УДК 004.421.2

МЕТОДИКА ОПРЕДЕЛЕНИЯ ЭФФЕКТИВНОСТИ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ

В.С. Выхованец
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
Email: valery@vykhovanets.ru

Ключевые слова: параллельные вычисления, конвейерные вычисления, кратные вычислений, ускорение вычислений, степень использования оборудования, комплексная эффективность вычислений.
Key words: parallel computing, pipeline computing, multiple computing, speed-up of program, equipment efficiency, and complex efficiency of computing.

Рассматривается задача определения эффективности параллельных вычислений. Показывается, что параллельные вычисления являются частным случаем кратных вычислений. Описывается методика определения комплексной эффективности кратных вычислений, учитывающая как эффективность по времени, так и пространственную (аппаратурную) эффективность.
MEASUREMENT OF PARALLEL COMPUTING EFFICIENCY / V.S. Vykhovanets (V.A. Trapeznokov Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117997, Russia, Email: valery@vykhovanets.ru). The paper considers the determination of parallel computing efficiency. Parallel computing as a kind of multiple computing is discussed. Measurement of complex efficiency, which takes into account time efficiency (speed-up of program) as well as space (equipment) efficiency, is described.

1. Введение

Обычно под вычислениями понимается выполнение последовательности операций (команд, процессов, подзадач), направленных на преобразование входных данных в выходные. Различают последовательные и параллельные вычисления. Последовательные вычисления представляются как линейно упорядоченная последовательность операций над данными, а параллельные – как частично упорядоченная.

При параллельных вычислениях имеются операции, которые выполняются независимо от других, в противоположность последовательным вычислениям, у которых все операции выполняются строго последовательно, т.е. результат предыдущей операции представляется необходимым для выполнения следующей. Очевидно, что параллельные вычисления могут выполняться последовательно, но не наоборот. Отсюда универсальными следует признать последовательные вычисления, а параллельные следует рассматривать как частный случай последовательных.

Возможности дальнейшего наращивания производительности вычислительных средств в рамках последовательного принципа обработки данных считаются практически исчерпанными. Поиск путей повышения производительности идет в направлении развития, как нетрадиционных принципов обработки данных, так и методов структурной организации вычислений: распараллеливания и конвейеризации. Последние предназначены для выполнения одних и тех же вычислений одновременно на нескольких наборах исходных данных, когда в один и тот же момент времени выполняются одновременно несколько операций (команд, процессов, подзадач). Причем, в случае параллельных вычислений, осуществляется введение избыточ-

ных вычислительных устройств, а при конвейерных вычислениях – их структурная реорганизация.

Определение эффективности параллельных вычислений сводится к оценке степени использования оборудования и максимально возможного ускорения процесса вычислений (сокращения времени решения задачи), определяемых относительно последовательных вычислений, направленных на решение той же самой задачи [1]. Формирование подобных оценок осуществляется применительно к одному и тому же вычислительному алгоритму, т.е. определяться оценка эффективности распараллеливания конкретного алгоритма решения задачи, а не задачи в целом. В этом случае используется модель вычислительного процесса, представленная в виде графа зависимостей типа «операции – операнды» [2].

Следует заметить, что в описанном выше подходе учитываются только временные ресурсы вычислительной системы, а аппаратурные ресурсы предполагаются, как правило, неиссякаемыми. Известны модифицированные методики оценки ускорения вычислений, в которых ресурсы ограничиваются некоторым «окном исполнения», определяемым как максимально допустимое количество параллельно выполняющихся операций [3]. Команды в «окне» могут исполняться параллельно, если между ними нет зависимости по данным и управлению. Однако в таких методиках точные аппаратурные затраты все также не учитывается в получаемых оценках эффективности, они используются только для ограничения количества параллельно выполняемых операций (команд, процессов, подзадач).

Поставим целью разработать методику определения эффективности параллельных вычислений, учитывающую не только временные, но и пространственные (аппаратурные) затраты. В качестве основной модели вычислений будем использовать модель многопоточного параллелизма, подразумевающую использование нескольких потоков команд и данных для достижения параллельного исполнения операций на нескольких процессорах или процессорных ядрах.

