УДК 004.021

ИССЛЕДОВАНИЕ ДИНАМИЧЕСКИХ ХАРАКТЕРИСТИК ПОТОКА ЗАДАЧ СУПЕРКОМПЬЮТЕРНОЙ СИСТЕМЫ

А. С. Антонов[1], С. А. Жуматий[1], Д. А. Никитенко[1], К. С. Стефанов[1], А. М. Теплов[1], П. А. Швец[1]

Представлена система мониторинга динамических характеристик потока задач суперкомпьютерной системы, реализованная в данный момент на суперкомпьютере СКИФ МГУ "Чебышев". Предложенный подход к анализу позволяет эффективно и технологически просто получить качественную оценку свойств реального потока задач, на основе которого можно судить об утилизации ресурсов суперкомпьютера, выделить проблемные места архитектуры и наметить возможные направления ее оптимизации.

Ключевые слова: суперкомпьютер, производительность, эффективность, динамические характеристики, загрузка, мониторинг, система управления потоком задач.

1. Введение. Эффективное использование суперкомпьютерной системы является крайне актуальной задачей, с необходимостью ее решения сталкиваются все современные держатели высокопроизводительных комплексов. Для того чтобы правильно оценить, что и каким образом следует улучшить в эксплуатации системы, для начала необходимо иметь корректную картину того, что реально происходит внутри суперкомпьютера.

Результаты исследования могут показать наличие серьезных ограничений аппаратной составляющей, в которые упираются пользовательские приложения. В таком случае руководство может принимать соответствующие решения по обновлению системы или же определять подобающим образом техническое задание и требования к новой системе, призванной прийти на смену ныне используемой.

Возможности современных средств сбора данных системного мониторинга (HOPSA/LAPTA, Ganglia, Collectd, ClustrXWatch и др.) достаточно широки. Здесь широту следует понимать как в смысле большого количества доступных для оценки метрик и достаточно высокой частоты их сбора при низких накладных расходах, так и в смысле существования описанных интерфейсов для доступа к собранным данным.

Для исследования динамических характеристик суперкомпьютерных приложений предлагается использовать подход, называемый Job Digest, созданный ранее коллективом разработчиков настоящего проекта [1, 2]. Этот подход применяется в Суперкомпьютерном комплексе МГУ [3]. Основная идея подхода заключается в получении и сохранении данных от датчиков систем мониторинга с момента начала работы приложения и до ее завершения. Реализующая Job Digest подсистема в результате взаимодействия с системой очередей суперкомпьютера определяет временной диапазон и список задействованных для приложения вычислительных узлов. Данные мониторинга сохраняются в СУБД (сейчас используются MongoDB [4] и Cassandra [5]. Подсистема Job Digest поддерживает интеграцию с системами очередей Cleo [6] и Slurm [7], установленными соответственно на суперкомпьютерах СКИФ МГУ "Чебышев" и "Ломоносов".

2. Ключевые динамические характеристики суперкомпьютерных приложений. Для исследования динамических характеристик потока задач суперкомпьютерной системы следует выделить набор характеристик, представляющий наибольший интерес. Таких характеристик заведомо не должно быть много, чтобы общая картина была более ясной. К тому же, зачастую на практике число снимаемых с процессора датчиков сильно ограничено, и приходится из них выбирать самые важные. С другой стороны, анализ на основе выбранных характеристик должен давать по возможности всестороннюю картину исследуемого потока задач.

На основании накопленного опыта по исследованию динамических характеристик суперкомпьютерных приложений с учетом данных системного мониторинга можно сделать выводы о необходимости при-

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Ленинские горы, д. 1, стр. 4, 119992, Москва; А. С. Антонов, вед. науч. сотр., e-mail: asa@parallel.ru; С. А. Жуматий, вед. науч. сотр., e-mail: serg@parallel.ru; Д. А. Никитенко, науч. сотр., e-mail: dan@parallel.ru; К. С. Стефанов, ст. науч. сотр., e-mail: cstef@parallel.ru; А. М. Теплов, мл. науч. сотр., e-mail: alex-teplov@yandex.ru; П. А. Швец, программист, e-mail: shvets.pavel.srcc@gmail.com

© Научно-исследовательский вычислительный центр МГУ им. М. В. Ломоносова

влечения к исследованиям данных, характеризующих следующие составляющие хода выполнения суперкомпьютерного приложения:

1) пользовательская загрузка процессора;

2) число выполненных арифметических операций;

3) число процессов, ожидающих вхождения в стадию счета;

4) данные о загрузке коммуникационной среды;

5) данные об использовании ввода-вывода (операции с диском или активность соответствующей транспортной сети, реализующей работу с сетевой файловой системой).

