Информационные системы и технологии

УДК 621.3
Д.И. Лазоренко

Объединение циклов для снижения энергопотребления

Предложен способ объединения циклов на уровне исходного текста с целью понижения энергопотребления проектируемых устройств за счет уменьшения обращений к основной памяти. Способ является основой для автоматизации трансформации кода программ на языках высокого уровня.

A method of loop fusion is presented in this paper. This method of code-to-code transformations serves for the purpose of power consumption reduction of digital systems through reduction of the number of accesses to the main memory. The algorithm is a basis for automation of code-to-code transformations.

Введение

Развитие полупроводниковых технологий привело к возникновению концепции Системы-на-Кристалле. Сложность современных приложений и использование субмикронных технологий обуславливают необходимость снижения энергопотребления таких систем путём применения оптимальных решений в процессе проектирования.

Современные цифровые системы, например, мультимедийные приложения, переносные телефоны, карманные персональные компьютеры, обрабатывают большие массивы данных по сложным алгоритмам. Развитие технологий переносных источников питания не успевает за увеличением энергопотребления новых приложений. Кроме того, пониженное энергопотребление позволяет упростить разводку шин питания на кристалле, приводит к уменьшению шумов на шинах питания, проявления эффекта электромиграции и электромагнитного излучения.

Источники энергопотребления в КМОП схемах

Подавляющее число современных микросхем производится с помощью КМОП (Комплементарный Металл-Оксид-Полупроводник) технологии, поэтому источники энергопотребления будут рассматриваться применительно к этой технологии. Существует четыре составляющие энергопотребления КМОП схем – токи короткого замыкания, статические токи, паразитные токи утечки, и динамическое рассеяние энергии.

$$P_{total} = P_{short} + P_{stat} + P_{leak} + P_{dyn}$$

Токи короткого замыкания P_{short}. Токи короткого замыкания существуют в процессе нормального функционирования логических схем. Под ними понимаются токи, протекающие через транзисторы из шины питания непосредственно в шину земли.

Статические токи P_{stat}. Под статическими понимаются постоянные токи.

Паразитные токи утечки P_{leak}. Под паразитными токами утечки понимаются подпороговые токи транзисторов и токи в подложке.

Динамическое рассеяние энергии P_{dyn}. Динамическое рассеяние энергии происходит из-за зарядки/разрядки узлов схемы, и может быть представлено следующей формулой:

$$P_{dyn} = C \cdot V_{dd}^2 \cdot \alpha \cdot f ,$$

где C – суммарная ёмкость в узлах схемы, V_{dd} – величина напряжения питания, f – частота переключений, α - коэффициент переключательной активности (среднее число логических вентилей, переключающихся в течение одного цикла сигнала синхронизации) [1].

Динамическое рассеяние энергии может составлять до 80% от полных потерь энергии. Переключательная активность в значительной мере определяется программным обеспечением цифровой системы [2,3].

В работе [4] обоснован вывод, что потенциальный выигрыш в энергопотреблении тем выше, чем выше уровень абстракции процесса проектирования, на котором принимается решение. Для системного уровня выигрыш может быть от 50% до 90%, на поведенческом уровне – от 40% до 70%, на RTL (Register Transfer Level) уровне – от 30% до 50%, на уровне вентилей – от 20% до 30%, на уровне транзисторов – от 10% до 20%, на уровне топологии – от 5% до 10%.

Современные цифровые системы используют большие объёмы памяти. Схемы памяти могут занимать от 50% до 80% площади полупроводникового кристалла. Известно, что схемам памяти присущи большие паразитные токи утечки. Кроме того, на обращение к памяти тратится много энергии, например, на операцию чтения из внешней памяти расходуется в 33 раза более энергии, чем на операцию 16-битного сложения. Согласно прогнозу международной организации International Technology Roadmap for Semiconductors схемы памяти будут занимать всё больше площади на полупроводниковых кристаллах: в 2008 г. – 83%, в 2011 г. – 90%, в 2014 г. – 94%. Также, величина динамического энергопотребления схем памяти в ближайшем будущем будет только увеличиваться. Например, по прогнозу на 2010 г. динамическое рассеяние энергии на схемах памяти будет в два раза больше, чем на логических схемах, к 2015 г. эта величина увеличиться до 2,5 раз, к 2020 г. – до 3 раз [5].

