МЕТОДЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ЭФФЕКТИВНОСТИ РЕАЛИЗАЦИИ ФРАГМЕНТОВ ПРИКЛАДНЫХ ПРОГРАММ НА ЯЗЫКЕ COLAMO

А.И. Дордопуло, И.И. Левин, В.А. Гудков

Одной из основных технических характеристик вычислительной системы (ВС) является ее производительность, под которой понимается число одновременно выполняемых операций над данными определенной разрядности и формы представления в единицу времени. Для оценки эффективности вычислительной системы и прикладных программ используются понятия пиковой, реальной и удельной производительности.

Под пиковой производительностью понимают теоретически возможное число одновременно выполняемых операций в вычислительной системе, что характеризует конструктивные и технологические потенциальные возможности вычислительной системы по обработке данных. Пиковая производительность является верхней оценкой достижимой производительности ВС. Реальная производительность, под которой понимается число одновременно выполняемых операций в единицу времени при решении прикладной задачи, в большей степени является показателем соответствия архитектуры системы структуре решаемой задачи. Удельная производительность определяется как отношение реальной производительности прикладной задачи к занимаемому аппаратному ресурсу и характеризует эффективность реализации прикладной задачи на ВС.

Для кластерных ВС высокие показатели реальной производительности достигаются, как правило, при решении слабосвязанных задач, у которых число информационных обменов между фрагментами задачи при ее решении сравнительно невелико. При решении сильносвязанных задач с интенсивными информационными обменами производительность кластерных ВС не превышает 10% от пиковой производительности.

Такой существенный разрыв между декларируемой пиковой и реальной производительностью кластерных ВС обуславливает применение реконфигурируемых вычислительных систем (РВС) для решения сильносвязанных задач. Для ряда предметных областей РВС демонстрируют высокие показатели реальной производительности при решении задач, составляющие не менее 60% от пиковой производительности системы. Для достижения таких высоких показателей производительности системы эффективность реализации прикладных задач на РВС является основополагающим фактором.

Как правило, причиной снижения производительности РВС являются неэффективные параллельные программы. Под неэффективной параллельной программой на языке высокого уровня COLAMO будем понимать программу, содержащую неэффективно используемые конструкции языка и/или неэффективно организованные потоки данных, приводящие к снижению степени распараллеливания параллельной программы или дополнительным аппаратным затратам.

Для достижения высокой реальной и удельной производительности РВС предлагается методика, направленная на автоматическое выявление неэффективного использования конструкций языка и выдачи рекомендаций по их устранению, состоящая из следующих этапов:

1. Поиск несбалансированных вычислительных структур фрагментов параллельной программы;
2. Поиск обратных связей и определение их параметров;
3. Поиск косвенной адресации при обращении к элементам массивов;
4. Поиск условных операторов и определение эффективности их реализации;
5. Поиск неэффективного использования вычислительных выражений;
6. Поиск неэффективной организации доступа к элементам потоковых массивов.

На каждом этапе выполнения методики реализующим ее программным компонентом выдается предупреждение о неэффективном использовании конструкций языка и рекомендации по их устранению.

Несбалансированность вычислительной структуры [4], когда темп обработки входной информации (на всех входах структуры) отличается от темпа выдачи информации, является одной из основных причин снижения степени распараллеливания программы.

Степень распараллеливания параллельной программы в языке COLAMO [1,2] определяется неявным образом при описании типа доступа к элементам массивов: параллельный (определяется ключевым словом Vector) и последовательный (определяется ключевым словом Stream) [3]. Тип доступа Vector указывает на необходимость размещения элементов массива в разных каналах памяти, доступ к которым может выполняться параллельно, а тип Stream указывает на то, что элементы массива располагаются в одном канале памяти, доступ к которым выполняется последовательно.

Снижение степени распараллеливания программы наиболее часто встречается при использовании в вычислительном фрагменте массивов с различными типами доступа, для которых при синтезе вычислительной структуры прикладной программы транслятором языка высокого уровня COLAMO осуществляется согласование скорости обработки потоков данных, что в случае несбалансированной вычислительной структуры, описанной программистом, приводит к снижению скорости обработки до скорости самого

медленного потока и непродуктивным затратам дополнительных аппаратных ресурсов на коммутацию информационных потоков данных и смещение их относительно друг друга.