2. Содержательная постановка задачи

Пусть задан вычислительное средство F (рис. 1), решающее некоторую элементарную задачу вычислительного характера $F(X) = Y$, где X – входные данные, а Y – выходные. Будем считать, что распараллеливание вычислений F уже выполнено или признано невозможным (нецелесообразным).

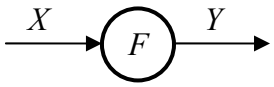


Рис. 1. Элементарные вычисления

В общем случае входные данные X могут быть представлены как состоящие из двух частей $X = [X', X'']$, где X' – программа (команда, операция), применяемая для обработки X'' – оставшейся части входных данных (аргументов, операндов). В рассматриваемом случае программа соотносится с крупноблочным уровнем вычислительного средства (компьютер), команда – со среднеблочным (процессор), а операция – с мелкоблочным (операционное устройство).

Задачу параллельных вычислений сведем к кратным вычислениям G (рис. 2) [4], реализующим вычисление F на $m > 1$ входных наборах данных $\mathbf{X} = [X^1, X^2, \ldots, X^m]$ для получения выходных наборов данных $\mathbf{Y} = [Y^1, Y^2, \ldots, Y^m]$, т.е. $G(\mathbf{X}) = \mathbf{Y}$.

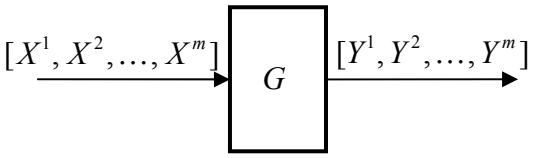


Рис. 2. Кратные вычисления

Для оценки комплексной эффективности E кратных вычислений G будем использовать временную T и пространственную V эффективности,

$$E = TV . \quad (1)$$

При расчете эффективностей за единицу измерения времени и пространства примем время и аппаратурные затраты при однократных вычислениях F. Временную эффективность T определим как ускорение вычислений, полученное в пересчете на один набор данных,

$$T = \frac{1}{\tau/m} = \frac{m}{\tau}, \quad (2)$$

где τ – время кратных вычислений G, а пространственную эффективность V определим отношение аппаратурных затрат в однократном случае F к аналогичным затратам при кратных вычислениях G,

$$V = \frac{1}{\nu}, \quad (3)$$

где ν – аппаратурные (пространственные) затраты кратных вычислений G.

В итоге имеем, что комплексная эффективность E определяет время вычисления на одном наборе данных в пересчете на единицу оборудования. Если $E > 1$, то получено повышение эффективности при кратных вычислениях G по отношению к F, а при $E < 1$ – уменьшение этой эффективности.

3. Организация кратных вычислений

При последовательных вычислениях входные наборы данных используются последовательно во времени, а при параллельных – параллельно в пространстве. Однако видится возможной организация параллельных вычислений не только последовательно во времени и параллельно в пространстве, но и последовательно в пространстве и параллельно во времени (табл. 1).

3.1. Последовательные вычисления во времени

При последовательных вычислениях во времени (повторная обработка) различные наборы данных подаются на вход одного вычислительного средства F в разные моменты времени t, отстоящие друг от друга на время однократного вычисления, т.е. временная последовательность входных наборов

$$X^1(1),\ X^2(2),\ \ldots, X^m(m)$$

порождает соответствующую временную последовательность результатов

$$Y^1(2),\ Y^2(3),\ \ldots, Y^m(m+1).$$

Тогда для повторной обработки имеем

$$\tau = m,\quad \nu = 1,\quad T = 1,\quad V = 1,\quad E = 1.$$

Примером реализации последовательных вычислений во времени на крупноблочном уровне является вычислительная система с одиночным потоком команд и одиночным пото-

ком данных, на среднеблочном – процессор скалярной обработки данных, на мелкоблочном – устройство выполнения арифметических операций.