Очевидно, в процессе проектирования нужно добиваться уменьшения объёма требуемой приложению памяти и количества обращений к ней. Для этого необходимо оптимизировать систему на поведенческом уровне. Важно, чтобы такая оптимизация про-

водилась перед разделением системы на программную и аппаратную части, поскольку это позволяет применить однообразный подход к обработке всей памяти. Циклы «for» представляют собой именно ту часть исходного кода приложения, которая ответственна за использование массивов в приложениях, обрабатывающих большие объёмы данных по сложным алгоритмам [6,7,8].

Для снижения объёма требуемой памяти необходимо уменьшить размер и количество временных массивов, создаваемых в процессе обработки данных. Уменьшения объёма памяти можно также добиться путём повторного использования одних и тех же адресов памяти разными массивами.

Алгоритм объединения многомерных циклов

Известно, что основным источником энергопотребления схем памяти являются операции чтения и записи. В данной статье предлагается алгоритм анализа циклов на возможность их объединения, что позволяет уменьшить количество обращений к памяти и её объём. Данный алгоритм использует представление исходного текста программы в виде графов и основан на предложенном в [9] графическом методе объединения одномерных циклов исходного текста описания цифровых систем. Рассмотрим далее формирование и преобразование упомянутых графов.

Для наглядности на рисунках будут использоваться двумерные массивы. Пусть есть некоторый исходный текст программы, содержащий многомерные циклы (рис. 1а). На рис. 1б изображён граф исходного текста программы. Его вершины L1 и L2 соответствуют вычислению элементов массивов **a1** и **a2**. Между циклами существует зависимость по данным, эта зависимость представлена в графе дугой направленной от вершины L1 к вершине L2. Дуге присваивается набор весов, количество которых равно размерности массивов. Сами величины весов вычисляются следующим образом. Допустим, элементы массива $a2[i_1,i_2,\ldots,i_n]$ вычисляются при помощи элементов массива $a1[i_1-d_1,i_2-d_2,\ldots,i_n-d_n]$. Вес, соответствующий i_k итерационной переменной, будет вычисляться, как разница k-х значений индексов для элементов массива **a2** и **a1**: $i_k-(i_k-d_k)=d_k$. Тогда набор весов данной дуги будет таким - $(d_1,d_2,\ldots,d_n)$. Ярлык «f» нужен для обозначения типа зависимости по данным, поскольку ниже будет рассматриваться также тип связи по выходу.

В работе [9] описан способ трансформации графа представления зависимостей при вычислении одномерных циклов, благодаря которому становится возможным их объединение. Ниже предлагается основанный на упомянутом способе алгоритм преобразования графов для многомерных циклов.

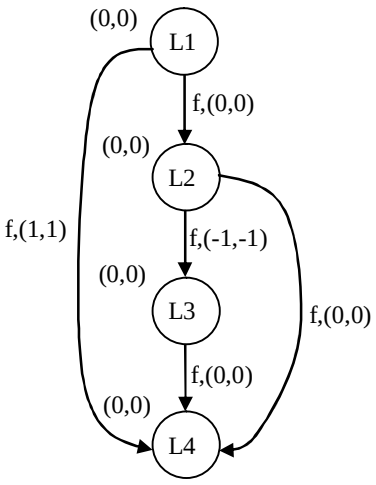
Перед началом преобразования графа присвоим каждой его вершине набор весов, количество которых равно размерности массивов. В исходном состоянии все значения весов в наборе будут нулевыми. Данные значения весов будет изменять по правилам приводимым ниже.

Рассмотрим преобразования графа на примере некоторого исходного текста программы на рис. 1а. На рис. 1б показан исходный граф.