Рассмотрим пример сочетания различных типов переменных в едином вычислительном контуре.

В программе на рис. 1 показано присвоение потоковому массиву С элементов векторного массива А, что приводит к несбалансированной вычислительной структуре и требует согласования потоков данных. Для корректной организации потоков данных в несбалансированной вычислительной структуре (рис.1) используются блок согласования (мультиплексор МХ) и блоки временных задержек (d^i – блок задержки на *i* отсчетов). Такая запись присваивания векторного массива к потоковому массиву приводит к синтезу несбалансированной структуры, падению реальной и удельной производительности вычислительной системы за счет снижения скорости обработки потоков данных и использования дополнительного оборудования.

```
Var a : Array Integer [10:Vector] Mem;
Var c : Array Integer [10:Stream] Mem;
Var I : Number;
Const k = 9;
Cadr summa;
    For i:=0 To k Do
        c[i] :=a[i];
Endcadr;
```

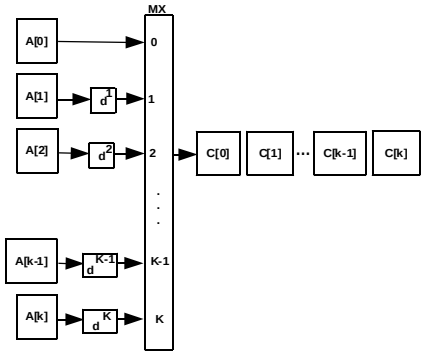

Рис. 1. Программа и эквивалентная ей вычислительная структура со степенью распараллеливания 10

Для рассматриваемого примера программисту необходимо изменить типы доступа к элементам используемых массивов параллельной программы или ввести дополнительные коммутационные переменные для организации единой степени распараллеливания для всех массивов в пределах оператора.

Несбалансированная вычислительная структура может синтезироваться также и в том случае, когда входящие в единый фрагмент параллельной программы вычислительные операторы по отдельности являются сбалансированными, но между собой не согласованы (одним из наиболее наглядных примеров является полный условный оператор, содержащий логическое выражение в условии, и вычислительные операторы в прямой и альтернативной ветвях, поскольку условные операторы более эффективно реализуются параллельно в виде множества независимых процессов).

Несбалансированность вычислительных структур прикладной параллельной программы может быть вызвана неэффективной организацией вычислений, большинство проявлений которой связано с организацией рекурсивных вычислений и неэффективным использованием распределенной памяти РВС.

Рекурсивные вычисления, под которыми понимается вычисление одного элемента массива на основе значения предыдущего (ранее вычисленного) элемента этого же массива, при структурной реализации представляют собой замкнутые вычислительные контуры (циклы) с отличным от других фрагментов темпом подачи данных на вход контура – скважностью. Скважность представляет собой натуральное число и является отношением количества тактов работы вычислительной структуры к количеству данных, которые можно подать на вход вычислительной структуры за это количество тактов, т.е. характеризует временной интервал в тактах, после которого можно подать на вход вычислительной структуры следующее данное. Для линейных вычислительных операторов скважность равна 1, при наличии в вычислительном контуре петель обратной связи скважность, как правило, увеличивается, а темп подачи, соответственно, снижается, что приводит к снижению реальной производительности. Скважность можно определить как ближайшее большее целое число отношения общей латентности вычислительного контура к значению временной задержки в цепи обратной связи.

Если в вычислительной структуре фрагмента параллельной программы существует несколько циклов, то для вычисления скважности *S* необходимо выполнить следующие действия:

- выполнить обход информационного графа и найти все циклы;
- для каждого цикла рассчитать его скважность S_i;
- определить общую скважность *S* для всех найденных циклов в информационном графе как максимум среди всех найденных значений S_i.

