ПРЕДПОСЫЛКИ АВТОМАТИЗИРОВАННОГО ПРОЕКТИРОВАНИЯ КЛАСТЕРНЫХ ЭВМ

К.С. Солнушкин

Производительность кластерных ЭВМ в России за последние 4 года растет экспоненциальными темпами. Свидетельством тому статистика списка “TOP 50” суперкомпьютеров России [10]. Однако конфигурация вычислительных систем – кластерных ЭВМ и инфраструктуры – порой выбирается спонтанно, без теоретического обоснования. Причиной этого во многом является отсутствие строгой, подтвержденной практикой теоретической базы принципов конструирования кластерных вычислительных систем.

Параллельные вычисления, в том числе с применением кластерных технологий, насчитывают уже не один десяток лет [5], однако кластерные ЭВМ, основанные на массово выпускаемых комплектующих, получили широкое распространение, по-видимому, начиная с проекта “Beowulf” (1993-1994 г.) [4]. Проектирование кластерных ЭВМ на научной основе с оценкой их стоимости стало предметом ряда исследований, однако зачастую служило решению конкретных вычислительных задач [9], либо рассматривался узкий диапазон аппаратных средств [6].

Как самостоятельный вид инженерной деятельности научно обоснованное проектирование кластерных ЭВМ из массовых компонентов получило наибольшее развитие в начале XXI века. Среди важных работ, развивающих применение методов автоматизированного проектирования и поддержки принятия решений, можно назвать [11; 1]. Математическая постановка задачи синтеза кластерных ЭВМ, оптимальных по критерию цены производительности, приведена в [2].

В данной статье обосновывается необходимость применения средств автоматизированного проектирования кластерных ЭВМ; рассматриваются две существующие программные системы и сравниваются подходы к проектированию.

НЕОБХОДИМОСТЬ АВТОМАТИЗИРОВАННОГО ПРОЕКТИРОВАНИЯ КЛАСТЕРНЫХ ЭВМ

Кластерные ЭВМ строят на основе массово выпускаемых компонентов. Это обусловливает широкое разнообразие компонентов в реально создаваемых ЭВМ. Так, по состоянию на май 2009 г. компания “Hewlett-Packard” выпускает кластерные ЭВМ на основе широкой номенклатуры центральных процессоров (ЦП) Intel Xeon и Itanium и AMD Opteron. Если ограничиться лишь современными сериями процессоров – Intel Xeon серий 5400, 7400 и 5500 (с 4 и 6 ядрами, с различной тактовой частотой, тепловыделением и т.д. – более 20 вариантов ЦП), Intel Itanium (с 2 ядрами – 2 варианта) и AMD Opteron серии 2300 (с 4 ядрами – более 10 вариантов), то общее количество типов процессоров составит свыше 30. Внутри одной серии процессоры отличаются тактовой частотой, размером кэш-памяти, тепловыделением и, разумеется, ценой; выбор модели процессора является нетривиальной задачей. Как справедливо отмечают авторы [11], лучший процессор для кластерной ЭВМ – не самый быстрый и не самый дешевый сам по себе, а тот, который доставляет экстремум интегральному критерию качества ЭВМ.

Для конструктивного исполнения кластерной ЭВМ компания “Hewlett-Packard” предлагает несколько вариантов: общепринятый стандарт исполнения сервера высотой 1 единицу (1U, от англ. "unit"), равную 1,75" (иногда применяют серверы высотой 2 единицы) и два варианта конструктивных блоков («шасси») для установки Blade-серверов – высотой 6U и 10U [3] (итого 4 варианта исполнения). Узлы и конструктивные блоки размещают в стойках, наиболее популярны стойки высотой 42U, применяют также и стойки высотой 22U.

Для построения коммуникационных сетей наиболее часто применяют технологии Gigabit Ethernet, Infiniband, Myrinet (итого 3 варианта); при этом сети, объединяющие узлы, могут иметь несколько вариантов топологий (звезда, дерево и т.д.)

Таким образом, даже если не учитывать другие параметры вычислительной системы, общее количество вариантов построения только узлов кластерной ЭВМ составит по самым консервативным подсчетам порядка нескольких сотен, и это только для одного производителя ЭВМ.