Аналогично тому, как было сделано в работе [9], необходимо трансформировать граф таким образом, чтобы в нём не осталось f-дуг с отрицательными значениями весов. Для вершин и f-дуг изменение каждого веса в наборе, соответствующего одной из итерационных переменных, производится точно так же, как и в одномерном случае. Если вес, соответствующий

```
for (int i = 0; i < N; i++)
 for (int j = 0; j < N; j++)
  a1[i][j] = f1(i,j);
…
for (int i = 0; i < N; i++)
 for (int j = 0; j < N; j++)
  a2[i][j] = f2(a1[i][j]);
…
for (int i = 0; i < N - 1; i++)
 for (int j = 0; j < N - 1; j++)
  a3[i][j] = f3(a2[i+1][j+1]);
…
for (int i = 1; i < N; i++)
 for (int j = 1; j < N; j++)
a4[i][j] = f4(a1[i-1][j-1], a2[i][j],a3[i][j]);
```

а) исходный текст программы



б) граф вычислений

Рис. 1

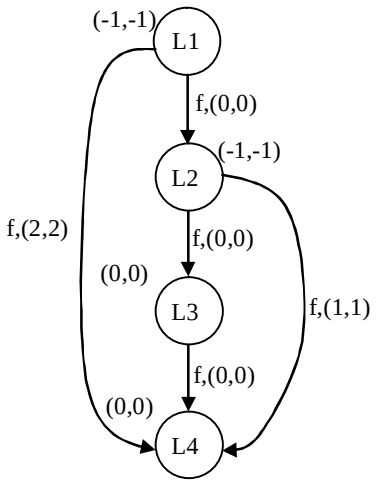
k-ой итерационной переменной, какой-либо дуги увеличивается на определённую величину, то вес, соответствующий k-ой итерационной переменной, вершины, из которой выходит данная дуга, уменьшается на данную величину. Вес, соответствующий k-ой итерационной переменной, дуг, входящих в эту вершину, также уменьшается на ту же величину, а вес, соответствующий k-ой итерационной переменной, дуг, выходящих из этой вершины, увеличивается на упомянутую величину.

Затем, двигаясь к вершине L1 (началу графа), проводим аналогичные преобразования для всех весов в наборах. На рис. 2 показан конечный преобразованный граф и текст объединённого цикла.

Вес вершин графа используется при формировании текста объединённого цикла.

Теперь рассмотрим объединение циклов, если между ними присутствует связь по выходу. На рис. 3а представлен некоторый исходный текст программы, содержащий три цикла. На рис. 3б представлен граф, соответствующие исходному тексту программы на рис. 3а. Каждой вершине присваивается набор нулевых весов. Веса f-дуг вычисляются так же, как это было сделано выше. Дуга с ярлыком «o» отображает наличие связи по выходу и не имеет веса. Между первым и третьим циклами существует связь по выходу. В исходном виде данные три цикла объединить невозможно. Например, значение элемента a[1,1] необходимо для вычисления b[2,2], но до этого оно уже перезаписывается во время выполнения третьего цикла. Итоговый граф для данного случая показан на рис. 4а, а текст объединённого цикла – на рис. 4б.

```
for (i = 2; i < N; i++)
 for (i = 2; j < N; j++)
{
  a1[i][j] = f1(i,j);
  a2[i][j] = f2(a1[i][j]);
  a3[i-1][j-1] = f3(a2[i][j]);
  a4[i-1][j-1] =
 f4(a1[i-2][j-2],a2[i-1][j-1],a3[i-1][j-1]);
}
```

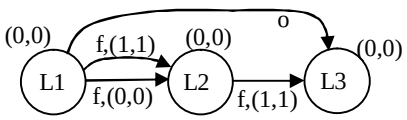


а) текст объединённого цикла **б) итоговый граф**

Рис. 2

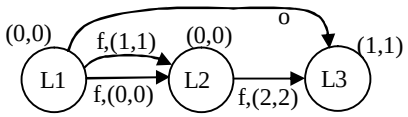
```
…
for (int i = 0; i < N; i++)
 for (int j = 0; j < N; j++)
  a[i][j] = f1(i,j);
…
for (int i = 1; i < N; i++)
 for (int j = 1; j < N; j++)
  b[i][j] = f2(a[i][j],a[i-1][j-1]);
…

for (int i = 0; i < N - 1; i++)
 for (int j = 0; j < N - 1; j++)
  a[i+1][j+1] = f3(b[i][j]);
…
```