Если вычисленная скважность *S* не равна «1», то темп подачи данных при синтезе вычислительной структуры будет снижен в *S* раз. Рассмотрим пример программы, содержащей фрагмент с обратной связью, и ее вычислительную структуру на рис. 2.

```
Var a,c,d,e : array Integer [100 : Stream] Mem;
Var b : array Integer [100 : Stream] Com;
Var i, n : Number;
Define n=70;
Cadr Example;
```

```
For i := 30 To n Do
Begin
b[i] := b[i - 10] + a[i] - c[i];
e[i] := b[i] + d[i];
End;
Endcadr;
```

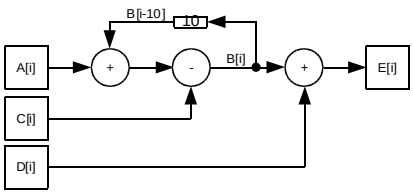

Рис. 2. Программа и эквивалентная ей вычислительная структура с обратной связью

Общая латентность кольца в вычислительной структуре на рис. 2 складывается из латентности сумматора (равна 22), вычитателя (равна 21) и элемента задержки (латентность равна 10) и равна 53. Значение временной задержки между элементами, образующими обратную связь N, определяется временной задержкой, установленной для организации обращения к элементам b[i] и b[i-10], и равно 10. Следовательно, скважность в вычислительной структуре равна 6.

Признаком неэффективного использования распределенной памяти РВС также является обращение к ней по произвольному адресу (косвенная адресация), которое используется в прикладных задачах для доступа к массивам данных или табличным значениям параметров задачи. Признаками использования косвенной адресации при обращении к элементам массива может быть использование массивов, функций или индексных переменных.Использование косвенной адресации в одном из каналов памяти ведет к снижению темпа подачи данных для этого канала памяти, что замедляет темп подачи данных для всех каналов памяти независимого от того, какой тип адресации для них используется.

На рис. 3 показаны параллельная программа с использованием косвенной адресации и эквивалентная ей вычислительная структура.

```
Var a,b,c,d : Array Real [10 : Stream] Mem;
Var i : Number;
Cadr ExampleIndexConst;
  For i := 0 to 9 do
    Begin
      a[d[i]] := b[i] + c[i];
    End;
EndCadr;
```

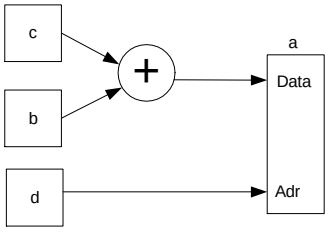

Рис. 3. Программа с использованием косвенной адресации и эквивалентная ей вычислительная структура

Обращение к переменной A является примером ярко выраженной косвенной адресации, так как адрес записи данных переменной A определяется значением элемента массива d[i].

Использование нескольких индексных переменных при обращении к элементам массива, показанное на рис. 4, также расценивается как косвенная адресация, т.к. использование нескольких индексных переменных не позволяет однозначно определить шаг изменения адреса памяти на каждой итерации циклов, поскольку он, в общем случае, не является величиной постоянной, хотя и обладает некоторой закономерностью изменений.

```
Var a,b,c,d : Array Real [100 : Stream] Mem;
Var i, j : Number;
Cadr ExampleTwoIndex;
  For i := 2 to 9 do
    For j := 5 to 90 do
      Begin
        a[i * j] := b[j] + c[j - 1];
      End;
EndCadr;
```

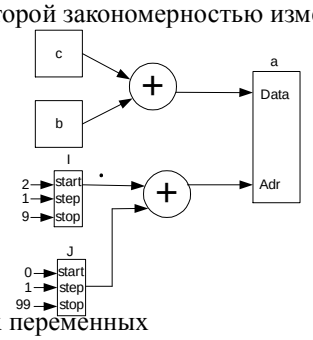

Рис. 4. Использование нескольких индексных переменных

Автоматическое выявление косвенной адресации при обращении к элементам массивов в параллельной программе при трансляции позволяет выдать программисту предупреждения об использовании косвенной адресации при обращении к элементам массивов.