2.1. Пользовательская загрузка процессора. Из набора традиционных индикаторов в операционной системе UNIX, отражающих характер использования центрального процессора, наибольший интерес представляет доля пользовательских процессов в общей загрузке процессора — CPU user, в наибольшей мере отражающая загрузку процессора приложением. Для более подробного исследования поведения приложения можно использовать и другие датчики, однако при общих исследованиях, в том числе при определении типовых профилей использования суперкомпьютерных систем, предполагающих всестороннюю оценку как приоритетную, достаточно ограничиться включением CPU User в список ключевых характеристик.

2.2. Число арифметических операций. Абсолютное большинство современных вычислительных алгоритмов основываются либо на целочисленной арифметике, либо на вещественной арифметике с определенной точностью. В связи с этим, основные показатели числа совершенных операций традиционно измеряются в Iops (Integer operations per second) и Flops (Floating point operations per second), т.е. число совершенных целочисленных операций и операций с плавающей точкой соответственно. Целочисленная арифметика свойственна таким задачам, как, например, задача перебора или обхода графа, в то же время бóльшая часть научных расчетов основывается на операциях с плавающей точкой. Такие задачи занимают львиную долю загрузки вычислительных мощностей суперкомпьютерных комплексов, предоставляющих ресурсы для научных расчетов. Следовательно, в список ключевых характеристик следует включить именно счетчик операций с плавающей точкой, а индикатор использования целочисленной арифметики использовать при исследовании специального класса задач.

2.3. Число процессов, ожидающих вхождения в стадию счета. Особый интерес всегда представляет степень используемого параллелизма на вычислительном узле. В идеальной ситуации на смену текущему процессу всегда должен быть готов прийти другой процесс, причем только один. Если таких процессов больше, то происходит простой процессов в ожидании ресурса, а если меньше — простой ресурса в ожидании процесса.

Таким образом, если некоторая задача выполняется, например, на узлах вычислительного кластера, каждый из которых включает в себя K N-ядерных процессоров, то идеальный показатель загрузки (Load Average) стремится к $K * N$. Например, для узлов из двух четырехъядерных процессоров показатель Load Average должен быть близок к $2 * 4 = 8$. Если значение Load Average значительно превышает такое значение, то это не обязательно говорит о проблемах в эффективном построении приложения, но в некоторых случаях, например, о большом числе порожденных процессов при использовании технологии OpenMP.

2.4. Данные о загрузке коммуникационной среды. Коммуникационная среда — важнейшая составляющая современной вычислительной системы. Именно благодаря ей отдельные процессы обмениваются сообщениями, необходимыми для вычислений. В последнее время чаще всего в качестве коммуникационной сети вычислительных кластеров используется Infiniband или Gigabit Ethernet. Существует множество характеристик для сетевой инфраструктуры: задержки передачи, число ошибок, средняя скорость, объем переданных данных и др. Тонкое исследование коммуникационных профилей конкретных приложений — отдельная серьезная задача, при которой и потребуется все множество доступных метрик. Для рассмотрения же поведения коммуникационной среды в целом в наибольшей мере подходит скорость передачи данных, отражающая как интенсивность передачи данных, так и степень загруженности пропускной способности канала.

2.5. Данные об использовании ввода-вывода. Ввод-вывод в суперкомпьютерных системах — одна из наиболее долгих, а потому “дорогих” в вычислительном смысле задач. В случае использования на узлах локальных дисков доступен целый ряд датчиков, отражающих как состояние дисков, так и объем переданных данных, промахи при обращении в кэш дисков и др. Однако, как следствие высокой стоимости, бóльшая часть вычислительных узлов современных систем не оснащается локальными дисками. Вместо этого используется сетевая файловая система, а данные передаются по специальной транспортной сети в целях минимизации наведенных эффектов на межузловую передачу данных в ходе счета. В условиях текущего роста масштаба вычислительных систем и роста степени параллелизма такой подход

будет применяться все шире, а потому именно его следует рассматривать как основной. В таком случае при использовании специальной сети для ввода-вывода справедливо выделение такой же ключевой характеристики, как и для коммуникационной сети, — скорости передачи данных.

3. Сбор данных для анализа. Источником данных для анализа могут служить стандартные средства получения данных системного мониторинга, такие как Ganglia, HOPSA и др., позволяющие сохранить их в формате CSV (Comma-Separated Values, последовательность значений с разделителями). Входные данные характеризуются составом, частотой съема и временем наблюдения.

Рекомендуемая временна́я гранулярность данных — 10 минут. Более частый сбор данных избыточен и приводит к увеличению времени получения собранных данных, при этом качественно представление профиля не становится существенно более точным. Это связано с тем, что наибольшую роль в формировании типового профиля использования системы играют большие задачи. При этом один только процесс их старта может занимать несколько минут в рамках ранее указанного рекомендованного шага по времени.

Рекомендуется исследовать временны́е интервалы длительностью не меньше среднего времени счета для конкретной вычислительной системы. Наиболее интересными для анализа являются следующие временны́е интервалы.

