерегулируемыми компонентами.. Таким образом диапазон управления энергопотреблением процессора составляет только от 19% до 36% потребления всего компьютера если включены или не-включены периферийные устройства в расчет.

Оптимизация приложений по Энергоэффективности

Оптимизация приложений по метрике Энергоэффективности (EE) начинается с определения параметра характеризующего результат работы приложения как функции времени. Это параметр С*корость приложения* в уравнении (2). Для сложного пакета вычислений, в отличие от простых бенчмарков, этот параметер может быть разным в разных фазах работы программы.

В приложении выбраном для исследования параметром характеризующим результат работы является скорость записи в единицах ГБ/с. Оптимизация производится как по времени (максимальная скорость записи уменьшает время использования), так и по энергопотреблению.

В проведенном исследовании была изолированна функция записи данных на отдельном (логическом) процессорном ядре, для чистоты эксперимента программа фиксировалась на этом ядре (CPU-0) и другие нагрузки на этом (физическом) ядре не допускались. Запись данных осуществлялась через библиотеку асинхронных вызовов, что бы максимально задействовать периферийное устройство и полностью разгрузить процессор. В таком исполнении пишущий процесс передает список готовых к записи буферов (размером 1 МБ) операционной системе и затем вызывает (блокирующую) функцию ожидания отчета о выполнении задания. При блокировке, в отсутствие других процессов на этом ядре, ОС включает счетчик времени простоя и мы считаем утилизацию по формуле (1).

Управление электропотреблением компьютера осуществлялось через установку рабочей частоты процессора (команда `cpufreq-set`). Показатели утилизации и скорости записи как производной от рабочей частоты процессора приведены на рисунке 2.

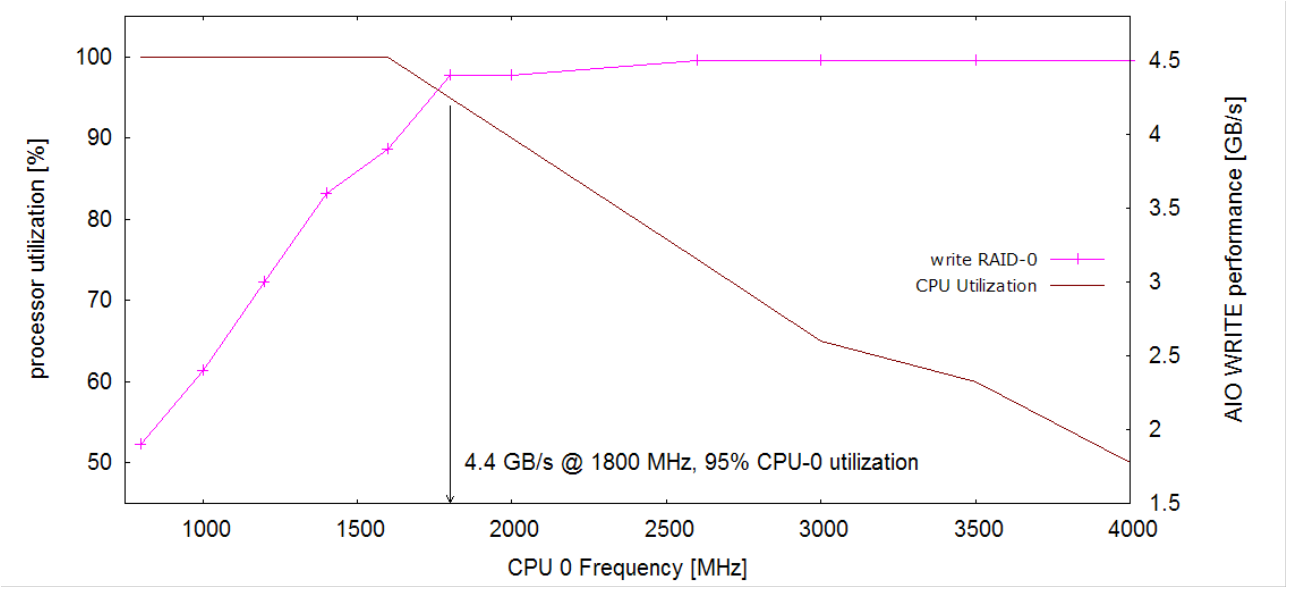

Рис. 2. Утилизация процессора и скорость записи на внешний носитель как функция рабочей частоты исполняющего процессора.