а) исходный текст программы **б) исходный граф**

Рис. 3



а) итоговый граф

```
for (int i = 2; i < N; i++)
 for (int j = 2; j < N; j++)
{
   a[i][j] = ƒ1(i,j);
   b[i][j] = ƒ2(a[i][j],a[i-1][j-1]);
   a[i-1][j-1] = ƒ3(b[i-2][j-2]);
}
```

б) текст объединённого цикла

Рис. 4

Чтобы сделать объединение циклов возможным, необходимо перезаписывать значение a[i,j] уже после того, как оно было использовано для вычисления b[i+1,j+1]. Для этого требуется трансформировать граф следующим образом. Увеличиваем каждый k-ый вес вершины L3 (на которой заканчивается o-дуга) на величину, равную максимальному значению из всех k-ых весов всех f-дуг, исходящих из вершины L1 (из которой исходит та же o-дуга). Например, из вершины L1 исходят две f-дуги, первый вес из набора одной из них равен 1, для другой – эта величина равна 0. Максимальное значение из приведённых двух равно 1, поэтому первый из набора весов вершины L3 должен увеличиться на 1. При этом веса всех f-дуг, входящих в L3, должны увеличиться на ту же величину, на которую увеличились соответствующие веса вершины L3, веса же всех исходящих из данной вершины f-дуг должны уменьшиться на упомянутую величину.

Выводы

Предлагаемый алгоритм представляет собой способ анализа циклов с точки зрения возможности их объединения. Метод, предложенный в [7], позволяет объединить только три их пяти циклов в тестовом примере, в то время как метод, предлагаемый в данной работе, позволяет добиться лучших результатов - объединить все циклы. Также, данный способ позволяет анализировать влияние связей внутри одного цикла на возможность объединения его с другим, что отсутствует в методе, предлагаемом в [7]. В результате объединения циклов, выполненного согласно предлагаемому методу, можно существенно сократить количество операций обращения к медленной памяти, и уменьшить её размер. Предлагаемый алгоритм позволяет также минимизировать количество данных, которые необходимо хранить в быстрой памяти в процессе вычислений.

Литература

1. *Veendrick H.J.M.* Deep-Submicron CMOS ICs. From Basics to ASICs. - Kluwer academic publishers, 2000. – 539 p.
2. *Poppen F.* Low Power Design Guide. - OFFIS Research Institute. http://www.offis.de
3. *Kim H.S.,* Irwin M.J., Vijaykrishnan N., Kandemir M. Effect of compiler optimizations on memory energy.//2000 IEEE Workshop on Signal Processing Systems. SiPS 2000. - 2000. – p. 663 - 672.
4. *Sproch J.* High Level Power Analysis and Optimization. Tutorial.//1997 International Symposium on Low Power Electronics and Design. - 1997.
5. International Technology Roadmap for Semiconductors. http://public.itrs.net/
6. *Fraboulet A., Huard G., Mignotte A.* Loop Alignment for Memory Accesses Optimization.//Twelfth International Symposium on System Synthesis. Proceedings (ISSS'99). IEEE Computer Society Press. - 1999. – P. 71–77.
7. *Fraboulet A.,* Kodary K., Mignotte A. Loop fusion for memory space optimization.//The 14th International Symposium on System Synthesis. Proceedings 2001. – 2001. – P. 95–100.
8. *Catthoor F.,* Franssen F., Wuytack S., Nachtergaele L., De Man H. Global communication and memory optimizing transformations for low power signal processing systems.//Workshop on VLSI Signal Processing, VII. – 1994. – P. 178 – 187.
9. *Лазоренко Д.И.* Алгоритм объединения одномерных циклов исходного текста описания цифровых систем с целью снижения их энергопотребления // Системи обробки інформації. – Х.:ХУПС, 2007. – Вип. 8(66). – С. 02–12.