Достижение экзафлопсной производительности в задачах глобальной оптимизации*

В.П. Гергель, К.А. Баркалов

Нижегородский государственный университет им. Н.И. Лобачевского

Эффективное использование огромных вычислительных возможностей экзафлопсных систем представляет собой глобальную проблему «вызова» всему спектру вычислительных наук. Условиями для успешного достижения экзафлопсной производительности являются: значительная вычислительная трудоемкость решаемых задач, высокий запас параллелизма вычислений, низкая интенсивность информационная зависимость вычислений, устойчивость вычислений к аппаратным сбоям вычислителей. В работе показывается, что все перечисленные условия достижимы для задач глобальной оптимизации.

1. Проблема организации эффективных экзафлопсных вычислений

Колоссальный вычислительный потенциал современных суперкомпьютерных систем позволяет приступить к решению многих сложнейших научно-прикладных проблем, анализ которых ранее находился за гранью возможного. Динамика развития производительности суперкомпьютеров опережает все самые смелые ожидания специалистов. Так, в 2008 г. был преодолен петафлопсный (10^{15} операций с плавающей запятой в сек.) рубеж скорости вычислений, в 2018-2020 гг. ожидается достижение экзафлопсного (10^{18}) уровня производительности. По оценкам специалистов, подобные суперкомпьютерные системы с рекордными вычислительными показателями будут существенно многопроцессорными (до миллиарда вычислительных ядер), гибридными с разными типами вычислительных устройств с многоуровневой структурой организации вычислений (распределенные вычислительные устройства – вычислительные узлы с общей разделяемой памятью – многоядерные процессорные элементы – ускорители вычислений) [1]. Эффективное использование огромных вычислительных возможностей экзафлопсных систем представляет собой глобальную проблему «вызова» всему спектру вычислительных наук. Условиями для успешного преодоления этой проблемы являются:

- значительная **вычислительная трудоемкость** решаемых задач (свыше 10^{18} операций);
- высокий запас параллелизма (**масштабируемость**) вычислений (вплоть до использования 10^{9} процессоров);
- низкая интенсивность информационных взаимодействий (**локальность**) параллельно выполняемых вычислений;
- **устойчивость** вычислений к аппаратным сбоям вычислителей, которые неизбежно будут происходить при столь больших количествах вычислительных элементов.

В работе показывается, что все перечисленные условия достижимы для задач глобальной оптимизации.

2. Оценка вычислительной трудоемкости задач глобальной оптимизации

Прежде всего, следует отметить, что задачи оптимизации имеют важное прикладное значение и чрезвычайно широкую область приложений – практически каждая задача проектирования новых устройств, изделий, систем включает в себя этап выбора оптимальных вариантов. В наиболее сложных ситуациях проблема выбора формулируется как задача глобальной оптимизации, когда критерии оптимальности могут быть многоэкстремальными, т.е. иметь множество

* Исследование выполнено при поддержке РФФИ, грант № 11-01-00682-а, а также Министерства образования и науки Российской Федерации, соглашение № 14.B37.21.0393

локально-оптимальных решений, среди которых необходимо выделить наилучшие варианты. Допустимость многоэкстремальности существенно повышает сложность решения оптимизационной задачи, ибо, если для подтверждения локального минимума достаточно исследования локальной окрестности, то глобальный минимум является интегральной характеристикой решаемой оптимизационной задачи и требует исследования всей области глобального поиска. Как результат, поиск глобального оптимума сводится к построению некоторого покрытия (сетки) в области вариации параметров. Поэтому на сложность решения задач рассматриваемого класса решающее влияние оказывает размерность: они подвержены “проклятию размерности”, состоящему в экспоненциальном росте вычислительных затрат при увеличении размерности. Так, если вычисление одного значения оптимизируемой функции требует m операций, а количество узлов сетки при решении одномерной задачи равно k, то общая оценка общего объема вычислений для N-мерной задачи глобальной оптимизации имеет вид

$$T = mk^N.$$

Приняв, например, $m=10^3$, $k=10$, $N=20$, общий объем вычислений получается равным $T=10^{23}$, что требует порядка суток вычислений на суперкомпьютере экзафлопского уровня производительности.

Снижение объема вычислений может быть достигнуто при построении существенно неравномерных покрытии областей глобального поиска – эти сетки должны быть достаточно плотными в окрестности глобального оптимума и более разреженными вдали от искомого решения. Построение таких оптимальных покрытий обеспечивается при повышении сложности самих численных методов глобального поиска.