Сутки. Здесь следует учесть возможную существенную разницу между днями недели. Это связано с тем, что многие исследователи, подготовив задачу, ставят ее на счет перед выходными или вечером, чтобы получить результат к моменту начала следующего рабочего дня или недели.

Неделя. Наиболее интересный для исследования вариант. Сложность состоит в непрерывности наблюдений. На практике вычислительные системы могут приостанавливаться для профилактики, обновлений и т.п., что с некоторой вероятностью может попасть даже на недельный интервал.

Для оценки выделенных ключевых характеристик 1 и 5 мы будем исследовать абсолютные значения соответствующих датчиков, а для характеристик 2–4 — скорости изменения датчиков за секунду, так как значениями этих датчиков, получаемыми от системного мониторинга, являются суммарные объемы данных с какого-то момента в прошлом (чаще всего с момента загрузки узла-источника этих данных), что для наших целей не несет ценной информации.

Данные мониторинга представляются в виде последовательностей пар (T_i, V_i), где T_i — момент времени, которому соответствуют данные мониторинга, а V_i — значение в этот момент времени. Среднее значение за промежуток времени (T_0, T_N) для характеристик 1 и 5 вычисляется по формуле

$$\sum_{i=1}^{N} V_{i-1}(T_i - T_{i-1}),$$

а для характеристик 2–4 — по формуле:

$$\sum_{i=1}^{N} V_i(T_i - T_{i-1}).$$

Разница в методах вычисления объясняется особенностями работы механизма прореживания данных используемой системы мониторинга. Указанным способом проводится усреднение отдельно для каждого узла (для датчиков 3–5) или для каждого процессорного ядра (для датчиков 1–2). Затем усредненные данные для каждого узла (или для каждого ядра процессора) усредняются по всем узлам (ядрам), от которых в течение указанного периода было получено хотя бы одно значение.

Полученные таким образом значения мы считаем средним значением указанного датчика для данного временно́го периода для исследуемого суперкомпьютера. Затем разобьем весь период наблюдений на равные периоды времени и для каждого периода за весь период наблюдений получим среднее значение. Полученная последовательность средних и будет составлять статистические данные по загрузке исследуемого суперкомпьютерного комплекса за период наблюдения.

4. Полученные результаты. Динамические характеристики потока задач исследовались на примере суперкомпьютера СКИФ МГУ “Чебышев” [8] с пиковой производительностью 60 Tflops. Эта вычислительная система представляет собой классический кластер из более чем 600 узлов, каждый из которых оснащен двумя четырехъядерными процессорами; итого около 5000 вычислительных ядер. Некоторые узлы оснащены локальными дисками и увеличенным объемом памяти, в качестве коммуникационной сети используется наиболее распространенная ее реализация — Infiniband, а в качестве транспортной сети, по которой осуществляется доступ к сетевой файловой системе, — Gigabit Ethernet. Рассматриваемая вычислительная система входит в состав Суперкомпьютерного комплекса МГУ им. М. В. Ломоносова и активно используется научным сообществом для решения актуальных вычислительно емких задач.

Данные собирались за недельный промежуток с 21 октября 2013 г. по 27 октября 2013 г. с усреднением по интервалу 10 минут. По каждой из выделенных ключевых характеристик по описанной методике были получены графики, показывающие динамику их изменения в указанный период. Если суммировать данные за весь период наблюдений, получим результаты, приведенные в таблице.

Пользовательская загрузка процессора, %	70.16
Число операций с плавающей точкой, Mflops	420.5
Число процессов, ожидающих вхождения в стадию счета (Load Average)	3.77
Скорость приема данных по коммуникационной сети, Мбайт/с	63.5
Скорость приема данных по транспортной сети, Мбайт/с	165.5
Скорость передачи данных по транспортной сети, Мбайт/с	336.3

Показатель средней пользовательской загрузки процессора более 70% говорит о хорошем уровне использования предоставленных ресурсов, по крайней мере с точки зрения их простоя. Отметим, что при анализе данных учитывались все узлы суперкомпьютера, в том числе и те, которые по каким-то причинам выведены из счетного поля, а также те, на которые не распределяются задачи для обеспечения запуска какой-то стоящей в очереди задачи, требующей значительных ресурсов.

Среднее количество операций с плавающей точкой 420.5 Mflops является крайне невысоким, учитывая, что теоретический максимум для используемых в “Чебышеве” процессоров составляет 12 Mflops, т. е. мы достигли всего лишь 3.5% от максимального значения. Однако именно такова ситуация в среднем в области высокопроизводительных вычислений: ресурсы процессора используются большинством программ в весьма незначительной степени.