Каждый вариант ЭВМ характеризуется стоимостью и производительностью. Выбор из множества вариантов необходимо делать с учетом ограничений – в первую очередь, на стоимость вычислительной системы, ее массогабаритные характеристики, энергопотребление и т.д. Очевидно, что при таком количестве комбинаций оценить перспективность каждого варианта сложно даже опытному проектировщику.

Вычислительная система, кроме собственно кластерной ЭВМ, включает в себя инфраструктурные компоненты: системы электропитания, охлаждения, хранения данных. В случае, если на месте установки ЭВМ эти компоненты отсутствуют, их необходимо проектировать совместно с ЭВМ, при этом должны выполняться все имеющиеся ограничения. Для каждого проекта вычислительной системы необходимо иметь обоснованный прогноз совокупной стоимости владения (“Total Cost of Ownership”, “TCO”) [3].

Из сказанного следует, что труд инженера-проектировщика необходимо облегчить, снабдив его системой автоматизированного проектирования кластерных ЭВМ, комбинированной с системой поддержки

принятия решений. Это позволит рассматривать больше вариантов компоновки ЭВМ, чем это возможно вручную [11], что будет способствовать выбору оптимального варианта, уменьшит вероятность человеческой ошибки, переложит монотонную работу на плечи компьютера, освободив время проектировщика для выработки новых, творческих подходов к конструированию.

СУЩЕСТВУЮЩИЕ ПРОГРАММНЫЕ СИСТЕМЫ

Существует целый ряд систем автоматизированного проектирования кластерных ЭВМ. Часть их разработана производителями аппаратного обеспечения. Так, компания “Hewlett-Packard” предлагает ряд разрозненных утилит, некоторые из них доступны через Web-интерфейс, а другие устанавливаются на рабочее место проектировщика (например, средство “HP BladeSystem Sizer”); компания “IBM” предлагает средство “Standalone Solutions Configuration Tool” (“IBM SSCT”) [3]. Подобные системы проектирования – «конфигураторы» – делают упор на подборе гарантированно совместимых компонент и оценивают в первую очередь технологические характеристики проектируемых ЭВМ – массу, габариты, энергопотребление, а также стоимость.

Что же касается производительности кластерных ЭВМ различной конфигурации, то необходимо отметить, что значительные усилия мирового научного сообщества направлены на прогнозирование производительности ЭВМ с помощью тех или иных математических моделей; и вместе с тем, до сих пор модели производительности и стоимости редко применялись совместно для проектирования ЭВМ. Одной из немногих успешных попыток следует назвать, по-видимому, систему проектирования “Cluster Design Rules” [7], представленную на конференции “Supercomputing–2002” (цит. по [11]). Это комбинированная система автоматизированного проектирования и система поддержки принятия решений с Web-интерфейсом. Научные основы построения системы были изложены в 2005 г. в [11].

В основе системы лежит предположение о том, что для проектирования кластерной ЭВМ необязательно быть экспертом в области аппаратного обеспечения, однако необходимо знать потребности в вычислительных ресурсах того программного обеспечения, которое будет исполняться на проектируемой ЭВМ. Система, таким образом, предназначена для специалистов в предметной области будущего использования ЭВМ (физиков, химиков и т.д.) Проектировщику задается ряд вопросов, на основе которых система автоматически синтезирует оптимальную ЭВМ.

Первая группа вопросов относится к параметрам аппаратного обеспечения, например, сколько оперативной памяти необходимо иметь в будущем кластере, какова минимально допустимая удельная пропускная способность оперативной памяти вычислительного узла в гигабайтах в секунду в расчете на 1 GFLOPS пиковой производительности ЦП узла, какова максимально допустимая латентность коммуникационной сети и т.д. Ответы на эти вопросы следует давать исходя из потребностей приложения.

Ситуация, при которой на ЭВМ будет выполняться несколько приложений, не предусмотрена. Очевидно, в этом случае ответ на каждый вопрос следует давать таким образом, чтобы обеспечить потребности любого из приложений.

Вторая группа вопросов посвящена конструктивному исполнению и параметрам инфраструктуры, в частности, максимально допустимое количество стоек с оборудованием, максимально допустимая потребляемая мощность ЭВМ, а также холодопроизводительность имеющейся системы охлаждения.

