Оценка энергопотребления в программе комплексного тестирования производительности вычислительного кластера

О. И. Вдовикин[1], П. Н. Телегин[2], Б. М. Шабанов[3]

[1]МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН, НИЦ «Курчатовский институт», Москва, Россия, vdovikin@jscc.ru;

[2]МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН, НИЦ «Курчатовский институт», Москва, Россия, telegin@jscc.ru;

[3]МСЦ РАН – филиал ФГУ ФНЦ НИИСИ РАН, НИЦ «Курчатовский институт», Москва, Россия, shabanov@jscc.ru.

Аннотация. Для сравнительной оценки производительности вычислительных кластеров используются различные группы тестов. Для определения параметров эффективности кластерных решений с большим числом ядер разработана программа комплексного тестирования производительности Laptest. Из-за высокого потребления электроэнергии суперкомпьютерами энергоэффективность при выполнении задач стала являться одним из важнейших свойств вычислительных систем. В статье описывается реализация оценки потребления энергии в программе Laptest.

Ключевые слова: суперкомпьютер, вычислительный кластер, тестирование производительности, энергоэффективность, RAPL

1. Введение

Эффективность работы вычислительных систем зависит от архитектуры систем и базового программного обеспечения (ПО): системного, промежуточного, прикладного. Для каждой архитектуры и ПО производительность определяется классами задач и реализацией прикладных программ. В связи с этим для оценки эффективности систем используются различные группы тестов.

Традиционно наиболее важным показателем производительности вычислительных систем являются операции над 64-разрядными числами. Обычно используются 3 вида тестов: стандартные тесты, программы пользователей и синтетические тесты. При этом тестируются три уровня, представленные в Таблице 1.

Тесты SPEC (Standard Performance Evaluation Corporation [1]) оценивают производительность вычислительной системы. Производительность на операциях с плавающей запятой оценивается наборами тестов SPECfp2000 (14 тестовых программ), SPECfp2006 (17 программ), SPEC CPU2017 (10 тестов, поддерживающих OpenMP).

Таблица 1. Уровни тестов

Уровень	Реализация	Тесты
Операции над структурами данных	Ядро процессора	STREAM Memory Test Suite SPEC
Нити (threads)	Узел кластера	SPEC DGEMM
Процессы	Кластер	NAS HPL программы пользователей

Тесты NAS Parallel Benchmarks [2] оценивают производительность параллельных вычислительных систем на программах из аэрокосмической области, в том числе с использованием среды MPI расширения языков OpenMP.

Тест HPL [3] используется при составлении рейтингов TOP500 [4], и TOP50 [5]. Тест основан на решении системы линейных уравнений методом LU-разложения. HPL использует среду MPI, библиотеку линейной алгебры BLAS [6] и применяется для сравнения суперкомпьютеров [7]. В рейтинг TOP500 также используется тест HPCG (High Performance Conjugate Gradients) [8]

для работы с разреженными матрицами, что характерно для многих реальных задач [9]. Тем не менее в списке Top500 представлены результаты только для 118 суперкомпьютеров из 500.

Важное значение для производительности имеет то, как система работает с оперативной памятью. Тесты STREAM, Memory Test Suite выполняют как простые операции чтения/записи, так и операции с векторной обработкой данных [10].

Подпрограмма умножения матриц DGEMM из библиотеки BLAS также часто используется в качестве теста, так как эффективно использует возможности аппаратуры и реализована практически для всех платформ.

В набор HPC Challenge Benchmark [11] были собраны тесты, наиболее часто используемые для оценки и анализа компонентов высокопроизводительных систем, включая вычислительные тесты (HPL, DGEMM, FFT), тесты работы с памятью и межпроцессорных коммуникаций.

Стоит также отметить, что в связи с бурным развитием технологий искусственного интеллекта был разработан тест HPL-MxP (Mixed-Precision Benchmark) [12], использующий в работе 16-разрядные числа. При этом результаты использования теста существенно зависят от входных данных [13].

Из-за большой электрической мощности суперкомпьютеров все более актуальными являются вопросы энергопотребления [14]. Например, в списке Top500 отдельно ведется список Green500 – наиболее энергоэффективных суперкомпьютеров. В набор тестов SPEC CPU2017 добавлено измерение максимальной и средней мощности, а также потребленной задачами энергии.

Главным недостатком стандартных тестов является то, что задачи пользователей, как правило, не укладываются в рамки тестов ни по используемым алгоритмам, ни по реализации.