Одной из важнейших конструкций языка COLAMO неэффективное использование которой может привести к снижению реальной и удельной производительности вычислительной системы является условный оператор. Синтезируемая для условного оператора вычислительная структура зависит от способа хранения переменных, используемых в условном выражении. Рассмотрим полный условный оператор, содержащий как прямую (Then), так и альтернативную (Else) ветви, каждая из которых содержит один оператор присваивания, информационный граф структурной реализации условного оператора будет выглядеть следующим образом (рис. 5)

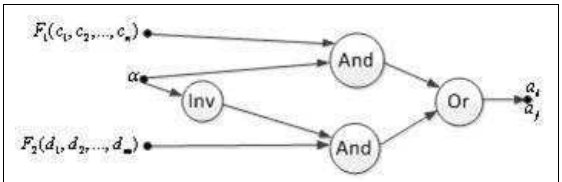

Рис. 5. Информационный граф структурной реализации условного оператора

Для реализации условного оператора требуются два блока логического умножения (And), один блок инверсии (Inv) и один блок логического сложения (Or).

Наиболее часто при структурной реализации условного оператора неэффективные затраты оборудования встречаются в следующих случаях:

- если в одной из веток коммутационной переменной присваивается значение, равное «0»;
- если условный оператор является неполным (т.е. отсутствует альтернативная ветка);
- если в ветках условного оператора используются разные коммутационные переменные в качестве результата присваивания.

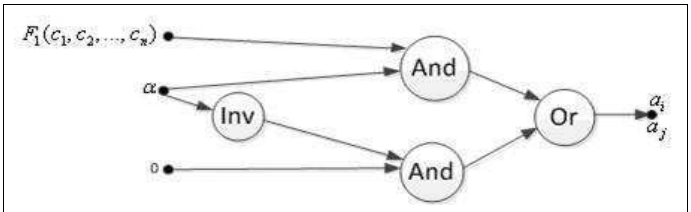

Рис. 6. Информационный граф условного оператора

Рассмотрим первый случай, информационный граф которого представлен на рис. 6.

Из информационного графа (рис. 6) видно, что в случае невыполнения условия α коммутационной переменной a_j будет присвоено значение «0». Однако, согласно особенностям коммутационной переменной, если в нее не выполняется запись, ее значение всегда будет иметь значение «0» в этом числе и в случае невыполнения условия α. Поэтому блок инверсии и блок логического умножения можно не использовать, а просто подать константу «0» на выход.

Удаление альтернативной ветки Else из программы на рис. 6 не приведет к сокращению аппаратных затрат, поскольку структурная схема, реализующая условный оператор, всегда включает в себя реализацию обеих веток условного оператора и аппаратные затраты на реализацию обеих веток (рис. 7).

При отсутствии альтернативной ветки Else на вход блока логического суммирования будет подаваться значение «0», что приводит к нерациональному использованию оборудования (блок инверсии и блок логического умножения), как и в первом случае ().

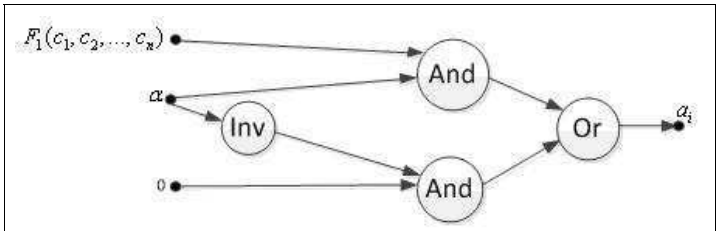

Рис. 7. Неполный условный оператор

Рассмотрим третий вариант использования условного оператора на примере программы на рис. 8.

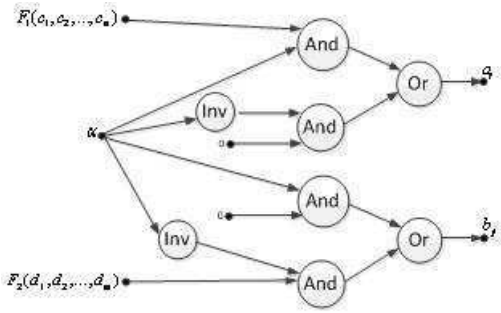

Рис. 8. Программа и эквивалентная ей вычислительная структура

Использование в ветках условного оператора разных переменных в результате присваивания приводит к структурной реализации условного оператора для обеих коммутационных переменных, расположенных в ветках условного оператора. Из вычислительной структуры на рис. 8 видно, что для вычисления массива *a* на вход блока логического суммирования подается значение «0», что соответствует отсутствию альтернативной ветки *Else*. Для вычисления массива *b* на вход блока логического суммирования также подается значение «0», т.к. в операторе условного оператора в прямой ветке отсутствует присваивание переменной *b*.