Третья группа вопросов описывает экономические параметры проектируемой ЭВМ. Имеется возможность указать максимально допустимые величины стоимости закупки и эксплуатационных затрат в расчете на один год. В стоимость закупки входит стоимость лицензий на программное обеспечение, для которой проектировщик может отдельно указать фиксированную часть и переменную часть в расчете на один вычислительный узел, ЦП или ядро ЦП. Эксплуатационные расходы система проектирования подсчитывает как стоимость потребляемой электроэнергии и стоимость аренды площади в центре обработки данных. Количество электроэнергии, которое потребляет ЭВМ, система подсчитывает автоматически для каждой синтезируемой ею конфигурации ЭВМ. Количество электроэнергии, потребляемое системой охлаждения, также подсчитывается автоматически, исходя из сделанного проектировщиком выбора одного из пяти предложенных типов систем охлаждения.

Четвертая группа вопросов предназначена для задания ограничений, связанных с производительностью. Здесь можно указать минимально допустимые значения пиковой производительности и реальной производительности на тестах “High Performance LINPACK” (HPL) и “SWEEP3D” [8].

Далее предлагают выбрать критерий эффективности ЭВМ. Им может служить реальная производительность на тестах “HPL” или “SWEEP3D”, либо совокупная стоимость владения, либо, как наиболее гибкий вариант, взвешенная сумма частных критериев эффективности ЭВМ. Путем назначения весов по усмотрению проектировщика можно реализовать сколь угодно гибкий критерий эффективности.

Для составления взвешенной суммы используют следующие характеристики, увеличение которых ведет к росту эффективности ЭВМ: объем оперативной памяти ЭВМ; удельная пропускная способность оперативной памяти; объем жестких дисков узлов; пропускная способность коммуникационной сети (bisectional bandwidth); пиковая производительность. Используются также характеристики, уменьшение которых ведет к

росту эффективности: латентность коммуникационной сети; совокупная стоимость владения; эксплуатационные затраты; энергопотребление; холодопроизводительность системы охлаждения.

Как видно из приведенного описания, система проектирования намеренно изолирует проектировщика от выбора вручную конкретных моделей элементов аппаратного обеспечения ЭВМ, производя синтез автоматически на основе предъявленных формальных требований (даже выбор количества вычислительных узлов производится автоматически). Единственным способом как-либо повлиять на выбор конкретного типа компонентов является использование фильтров, разрешающих или запрещающих применение компонента в случае присутствия в его названии определенной подстроки; эти фильтры устроены чрезвычайно негибко.

Синтез конфигураций кластерной ЭВМ производится методом поиска в глубину с отсечением ветвей дерева решений [11]. Для обеспечения интерактивности время поиска ограничено 16 секундами, что накладывает жесткие ограничения на эффективность алгоритма поиска. Отсечение производится на основании несоответствия генерируемых решений формальным требованиям, предъявленным проектировщиком: минимальной пиковой производительности, максимальной стоимости закупки и т.д. Для каждой конфигурации, соответствующей требованиям, вычисляется значение интегрального критерия эффективности, и проектировщику выдают первые 32 варианта конфигурации с наибольшим значением критерия.

По умолчанию единственным слагаемым взвешенной суммы – интегрального критерия эффективности ЭВМ – является пиковая производительность; веса при остальных характеристиках приняты нулевыми. Это соответствует задаче синтеза ЭВМ максимальной производительности по методу главного критерия.

Ограничением взвешенной суммы является невозможность реализовать критерий «цена [единицы] производительности» [2]. В этом заключено различие в философском подходе: с использованием взвешенной суммы можно построить более гибкий критерий, однако возникает проблема субъективизма в назначении весов.

Система проектирования подбирает компоненты проектируемой ЭВМ на уровне отдельных устройств (материнские платы, ЦП, модули памяти и т.д.) В систему заложена информация о совместимости устройств различных типов по формальным признакам, например, по типу разъема ЦП и материнской платы. Важным достоинством системы является автоматический анализ множества вариантов построения коммуникационной сети, с различным оборудованием и разнообразными топологиями; для облегчения восприятия информации предусмотрено формирование графической условной схемы подключения узлов, являющейся подспорьем при сборке ЭВМ.

Отечественной аналогом рассмотренной системы является описанная в [1] система “ClusterDesign” (в сети Интернет отсутствует информация об этой системе). Авторы утверждают, что разработанная на языке Java система проектирования позволяет проводить проектирование на трех уровнях детализации: поэлементном (выбор проектировщиком таких устройств, как материнские платы, ЦП и т.д.), покомпонентном (выбор из массива готовых, протестированных на совместимость компонентов (вычислительных узлов и др.) и шаблонном (выбор из массива готовых вычислительных систем). Для оценки реальной производительности проектируемой ЭВМ применяется аналитическая модель производительности теста “High Performance LINPACK”. Сообщается также, хотя и без подробностей, что имеется возможность спроектировать совместно с ЭВМ системы электропитания и охлаждения.