3.2. Параллельные вычисления в пространстве

При параллельных вычислениях в пространстве (распараллеливание или многоэлементная обработка) используется m вычислительных средств F, каждое из которых выполняет вычисление на своем наборе. В этом случае результаты появляются на выходе через время однократного вычисления, но при этом требуется дублирование оборудования, равное кратности вычислений m. Входные наборы данных образуют пространственную последовательность и подаются на входы F одновременно:

$$X^1(1),\ X^2(1),\ \ldots, X^m(1),$$

и одновременно появляются на выходе соответствующие результаты:

$$Y^1(2),\ Y^2(2),\ \ldots, Y^m(2).$$

Комплексная эффективность многоэлементной обработки, как и при последовательных вычислениях во времени, будет равна единице:

$$\tau = 1,\quad \nu = m,\quad T = 1,\quad V = 1,\quad E = 1.$$

Примером реализации последовательных вычислений во времени на крупноблочном уровне является вычислительная система с множественным потоком команд и множественным потоком данных, на среднеблочном – процессор векторной обработки данных, на мелкоблочном – устройство выполнения поразрядных логических операций.

3.3. Последовательные вычисления в пространстве

При последовательных вычислениях в пространстве (конвейеризация или многостадийная обработка) различные наборы данных подаются на вход также в разные моменты времени, но отстоящие друг от друга на время $1/s$, которого достаточно для вычислений, осуществляемых каждой из s ступеней (стадий) конвейера.

Для последовательных вычислений в пространстве вычислительное средство F разделяется на s последовательных частей (стадий), таким образом, чтобы результат вычисления предыдущей стадии являлся входными данными для вычисления следующей стадии. В этом случае временная последовательность входных наборов

$$X^1(1),\ X^2(1+1/s),\ \ldots, X^m(1+(m-1)/s)$$

и порождает соответствующую временную последовательность результатов

$$Y^1(2),\ Y^2(2+1/s),\ \ldots, Y^m(2+(m-1)/s),$$

откуда получаем выражение для комплексной эффективности:

$$\tau = 1+(m-1)/s,\quad \nu = 1,\quad T = \frac{m}{1+(m-1)/s},\ V = 1,\ E = \frac{sm}{s+m-1} \geq 1. \qquad (4)$$

Заметим, что при $s = 1$ имеем ранее рассмотренную повторную обработку с единичной комплексной эффективностью. Однако при числе стадий $s > 1$ комплексная эффективность кратных вычислений будет больше единицы, а в пределе стремиться к s. Более того из формул (4) непосредственно следует, что при кратности вычислений $m = 1$ последовательные вычисления в пространстве трансформируются в последовательные вычисления во времени.

Примером реализации последовательных вычислений во времени на крупноблочном уровне является перенаправление выходного потока данных одной программы во входной поток другой, на среднеблочном – процессор конвейерной обработки данных, на мелкоблочном – конвейеризованное устройство умножения-сложения чисел в формате с плавающей запятой.

3.4. Параллельные вычисления во времени

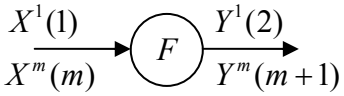
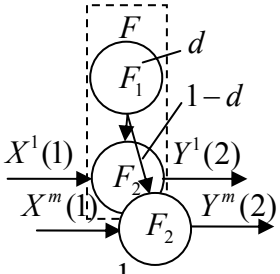
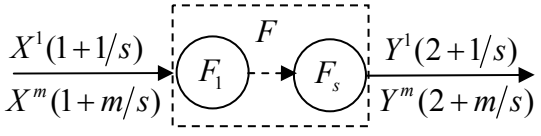
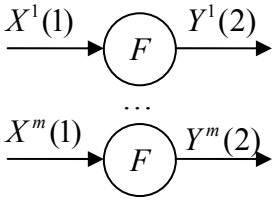
При параллельных вычислениях во времени используется m вычислительных средств F, но не в полном объеме: общая часть оборудования F_1 с долей d не дублируется, а используется в одном экземпляре. В этом случае результаты появляются на выходе через время однократного вычисления, но при этом требуется дублирование оборудования параллельных каналов F_2 с кратностью m. Входные наборы данных, как и при параллельных вычислениях в пространстве, образуют пространственную последовательность и подаются на входы F одновременно, одновременно снимаются и результаты вычислений, но с задержкой на время однократного вычисления. Следовательно, для параллельных вычислений во времени

$$\tau = 1, \quad \nu = \frac{1}{m(1-d)+d}, \quad T = m,\ V = \frac{1}{m(1-d)+d},\ E = \frac{m}{m(1-d)+d} \geq 1. \qquad (5)$$