Применение неравномерных покрытий позволяет повысить размерность решаемых задач глобальной оптимизации в 2-3 раза, что является критичным в приложениях. Получаемые оценки размерности (25-50) позволяют охватить большинство научно-прикладных задач, поскольку, как правило, количество варьируемых параметров, по которым проявляется многоэкстремальность оптимизируемых критериев, ограничено.

3. Основные принципы организации вычислений в задачах глобальной оптимизации

3.1 Класс решаемых задач

Задача многомерной многоэкстремальной оптимизации может быть определена как проблема поиска наименьшего значения действительной функции $\varphi(y)$ [1]

$$\varphi(y^*) = \min\{\varphi(y): y \in D\},\ D = \{y \in R^N: a_i \le y_i \le b_i,\ 1 \le i \le N\}, \quad (1)$$

где a, $b \in R^N$ есть заданные векторы.

Численное решение задачи (1) сводится к построению оценки $y_\varepsilon^* \in D$, отвечающей тому или иному понятию близости к точке y^* на основе конечного числа k значений функционалов задачи, вычисленных в точках области D.

Относительно класса рассматриваемых задач предполагается выполнение двух важных принципиальных требований:

- процедуры проведения *испытаний* (вычисления значений целевой функции) является *вычислительно-трудоемкими*,
- функция $\varphi(y)$ должна удовлетворять *условию Липшица* (ограниченность изменения значений функции при ограниченной вариации параметров).

Следует отметить, что именно выполнимость условия Липшица позволяет строить численные оценки искомого глобального решения по конечному набору вычисленных значений функции $\varphi(y)$.

[1] Для уменьшения сложности изложения материала задача глобальной оптимизации в данной работе рассматривается при отсутствии функциональных ограничений.

Отметим также, что все далее излагаемые результаты получены в рамках информационно-статистической теории многоэкстремальной оптимизации [2, 3], являющейся одним из наиболее активно развиваемых направлений исследований в области глобального поиска.

3.2 Редукция размерности

Для снижения сложности алгоритмов глобальной оптимизации, формирующих эффективные покрытия области поиска, в теории и практике многоэкстремальной оптимизации широко используются различные схемы редукции размерности, которые позволяют свести решение многомерных оптимизационных задач к семейству задач одномерной оптимизации. Редукция размерности позволяет существенно снизить сложность разрабатываемых алгоритмов глобального поиска. Кроме того, данный подход позволяет задействовать весь имеющийся аппарат одномерной многоэкстремальной оптимизации для построения эффективных многомерных методов глобального поиска.

Один из наиболее общих методов редукции размерности состоит в применении разверток единичного отрезка вещественной оси на гиперкуб. Роль таких разверток играют непрерывные однозначные отображения типа кривой Пеано, называемые также кривыми, заполняющими пространство.

Использование развертки Пеано $y(x)$, однозначно отображающей единичный отрезок вещественной оси на единичный гиперкуб, позволяет свести многомерную задачу минимизации в области D к одномерной задаче условной минимизации на отрезке $[0,1]$

$$\varphi(y(x^*))=\min\{\varphi(y(x)):x\in[0,1]\}. \qquad (2)$$

Вопросы численного построения отображений типа кривой Пеано и соответствующая теория подробно рассмотрены в [2, 3].

3.3 Общая схема алгоритмов глобального поиска

Успехи в разработке одномерных алгоритмов глобального поиска в рамках информационно-статистической теории глобального поиска позволили получить чрезвычайно значимый результат – к формулированию общей схемы представления алгоритмов, в рамках которой с единых позиций можно провести анализ свойств методов и даже ставить задачу построения алгоритмов с заданными свойствами.

Данная общая схема, получившая наименование *схемой характеристической представимости алгоритмов глобального поиска* (см., например, [2]), состоит в следующем.

Алгоритм глобального поиска называется *характеристическим*, если, начиная с некоторого шага поиска $k_0\geq 1$, выбор точки x^{k+1} очередной итерации (*решающее правило алгоритма*) включает выполнение следующего набора действий:

1. Упорядочить поисковую информацию в порядке возрастания значений точек ранее выполненных итераций (новый порядок расположения данных в поисковой информации отражается при помощи нижнего индекса):

$$x_1 \leq x_2 \leq \ldots \leq x_k ,$$

2. Вычислить для каждого интервала $(x_{i-1}, x_i), 1 < i \leq k$, величину $R(i)$, называемую далее *характеристикой* интервала,

3. Определить интервал (x_{t-1}, x_t), которому соответствует максимальная характеристика $R(t)$, т.е.