Для эффективной структурной реализации программ в случаях, показанных на рис. 6-8, рекомендуется условный оператор привести к логическому умножению, существенно сократив аппаратные затраты. Для организации выбора значения коммутационной переменной по условию программа должна иметь вид, показанный на рис. 9.

$$a_i = F_1(c_1, c_2, \ldots, c_n) \,\&\, \alpha$$

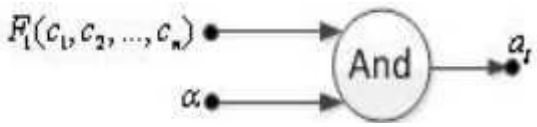

Рис. 9. Программа и эквивалентная ей вычислительная структура

Эффективность параллельной программы зависит не только от грамотного использования условных операторов, но и от эффективного использования вычислительных операторов. Для вычислительных выражений наиболее распространенным примером неэффективной реализации является повторная реализация вычислительного выражения (программа 1.4)

```
Var a,b,c,d,e,r : Array Integer [100 : Stream] Mem;
Var i, j, n : Number ;
Define n=99 ;
Cadr EqualPart;
  For i := 2 to n do
    Begin
      a[i] := F1(b[i], c[i]) - d[i] + e[i];
      r[i] := F1(b[i], c[i]) + (d[i] * e[i]);
    End;
Endcadr;                                                    (4)
```

Согласно принципам языка COLAMO в вычислительной структуре (рис. 10) программы (4) будет реализовано две функции F_1(b[i], c[i]), несмотря на то, что они одинаковы, что приводит к необоснованным затратам оборудования.

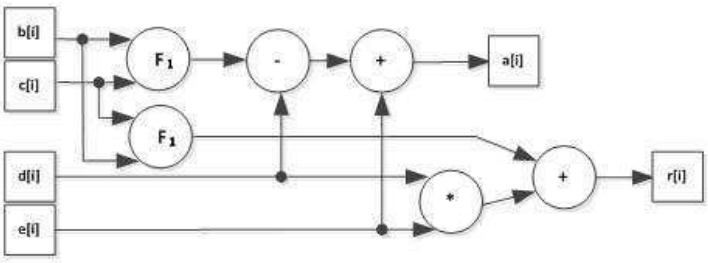

Рис. 10. Вычислительная структура программы (1.4)

Для эффективной реализации программы (4) необходимо использовать коммутационную переменную, которой будет присвоено выражение F_1(b[i], c[i]), а сами выражения в операторах должны быть заменены на коммутационную переменную. После выполнения данных преобразований программа (4) будет выглядеть следующим образом:

```
Var a,b,c,d,e,r : Array Integer [100 : Stream] Mem;
Var t : Integer Com;
Var i, j, n : Number ;
Define n=99 ;
Cadr EqualPart;
  For i := 2 to n do
    Begin
      t := F1(b[i], c[i]);
      a[i] := t - d[i] + e[i];
      r[i] := t + (d[i] * e[i]);
    End;
```

Endcadr; (5)

Поскольку в операторах программы (5) используется только один вызов функции F_1, то и в вычислительной структуре (рис. 11) будет реализована только одна функция, что позволяет сократить

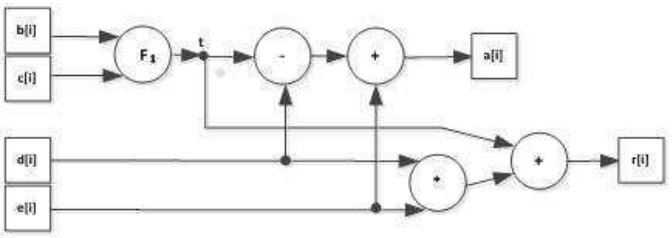

Рис. 11. Вычислительная структура программы (1.5)

аппаратный ресурс, необходимый для реализации задачи.