Заметим, что при $d = 0$ имеем рассмотренную ранее многоэлементную обработку с единичной комплексной эффективностью. Однако при доле общего оборудования $d > 0$ комплексная эффективность кратных вычислений будет больше единицы, а в пределе стремиться к $1/(1-d)$. Более того из формул (5) непосредственно следует, что при кратности вычислений $m = 1$ параллельные вычисления во времени трансформируются в параллельные вычисления в пространстве.

Примером реализации параллельных вычислений во времени на крупноблочном уровне является вычислительная система с одиночным потоком команд и множественным потоком данных или с множественным потоком команд и одиночным потоком данных, на среднеблочном – процессор суперскалярной обработки данных, на мелкоблочном – многоканальный аналогово-цифровой преобразователь.

Таблица 1. Методы организации кратных вычислений

	Последовательно	Параллельно
Во времени	 $T = 1,\ V = 1,\ E = 1$	 $T = m,\ V = \frac{1}{m(1-d)+d},\ E = \frac{m}{m(1-d)+d}$
В пространстве	 $T = \frac{m}{1+(m-1)/s},\ V = 1,\ E = \frac{sm}{s+m-1}$	 $T = m,\ V = \frac{1}{m},\ E = 1$

4. Методика определения эффективности

Определение эффективности параллельных вычислений возможно двумя методами: априорным и апостериорным.

4.1. Априорная эффективность

При априорном определении эффективности параллельных вычислений выполняется расчет комплексной эффективности на основе знания структурной организации планируемых кратных вычислений. В этом случае для последовательных вычислений во времени и параллельных вычислений в пространстве комплексная эффективность принимается равной единице, а для последовательных вычислений в пространстве и параллельных вычислений во времени используются формулы (4) и (5) соответственно.

Некоторую сложность представляет определение доли общего оборудования d для нескольких параллельных каналов обработки данных. Для оценки d могут быть использованы такие характеристики вычислительных средств и их частей как количество логических элементов, пространственный объем, площадь на кристалле, и т.п.

В случае, когда кратность вычислений m неизвестна, производится подсчет асимптотической (максимальной) комплексной эффективности, получаемой при устремлении кратности вычислений m к бесконечности. В этом случае вместо формул (4) и (5) используются формулы $E = s$ и $E = 1/(1-d)$ соответственно.

Если структура кратных вычислений такова, что одновременно используются последовательные и параллельные вычисления, то комплексная эффективность такой организации принимается равной произведению их комплексных эффективностей. В этом случае наибольшая комплексная эффективность достигается при совмещении последовательных вычислений в пространстве с параллельными вычислениями во времени (рис. 3),

$$E = \frac{sm}{(s + m' - 1)(m''(1-d) + d)},$$

где m – общая кратность вычислений, m' – кратность последовательных вычислений, m' – кратность параллельных вычислений, $m = m'm''$.

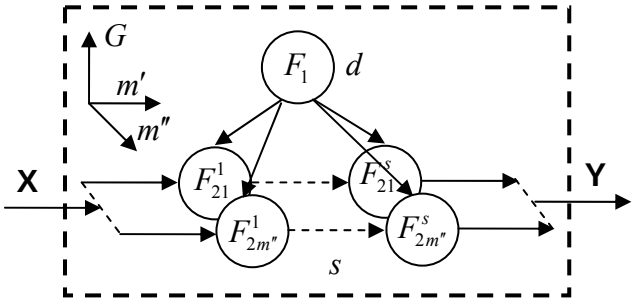


Рис. 3. Совмещенная организация вычислений

4.2. Апостериорная эффективность

При апостериорном определении эффективности параллельных вычислений выполняется вычисление комплексной эффективности на основе измерения временных и пространственных характеристик выполняемых вычислений.