Данные о числе процессов, ожидающих вхождения в стадию счета (Load Average), также свидетельствуют о неплохой утилизации предоставленных ресурсов. При этом подавляющее большинство задач не запускают число процессов (нитей) больше, чем имеющееся число ядер процессора (в нашем случае 8). Анализ узлов, на которых были зафиксированы значения Load Average в диапазоне 9–11, показал, что чаще всего такие значения вызваны работой служебных процессов, обеспечивающих коммуникации, а также используемой системой мониторинга. Значительный объем узлов со значением параметра Load Average, равным нулю, может объясняться наличием в ядре Linux ошибки, которая приводит к неучитыванию работы процессов, часто переключающихся из режима счета в режим ожидания (см., например, [9]), а такие нередко встречаются при работе вычислительных задач с активными коммуникациями.

Мы говорим только о средней скорости приема данных по коммуникационной сети и не говорим о данных, переданных по той же сети. Это связано с тем обстоятельством, что эти величины с хорошей точностью (не менее 99.5%) совпадают на всем интервале наблюдения. Такое совпадение объясняется тем, что при нормальной работе системы данные, переданные одним узлом, принимаются каким-то другим узлом. Так как мы ведем наблюдение сразу за всеми вычислительными узлами суперкомпьютера, а по коммуникационной сети связаны между собой только они, то полученное совпадение говорит о правильности работы вычислительного комплекса. Имеющееся незначительное расхождение объясняется учетом передачи и приема некоторых данных в разных интервалах наблюдений.

Для транспортной сети совпадения объема полученных и переданных данных не наблюдается. Это объясняется тем, что основной объем данных, передаваемых по этой сети, приходится на коммуникацию вычислительных узлов с серверами сетевой файловой системы, данные мониторинга которой не учитываются в нашем исследовании.

Более чем двукратное превышение средней скорости передаваемых данных по сравнению с принятыми свидетельствует о том, что выполняемые на суперкомпьютере задачи обычно имеют выходной объем данных больше, чем поступает им на вход.

5. Заключение. Рассмотрена возможность мониторинга динамических характеристик потока задач суперкомпьютерной системы, реализованная в настоящее время на суперкомпьютере СКИФ МГУ “Чебышев”. Предложенный подход к анализу позволяет эффективно и технологически просто получить качественную оценку свойств реального потока задач. Исследуя типовые значения ключевых характеристик, можно сделать вывод о характере и масштабе использования ресурсов (процессор, коммуникационная сеть и др.) на текущем потоке задач. Результаты такого анализа позволяют выделить нормальный (т.е. средний) уровень использования и высокий уровень загрузки, вплоть до полного использования возможностей ресурса (например, полная загрузка пропускной способности канала). Анализ получаемых значений позволяет судить об утилизации ресурсов суперкомпьютера теми задачами, которые считаются на нем в рассматриваемый период времени, выделить проблемные места архитектуры и наметить возможные

направления ее оптимизации.

Работа выполняется при финансовой поддержке Министерства образования и науки Российской Федерации, государственный контракт № 14.514.11.4107, и гранта РФФИ № 13-07-00786.

СПИСОК ЛИТЕРАТУРЫ

1. *Адинец А.В., Брызгалов П.А., Воеводин Вад.В., Жуматий С.А., Никитенко Д.А., Стефанов К.С.* Job Digest — подход к исследованию динамических свойств задач на суперкомпьютерных системах // Вестн. Уфимского гос. авиационного техн. ун-та. 2013. **17**, № 2. 131–137.
2. *Adinets A.V., Bryzgalov P.A., Voevodin Vad.V., Zhumatii S.A., Nikitenko D.A., Stefanov K.S.* Job Digest: an approach to dynamic analysis of job characteristics on supercomputers // Numerical Methods and Programming: Advanced Computing. 2012. **13**, section 2. 160–166.
3. *Воеводин Вл.В., Жуматий С.А., Соболев С.И., Антонов А.С., Брызгалов П.А., Никитенко Д.А., Стефанов К.С., Воеводин Вад.В.* Практика суперкомпьютера "Ломоносов" // Открытые системы. 2012. № 7. 36–39.
4. MongoDB (http://www.mongodb.org/).
5. Cassandra (http://cassandra.apache.org/).
6. Cleo Cluster Batch System (http://sourceforge.net/projects/cleo-bs/).
7. SLURM: A Highly Scalable Resource Manager (https://computing.llnl.gov/linux/slurm/).
8. *Антонов А.С.* СКИФ МГУ — основа Суперкомпьютерного комплекса Московского университета // Вторая Международная научная конференция "Суперкомпьютерные системы и их применение" (SSA'2008). Минск: ОИПИ НАН Беларуси, 2008. 7–10.
9. Linux load averages, for example from top and uptime commands, can be massively incorrect on the low side (http://www.smythies.com/ doug/network/load_average/original.html).

Поступила в редакцию
30.10.2013