$$R(t) = \max\{R(i) : 1 < i \leq k\},$$

4. Вычислить точку x^{k+1} очередной итерации глобального поиска в интервале с максимальной характеристикой в соответствии с некоторым правилом

$$x^{k+1} = s(t) \in (x_{t-1}, x_t) .$$

4. Параллельные вычисления для поиска глобально-оптимальных решений

4.1 Основной принцип организации параллельных вычислений

Прежде всего следует отметить, что в случае многомерной многоэкстремальной оптимизации многие широко используемые подходы для распараллеливания или не обеспечивают должного эффекта или вообще не применимы. В частности, один из основных способов организации параллельных вычислений состоит в разделении области решения задачи между процессорами и параллельного решения задачи в этих подобластях. В случае решения оптимизационных задач такой подход обладают низкой эффективностью, поскольку при разделении области поиска только небольшая часть процессоров будет обладать подобластью с искомым глобальным минимумом, все же остальные процессоры будут работать в подобластях, в которых отсутствует глобальный минимум исходной задачи.

Плохо применим в данном случае и подход, связанный с распараллеливанием непосредственно алгоритма решения задачи, ибо в силу сформулированных предположений о классе рассматриваемых задач основной объем вычислений в процессе глобального поиска занимают расчеты, связанные с вычислением значений функционалов оптимизационной задачи. В этом плане, целесообразным представляется распараллеливание процедур расчета значений функционалов, однако такой подход не обладает общностью и требует проведения действий по распараллеливанию для каждой оптимизационной задачи в отдельности.

С учетом выполненного обсуждения, в качестве основного принципа построения параллельных методов многоэкстремальной оптимизации в диссертации будет применяться схема параллельных вычислений, в которой будет обеспечиваться возможность одновременного (параллельного) расчета значений функционалов оптимизационной задачи в нескольких разных точках области поиска (см., например, [3, 4]). Такой подход характеризуется *эффективностью* (распараллеливается именно та часть вычислительного процесса, в котором выполняется основной объем вычислений) и *общностью*, поскольку применим для всех вычислительно-трудоемких задач многоэкстремальной оптимизации.

4.2 Параллельные вычисления для систем с распределенной памятью

Новый подход для организации параллельных вычислений в задачах глобальной оптимизации состоит в использование для редукции размерности множества отображений

$$Y_L(x)=\{y^1(x),\ldots,\ y^L(x)\}$$

вместо применения единственной кривой Пеано $y(x)$ [3, 4]. Использование множества отображений $Y_L(x)$ приводит к формированию соответствующего множества одномерных многоэкстремальных задач

$$\min\{\ \varphi(y^l(x)):x\in[0,1]\ \},\ 1\le l\le L. \qquad (3)$$

Каждая задача из данного набора может решаться независимо, при этом любое вычисленное значение $z=\varphi(y^i(x'))$ функции $\varphi(y)$ в i-й задаче может интерпретироваться как вычисление значения $z=\varphi(y^s(x''))$ для любой другой s-й задачи без повторных трудоемких вычислений функции $\varphi(y)$. Подобное информационное единство позволяет решать исходную задачу (1) путем параллельного решения L задач вида (4) на наборе отрезков [0,1]. Каждая одномерная задача решается на отдельном процессоре. Результаты вычислений в точке x^k, полученные конкретным процессором для решаемой им задачи, интерпретируются как результаты вычислений во всех остальных задачах (в соответствующих точках $x^{k1},\ldots,x^{kL}$). При таком подходе вычисления в точке $x^k\in[0,1]$, осуществляемое в s-й задаче, состоит в последовательности действий:

1. Определить образ $y^k=y^s(x^k)$ при соответствии $y^s(x)$.
2. Проинформировать остальные процессоры о начале проведения испытания в точке y^k (*блокирование* точки y^k).
3. Вычислить значение оптимизируемой функции $z^k=\varphi(y^s(x^k))$.
4. Определить прообразы $x^{kl}\in[0,1]$, $1\le l\le L$, точки y^k, и интерпретировать значение функции $\varphi(y)$, вычисленное в точке $y^k\in D$, как одновременные вычисления в L точках

$$x^{k1},\ldots,\ x^{kL},$$

с одинаковыми результатами

$$\varphi(y^1(x^{k1}))=\ldots=\ \varphi(y^L(x^{kL}))=z^k.$$

5. Проинформировать остальные процессоры о результатах испытания в точке *yk*.