2. Программа комплексной оценки производительности вычислительного кластера

В МСЦ РАН создавались вычислительные системы серии МВС: МВС-1000М, МВС-15000ВМ, МВС-6000IM, МВС-100К, МВС-10П/*. При выборе системных решений одним из важнейших параметров были результаты тестов производительности вычислительных платформ. В разработанной методике выбора суперкомпьютерных платформ предусмотрено тестирование компонентов до создания вычислительных систем [15].

Стандартные тесты не всегда дают необходимые данные для выбора платформ, так как реальные программы вариативны. Поэтому был разработан тест LAPTEST [16], оценивающий стандартные действия на компонентах вычислительный системы, реализующий методику [15]. Наиболее близкий тест – это HPC Challenge Benchmark. Однако, потребовалась большая полнота тестирования [17]. Кроме того, для имитации реальной нагрузки желательно иметь возможность обеспечивать одновременно разную загрузку вычислительных узлов и их групп. Тест ориентирован на анализ вычислительных кластеров с многоядерными процессорами.

Стандартные действия разделены на типовые элементы по тестированию как оперативной памяти, так и связки процессор-память, дисковых операций, сетевых обменов, загрузки процессора.

Рассмотрим детальнее.

Тестирование производительности памяти. Возможны следующие варианты выполнения:

- чтение;
- запись;
- одновременные чтение/запись.

Операции могут выполняться над регулярными или иррегулярными данными, с одним или двумя потоками данных при чтении.

Тестирование производительности коммуникационной сети с использованием среды MPI. Варианты выполнения:

- передача данных (send);
- прием данных (recv);
- одновременные передача/прием (sendrecv);
- широковещательная передача данных (bcast).

Тестирование производительности при работе с дисковой подсистемой. Варианты выполнения:

- чтение;
- запись;
- одновременные чтение/запись.

Тестирование производительности ядер под нагрузкой. Вызывается подпрограмма DGEMM. Данная подпрограмма сильно оптимизирована практически для всех вычислительных платформ, поэтому показывает близкую к максимально возможной производительность, а также задействует значительную часть исполнительных устройств процессора, что повышает его температуру и позволяет, кроме прочего, оценить реализацию систем охлаждения. Возможно тестирование с использованием внешних подключаемых библиотек. В настоящее время реализован интерфейс для одновременного использования до восьми реализаций DGEMM. Библиотечные функции настраиваются в конфигурационном файле.

Ожидание (sleep). Используется для организации простоя процессора.

Тест предназначен для оценки характеристик вычислительных кластеров с большим числом ядер, в том числе коэффициентов замедлений при одновременной работе различного числа ядер.

Тесты имеют гибкую структуру и настраиваются с помощью конфигурационного файла. В нем можно задавать набор тестов, их раскладку по узлам и ядрам, варианты тестирования, длительность и объем данных.

3. Измерение энергопотребления

Так как рост производительности вычислительных систем носит не только интенсивный, но и экстенсивный характер, становится необходимым кроме измерения параметров, непосредственно описывающих производительность, оценивать также и энергоэффективность вычислительных систем. В связи с этим в синтетический тест Laptest были добавлены средства измерения энергопотребления при выполнении как задач, так и отдельных операций.

3.1. Способы измерения энергопотребления

В настоящее время кроме «аналогового» способа измерения энергопотребления с использованием ваттметра или токовых клещей, производителями высокопроизводительных систем стали использоваться интегрированные в платформу варианты для измерения потребления как системы в целом, так и отдельных ее компонент, таких как центральный микропроцессор и оперативная память. Прикладное ПО, работающее в среде ОС Linux, имеет возможность получить данные об энергопотреблении несколькими способами [18], например:

- с использованием интерфейсов Running Average Power Limit (RAPL);
- прямым чтением специальных регистров микропроцессоров (MSR);
- с использованием библиотек Performance Application Programming Interface (PAPI).
- посредством обращения к Baseboard Management Controller (BMC) или Intelligent Platform Management Interface (IPMI).

Наиболее простым и универсальным вариантом является использование программных интерфейсов RAPL, который реализуется одной из подсистем ОС Linux. Он обеспечивает платформонезависимый доступ к показателям энергопотребления и предоставляет информацию о потребленной энергии, измеряемой в мДж. При этом измерение производится косвенно при помощи программной модели и основано на данных специальных регистров (MSR).

RAPL также используется и в PAPI [19] как один методов для получения данных о энергопотреблении.