Теоретической основой апостериорного метода служат следующие свойства кратных вычислений:

1) последовательные вычисления во времени и параллельные вычисления в пространстве, будучи использованными в любой форме, не влияют на комплексную эффективность кратных вычислений, т.к. их комплексная эффективность всегда равна единице;

2) в однократном случае ($m=1$) последовательные вычисления в пространстве трансформируются в последовательные вычисления во времени, а параллельные вычисления во времени – в параллельные вычисления в пространстве.

Возможность определения значений единиц измерения времени и пространства непосредственно следует из свойств 1) и 2). Для этого выполняются однократные вычисления и замеряются соответствующие характеристики выполненного преобразования данных. Единицей измерения времени в этом случае может служить временной интервал τ_1, затраченный вычислительным средством на выполнение однократных вычислений, а единицей измерений пространства – дополнительная энергия v_1, потребленная вычислительным средством при однократных вычислениях.

Аналогичные измерения τ_m и v_m производятся при m-кратных вычислениях таких, что задействуются как конвейер, так и все параллельные каналы вычислительного средства. Для этого необходимо выбрать m достаточно большим. В этом случае и с учетом ранее определенных единиц измерения времени τ_1 и пространства v_1 эффективности выполненных кратных вычислений в соответствии с (1-3) могут быть оценены по следующим формулам:

$$T=\frac{\tau_1}{\tau_m/m}=m\frac{\tau_1}{\tau_m},\quad V=\frac{v_1}{v_m},\quad E=m\frac{\tau_1 v_1}{\tau_m v_m},$$

где τ_m – время, затраченное на m-кратные вычисления, v_m – дополнительная энергия, потребленная при m-кратных вычислениях.

5. Заключение

Таким образом, используя введенные критерии эффективности, появляется априорная и апостериорная возможность оценить, за счет чего получено ускорение параллельных вычислений: при $E=1$ – за счет пропорционального увеличения аппаратурных затрат (экстенсификация вычислений), или при $E>1$ – за счет эффективной организации вычислительного средства (интенсификация вычислений). Следует также рассматривать случай деградации вычислений, когда получаемая комплексная эффективность меньше единицы, $E<1$. В этом случае, как правило, добиваются ускорения параллельных вычислений, не считаясь с пространственными (аппаратурными) затратами.

Из приведенного анализа также следует, что последовательные вычисления в пространстве (конвейеризация) являются одной из форм параллельных вычислений, однако осуществляемых в скрытой форме. Следует также обратить внимание на параллельные вычисления во времени, которые являются интенсивной формой параллельных вычислений по определению [5]. В противоположность конвейерным вычислениям, где исходная задача F декомпозируется на последовательно выполняемые подзадачи $F_1-F_2-...-F_s$, при параллельных вычислениях во времени осуществляется параллельная декомпозиция $F=F_1|F_2$, заключающаяся в выделении общего оборудования F_1 (элементов, устройств, областей памяти, и т.п.) с долей d от всего оборудования F, которое используется для нескольких параллельных каналов обработки данных F_2.

Список литературы

1. Roosta S.H. Parallel Processing and Parallel Algorithms: Theory and Computation. – NY: Springer-Verlag, 2000.
2. Bertsekas D.P., Tsitsiklis J.N. Parallel and Distributed Computation. Numerical Methods. – Prentice Hall: Englewood Cliffs, 1989.
3. Smith J.E., Pleszkun A.R. Implementation of precise interrupts in pipelined processors // Proceedings of the 12th annual international symposium on Computer architecture. – Boston, 1985. P. 36-44.
4. Выхованец В.С., Малюгин В.Д. Кратные логические вычисления // Автоматика и телемеханика. – 1998, № 6. – С. 163-171.
5. Выхованец В.С. Параллельные вычисления во времени // Автоматика и телемеханика. – 1999, № 12. – С. 155-165.