Каждый процессор при таком подходе должен свою копию программных средств, реализующих вычисление функций задачи, и решающее правило алгоритма. Для организации взаимодействия на каждом процессоре создается L очередей, в которые процессоры помещают информацию о выполненных итерациях (точки очередных итераций и значения оптимизируемой функции).

Рассмотренный подход позволяет использовать для решения задач глобальной оптимизации порядка нескольких тысяч процессоров – ограничением возможного количества используемых вычислительных элементов является трудоемкость информационного взаимодействия при организации обменов результатов итераций глобального поиска между процессорами.

4.3 Параллельные вычисления для систем с общей разделяемой памятью

Ранее рассмотренная схема применима и для систем с общей разделяемой памятью, тем более что, что в этом случае будут отсутствовать затраты на обмен результатов итераций глобального поиска между процессорами. Вместе с этим, для организации параллельных вычислений может быть применена и другая схема, при которой параллельно может решаться и отдельные задачи семейства (3). Для обоснования подхода можно отметить, что используемые в алгоритмах характеристики интервалов (см. выше общую схему методов глобального поиска) могут рассматриваться как некоторые меры важности интервалов на предмет содержания в интервалах искомого решения оптимизационной задачи. Следуя данному пониманию, для организации параллельных вычислений после выбора точки вычисления значения функции для первого процессора в точном соответствии с последовательным алгоритмом для второго процессора точку очередной итерации целесообразно выбирать из следующего по важности интервала (т.е. из интервала со следующей по порядку максимальной характеристикой) и т.д.

На данной основе может быть предложена общая схема *параллельных асинхронных характеристических методов*, определяемая следующим образом [5].

Выбор точек очередных итераций глобального поиска начиная с некоторого шага поиска $k_0 \geq 1$ включает выполнение следующего набора действий:

1) При отсутствии свободных процессоров алгоритм находится в состоянии ожидания. Если закончили проведение испытаний $1 \leq q = q(k) \leq p(k)$ процессоров, то устанавливается номер итерации $k = k + q$, и осуществляется переход к следующему правилу общей схемы.

2) Упорядочить поисковую информацию в порядке возрастания значений точек ранее выполненных итераций (новый порядок расположения данных в поисковой информации отражается при помощи нижнего индекса):

$$0 = x_0 < x_1 < \ldots < x_{\tau-1} < x_k = 1\,.$$

3) Вычислить для каждого интервала $(x_{i-1}, x_i), 1 < i \leq k$, величину $R(i)$, называемую далее *характеристикой* интервала,.

4) Упорядочить характеристики $R(i), 1 \leq i \leq \tau,$ в порядке убывания:

$$R(t_1) \geq R(t_2) \geq \ldots \geq R(t_{k-1}) \geq R(t_k) \qquad (4)$$

5) Выбрать в последовательности (4) q наибольших характеристик с номерами $t_j, 1 \leq j \leq q,$ и в интервалах, соответствующих этим характеристикам, выполнить испытания в точках

$$\bar{x}^j = d(t_j) \in (x_{t_j-1}, x_{t_j}),\ 1 \leq j \leq q.$$

Как можно заметить, рассмотренная общая схема практически полностью соответствует схеме характеристической представимости алгоритмов глобального поиска.

Отметим также, рассмотренная схема асинхронных параллельных вычислений полностью применима и для случая, когда вычислительные узлы с общей разделяемой памятью содержат ускорители вычислений вида современных графических процессоров. Отличие состоит лишь в

том, что определяемые точки итераций глобального поиска должны передаваться для использования ускорителям вычислений.

Предполагая наличие количество вычислительных элементов (ядер) на общей памяти порядка 100, использование параллельных алгоритмов в рамках рассмотренных схем параллельного глобального поиска, позволяет довести возможное число одновременно используемых вычислительных устройств суперкомпьютерных систем экзафлопсного уровня производительности до нескольких сотен тысяч.

Литература

1. International Exascale Software Project. – http://www.exascale.org
2. Стронгин Р.Г. Численные методы в многоэкстремальных задачах. М.: Наука, 1978.
3. Strongin R.G., Sergeyev Ya.D. Global optimization with non-convex constraints. Sequential and parallel algorithms. Kluwer Academic Publishers, Dordrecht, 2000.
4. Стронгин Р.Г., Гергель В.П., Баркалов К.А. Параллельные методы решения задач глобальной оптимизации // Известия высших учебных заведений. Приборостроение. 2009. Т. 52. № 10. С. 25–33.
5. Стронгин Р.Г., Гергель В.П., Гришагин В.А., Баркалов К.А. Параллельные вычисления в задачах глобальной оптимизации. – М.: МГУ, 2012.