В свою очередь IPMI и BMC являются стандартизованными механизмами для мониторинга системы, которые кроме энергопотребления измеряют также температуру, обороты вращения вентиляторов, напряжение и статус других компонент платформы.

3.2. Реализация оценки потребления энергии

Технически ОС Linux предоставляет доступ к RAPL через специальные файлы, расположенные в файловой системе /sys. На используемых в МСЦ РАН системах можно получить данные для четырех счетчиков энергопотребления: двух микропроцессоров и двух банков памяти. Для каждого из них используются специальные файлы energy_uj. Поскольку счетчики работают в накопительном режиме и имеют фиксированную разрядность, система предоставляет возможность получения максимального значения до переполнения (max_energy_range_uj), когда счетчик сбрасывается в ноль.

В прототипе Laptest добавлен новый модуль, который обеспечивает измерение потребляемой энергии во время выполнения тестов. В конфигурационном файле помимо параметров запуска тестов указывается список опрашиваемых счетчиков. В начале и конце работы каждого из тестов всеми выделенными для работы процессами осуществляется опрос датчиков энергопотребления и передача их «нулевому» процессу, где данные суммируются для всех датчиков, а затем вычисляется среднее значение потребленной мощности, которое попадает в отчет о работе теста. Для упрощения обработки данных и исключения необходимости модификации самих тестов, продолжительность их работы выбрана таким образом, чтобы исключить двойное переполнение счетчиков энергопотребления [20].

3.3. Оценка корректности измерений

Для оценки корректности измерения энергопотребления во время работы тестов одновременно произведено получение данных энергопотребления через IPMI для оперативной памяти и микропроцессоров:

ipmitool -c -b 0x06 -t 0x2c nm statistics power domain memory:

ipmitool -c -b 0x06 -t 0x2c nm statistics power domain cpu.

С помощью shell-скрипта данные IPMI получались каждые 2 секунды, а затем усреднялись для каждого из тестов.

3.4. Результаты измерений

Тестовые прогоны проводились на узлах, использующих два микропроцессора Intel Xeon E5-2697A (Broadwell) и 128 ГБ оперативной памяти, а также два микропроцессора Intel Xeon Gold 6154 (Skylake) и 192 ГБ оперативной памяти.

Были выбраны следующие тесты Laptest:

- S – Sleep – процессы «спят» заданное в конфигурации время;
- MR1 и MR2 – Memory Read – все ядра (MR1) или потоки (MR2) осуществляют одновременное чтение последовательных данных из оперативной памяти;
- MW1 и MW2 – Memory Write – все ядра (MW1) или потоки (MW2) осуществляют одновременную последовательную запись данных в оперативную память;
- BURN1 и BURN2 – Burn - все ядра (BURN1) или потоки (BURN2) одновременно выполняют операцию DGEMM над матрицами.

Полученные в результате тестовых прогонов данные представлены ниже (Таблицы 2 и 3).

Таблица 2. Результат запуска на узле Broadwell

Тест	Средняя потребляемая мощность (RAPL), Вт	Средняя потребляемая мощность (IPMI), Вт
S	39	38
MR1	259	258
MR2	282	281
MW1	250	249
MW2	265	263
BURN1	277	275
BURN2	326	322

Таблица 3. Результат запуска на узле Skylake

Тест	Средняя потребляемая мощность (RAPL), Вт	Средняя потребляемая мощность (IPMI), Вт
S	68	73
MR1	349	351
MR2	386	386
MW1	370	370
MW2	420	421
BURN1	371	373
BURN2	437	436

Таким образом, тестовые прогоны показали приемлемую точность измерений, проводимых с помощью данных, получаемых через интерфейсы RAPL, в сравнении с аналогичными, полученными через IPMI.

Публикация выполнена в рамках государственного задания по проведению фундаментальных исследований по теме FNEF-2024-0016.

Заключение

В программу комплексного тестирования производительности вычислительного кластера добавлена возможность оценки потребляемой мощности вычислительных узлов. Результаты исследования показали корректность и приемлемую точность измерений, проводимых с помощью реализованных средств.

Estimation of Power Consumption in a Comprehensive Computing Cluster Performance Testing Program

Oleg Vdovikin, Pavel Telegin, Boris Shabanov

Abstract. Different groups of tests are used to compare the performance of computing clusters. The Laptest comprehensive performance testing program was developed to determine the efficiency parameters of the computing cluster systems with a large number of cores. Due to the high energy consumption of supercomputers, energy efficiency of executing tasks has become one of the most important features of computing systems. The article describes the implementation of energy consumption estimation in the Laptest program.