Оптимизация выполнения задания по времени требует максимальной скорости записи. Сама скорость записи зависит от производительности внешнего устройства (макс. 4.4 ГБ/с). Как описано выше, в данном алгоритме программа формирует список буферов и блокируется в ожидании отчета о записи. Скорость формирования списка зависит от рабочей частоты процессора и для оптимальной работы должна соответствовать времени получения отчета о записи, что бы время блокировки было минимальным.

На ординате справа (рис. 2) показана скорость записи и красная кривая (обозначенная "write RAID-0") растет пропорционально частоте процессора до отметки 1.8 ГГц. На этой отметке скорость записи достигает максимума. При этом утилизация (ордината слева) максимальна и на данной отметке снижается до 95%, т.е. в таком режиме процессор успевает сформировать список, в то время как периферийное устройство записывает переданные раньше буфера, и время блокировки оказывается минимальным.

В таком режиме общее потребление компьютера оказывается около 100 W и приложение записи данных оптимизированно как по времени, так и по энергоэффективности.

В ходе эксперимента было установленно, что по умолчанию Операционная система поднимает частоту процессора до максимальных значений когда запускается программа пользователя, хотя это не обязательно для достижения максимальных характеристик производительности.

В данном эскперименте для автоматической регулировки частоты можно было ориентироваться на показатель утилизации. В современных ядрах Линукс (начиная с 4.0) есть возможность выбора планировщика, который учитывает

утилизацию процессора и регулирует тактовую частоту по этой характеристике [9]. Этого, однако, недостаточно для общего случая программ, активно использующих процессор. Процессор не будет находится в режиме ожидания, если ему предложена дополнительная нагрузка, а запись данных может идти в фоновом режиме. При определении оптимальной частоты работы процессора должны учитываться скорости по метрике определяемой конкретным приложением.