Отметим разницу в подходе между двумя рассматриваемыми системами. В первой системе от проектировщика не требуется быть экспертом по аппаратному обеспечению, а во второй это необходимо; это является недостатком второй системы (знание особенностей прикладных программ, которые будут выполняться на проектируемой ЭВМ, необходимо в обоих случаях).

К числу недостатков второй системы относится и то, что ни стоимость закупки, ни совокупная стоимость владения в ней, по-видимому, никак не оцениваются; это не позволяет выбрать при прочих равных условиях более выгодную с экономической точки зрения ЭВМ.

С другой стороны, достоинством второй системы является возможность совместного проектирования систем электропитания и охлаждения ЭВМ. Кроме того, декларируется возможность конструирования ЭВМ из набора протестированных на совместимость компонентов, тогда как в первой системе совместимость обеспечивается лишь по формальным признакам, т.е. возможно соединение на физическом уровне, но гарантии совместной работы нет.

ВЫВОДЫ

При проектировании кластерных ЭВМ необходимо рассмотреть большое количество вариантов конфигураций; при этом системы автоматизированного проектирования могут оказать неоценимую помощь. Существующие системы реализуют лишь малую долю возможных функций, но уже представляют собой полезный инструмент. Рассмотрена работа двух таких систем, проанализировано сходство и различие в подходах.

Среди дальнейших направлений работ можно назвать сравнительный анализ критериев эффективности ЭВМ для применения в системах проектирования, а также выработку требований к будущим системам проектирования.

Работа поддержана грантом Правительства Санкт-Петербурга для молодых ученых за 2008 г.

ЛИТЕРАТУРА

1. Аветисян А.И., Самоваров А.И., Михайлов Г.М., Рогов Ю.П. Среда проектирования и разработки кластерных вычислительных систем // В сб. трудов Всероссийской научной конференции "Научный сервис в сети Интернет: технологии параллельного программирования" (г. Новороссийск, 18-23 сентября 2006 года). – М.: Изд-во МГУ, 2006. С. 62-64
2. Солнушкин К.С. Постановка задачи синтеза кластерных ЭВМ, оптимальных по критерию цены производительности // Труды XV Всероссийской научно-методической конференции Телематика-2008. СПб.: Изд-во СПбГУ ИТМО, 2008.
3. Солнушкин К.С. Моделирование совокупной стоимости владения вычислительной системой // Научно-технические ведомости СПбГПУ. – СПб: Изд-во СПбГПУ, 2008. – N6(69). – С. 130-135.
4. Beowulf.org: The Beowulf Cluster Site. – Режим доступа: http://www.beowulf.org, свободный.
5. Cluster (Computing) – Wikipedia. – Режим доступа: http://en.wikipedia.org/wiki/Cluster_computing, свободный.
6. Jeffrey C. Becker, Bill Nitzberg, Rob F. Van Der Wijngaart and Maurice Yarrow. Predicting Price/Performance Trade-offs for Whitney: A Commodity Computing Cluster. The Journal of Supercomputing, 13(3), pp. 303–319, May 1999.
7. The Aggregate Cluster Design Rules. – Режим доступа: http://aggregate.org/CDR, свободный
8. The ASCI SWEEP3D README File. http://www.c3.lanl.gov/pal/software/sweep3d/ sweep3d_readme.html
9. Thomas Hauser, Timothy I. Mattox, Raymond P. LeBeau, Henry G. Dietz, P. George Huang, "High-Cost CFD on a Low-Cost Cluster," pp.55, ACM/IEEE SC 2000 Conference (Supercomputing 2000), 2000.
10. TOP 50. Суперкомпьютеры. – Режим доступа: http://www.supercomputers.ru/?page=stat, свободный.
11. William R. Dieter and Henry G. Dietz, Automatic Exploration and Characterization of the Cluster Design Space, Technical Report TR-ECE-2005-04-25-01, University of Kentucky, April, 2005. – Режим доступа: http://www.engr.uky.edu/~dieter/pub/TR-ECE-2005-04-25-01.pdf, свободный