Keywords: supercomputer, computing cluster, performance testing, energy efficiency, RAPL

Литература

1. Standard Performance Evaluation Corporation, 2024. URL: http://www.spec.org/. (Дата обращения: 18 07 2024).

2. NAS Parallel Benchmarks. URL: http://www.nas.nasa.gov/Resources/Software/npb.html. (Дата обращения: 18 06 2024).

3. A. Petitet, R. C. Whaley, J. Dongarra и A. Cleary. HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers, 2016. URL: http://www.netlib.org/benchmark/hpl/. (Дата обращения: 04 08 2018).

4. TOP500. The List, 1993-2024. URL: https://www.top500.org/. (Дата обращения: 17 08 2024).

5. Суперкомпьютеры top50, 2023. URL: http://top50.supercomputers.ru/?page=rating. (Дата обращения: 15 08 2024).

6. BLAS (Basic Linear Algebra Subprograms) 2024. URL: https://www.openblas.net/. (Дата обращения: 23 07 2024).

7. D. Papakyriakou, I. S. B. Barbounakis. High Performance Linpack (HPL) Benchmark on Raspberry Pi 4B (8GB) Beowulf Cluster. «International Journal of Computer Applications», Т. 185 (2023) №. 25, 11-19.

8. J. Dongarra, M. A. Heroux, P. Luszczek. HPCG Benchmark: a New Metric for Ranking High Performance Computing Systems. Technical Report, Electrical Engineering and Computer Sciente Department, UT-EECS-15-736, Knoxville, Tennessee, 2015.

9. А.В. Киселев, Е. А. Киселев и В. В. Корнеев. Анализ структуры информационного графа теста HPCG. «Научный сервис в сети Интернет: многообразие суперкомпьютерных миров. Труды Международной суперкомпьютерной конференции (22-27 сентября 2014 г., г. Новороссийск)» М.: Изд-во МГУ, 2014, 49-51.

10. Memory Test Suite 2020. URL: https://openbenchmarking.org/suite/pts/memory. (Дата обращения: 17 07 2024).

11. The HPC Challenge (HPCC) benchmark suite. ACM, Inc., 2024. URL: https://dl.acm.org/doi/10.1145/1188455.1188677. (Дата обращения: 18 07 2024).

12. rocHPL-MxP. 2024. URL: https://github.com/ROCm/rocHPL-MxP. (Дата обращения: 19 07 2024).

13. G. Henry, E. Petit, A. Lyashevsky, P. Caday. Deconstructing HPL-MxP Benchmark: A Numerical Perspective. «Euro-Par 2024: Parallel Processing». Berlin, Heidelberg, Springer-Verlag, 2024, pp. 47 - 60.

14. S. Devineni и G. Bhargavi. Energy-Efficient Computing and Green Computing Techniques. «Computer Science, Engineering and Technology» T.1 (2023), № 4, 37-45.

15. Б.М. Шабанов, Исследование, разработка и применение суперкомпьютерных вычислительных систем. Докторская диссертация. Москва, 2019, http://www.frccsc.ru/sites/default/files/docs/ds/002-073-02/diss/08-shabanov/ds02-08-shabanov_main.pdf?336. (Дата обращения: 25 12 2019).

16. М.С. Клинов, С. Ю. Лапшина, П. Н. Телегин, Б. М. Шабанов. Особенности использования многоядерных процессоров в научных вычислениях. «Вестник УГАТУ». Т. 1 (2012), № 6(51), 25-31.

17. Б.М. Шабанов, П. Н. Телегин, О. С. Аладышев. Особенности использования многоядерных процессоров. «Программные продукты и системы». (2008) № 2, 7-9, 2008.

18. F. Alizadeh Moghaddam, T. Geenen, P. Lago, P. Grosso. A user perspective on energy profiling tools in large scale computing environments «2015 Sustainable Internet and ICT for Sustainability (SustainIT)». Madrid, Spain, 2015, 1-5.

19. V.M. Weaver, M. Johnson, K. Kasichayanu, J. Ralph, P. Luszczek, D. Terpstra, S. Moore. Measuring Energy and Power with PAPI. «2012 41st International Conference on Parallel Processing Workshops». Pittsburgh, PA, USA 2012, 262-268.

20. K.N. Khan, M. Hirki, T. Niemi, J. K. Nurminen, Z. Ou. RAPL in Action: Experiences in Using RAPL for Power Measurements. «ACM Transactions on Modeling and Performance Evaluation of Computing Systems». T.3 (2018), №2, Article 9, 1-26.