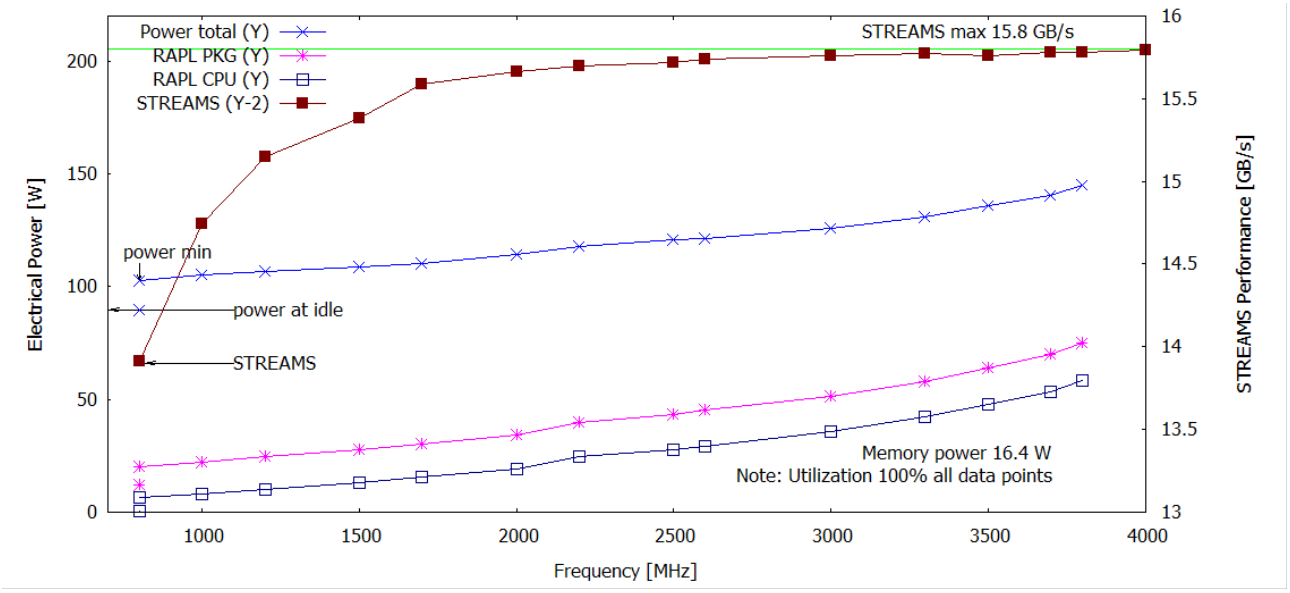

Рис. 3. Электрическая мощность и скорость доступа к памяти по замеру теста Triad из пакета STREAMS как функция тактовой частоты процессора.

График производительности теста STREAMS Triad: $a(i) = b(i) + qc(i)$ показывает (см. рис. 3), что 99% производительности памяти достигается при частоте 1.7 ГГц и не требует полной тактовой частоты процессора.В данном эксперименте было показано, что регулировка тактовой частоты должна производиться по характеристике *Скорости приложения* (см. уравнение 2), экономя при этом электроэнергию.

Для интерпретации полученных результатов были произведены замеры вычислений Triad в ситуации, когда вычисления производились из данных в регистрах процессора. В таком случае скорость вычислений прямо пропорциональна тактовой частоте процессора, т.к. полностью независимы от периферийных устройств.

К интерпретации полученных результатов следует добавить, что есть возможность управлять потреблением процессора через установку тактовой частоты, но потребление других компонент может зависеть от нагрузки и не регулируется из Операционной системы. Так, потребление системы памяти (DIMMS) в эксперименте составляла 10-16 Ватт, системы хранения данных NVMe на которые записываются данные около 50 Ватт, система графической обработки 35 Ватт (в режиме ожидания, т.к. ГПУ не использовалась в эксперименте).

Заключение

В работе показан метод оптимизации вычислительного устройства по времени (скорости) записи на внешний носитель, доступа к оперативной памяти и по энергоэффективности. Метод использует скалирование рабочей частоты процессора при неизменной характеристике выполняемой задачи. Это только часть нашего исследования по увеличению эффективности Высоконагруженных вычислений по времени выполнения задания и по энергоэффективности.

При практическом применении данного подхода необходимо разработать метод определения характеристики задачи, которая должна быть соблюдена при установлении тактовой частоты процессора.

References

1. L.A. Barroso, U. Holzle, The Case for Energy-Proportional Computing, Computer Vol. 40 Issue 12, 2007.
2. Y. Jin, Y. Wen, Q. Chen, Energy efficiency and server virtualization in data centers: An empirical investigation, IEEE Xplore 03 May 2012 DOI: 10.1109/INFCOMW.2012.6193474
3. D. Tsirogiannis, S. Harizopoulos, M. Shah, Analyzing the Energy Efficiency of a Database Server, SIGMOD’10, June 6-11, 2010.
4. https://www.top500.org/green500/list/2017/06/
5. http://icl.cs.utk.edu/hpcc/
6. https://graph500.org/
7. Intel 64 and IA-32 Architectures Software Developer’s Manual, Vol. 3B: System Programming Guide, Part 2, https://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-vol-3b-part-2-manual.pdf
8. McCalpin, John D., 1995: "Memory Bandwidth and Machine Balance in Current High Performance Computers", IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter, December 1995. Available at: http://www.cs.virginia.edu/stream/
9. https://www.kernel.org/doc/html/v4.14/admin-guide/pm/cpufreq.html