Таким образом, если в операторах параллельной программы найдены общие подвыражения, то выдается рекомендация о необходимости использования коммутационных переменных в качестве промежуточных переменных между операторами и выделенным выражениями.

Зачастую в операторах параллельной программы используется параллельный доступ к разным элементам одного и того же потокового массива. Однако элементы потокового массива могут обрабатываться только последовательно, поэтому для организации одновременного доступа к разным элементам массива необходимо использовать смещение на уровне информационных потоков данных (рис. 12).

```
Var a,b : Array Real [100 : Stream] Mem;
Var c : Array Real [100 : Stream] Com;
Var I : Number;
Cadr ExampleF;
  For i := 10 to 70 do
   Begin
      c[i] := b[i];
      a[i] := F (c[i], c[i-5], c[i-8]);
Endcadr;
```

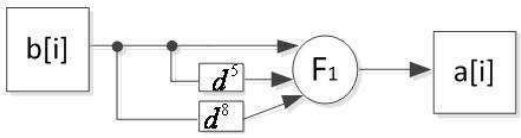

Рис. 12. Параллельная программа со смещением при адресации массивов и эквивалентная ей вычислительная структура

В программе на рис. 12 выполняется три одновременных доступа к потоковому массиву C. Для корректной организации одновременного доступа необходимо использовать две временные задержки (смещения) d^5 и d^8. При этом суммарную величину временных задержек, необходимых для организации потоков данных, можно сократить за счет использования коммутационных массивов и, тем самым, сократить аппаратные затраты.

Использование коммутационных массивов позволяет организовать одновременный доступ к элементам массива c[i], c[i-5], c[i-8], но при этом временные задержки, необходимые для выравнивания потока данных, будут организованы не параллельно, как показано на рис. 12, а последовательно (рис. 13).

```
Const n = 99;
Var a,b : Array Real [100 : Stream] Mem;
Var c,d,e : Array Real [100 : Stream] Mem;
Var I : Number;
Cadr Example;
  For i := 0 to n do
   Begin
      c[i] := b[i];
      d[i] := c[i-5];
      e[i] := d[i-3];
      a[i] := F (c[i], d[i], e[i]);
Endcadr;
```

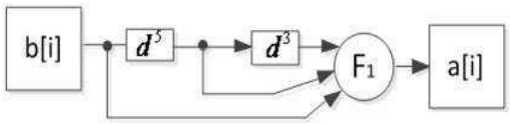

Рис. 13. Программа (рис. 12) с использованием коммутационных массивов и эквивалентная ей вычислительная структура

Количество коммутационных массивов, реализующих смещение, определяется величиной *n-1* , где *n* – число одновременных обращений к массиву с (рис. 12), а величина временной задержки между элементами коммутационных массивов рассчитывается на основании смещений между элементами массива *c*. При этом суммарная величина полученных временных задержек не должна превышать значения максимальной временной задержки.

Из вычислительных структур на рис. 12 и рис. 13 видно, что суммарная величина временной задержки между элементами на рис. 12 складывается из величин d^5 и d^8 и равна 13, а на рис. 13 суммарная величина временной задержки складывается из величин d^5 и d^3и равна 8.

Таким образом, при выявлении параллельной организации доступа к элементам потокового массива (рис. 12) необходимо перейти к последовательной организации доступа к элементам массива за счет использования промежуточных коммутационных массивов.

Рассмотренная методика, направленная на выявление неэффективного использования конструкций языка и выдачи рекомендаций по их устранению позволит создавать эффективные параллельные прикладные программы и как следствие, повысит реальную и удельную производительность реконфигурируемых вычислительных систем.

Исследования выполнены при поддержке научно-исследовательских работ по договору № 1301-Ю от 01.04.2011 г. во исполнение государственного контракта № 07.514.12.4001 от 24.12.2010 г.Министерства образования и науки РФ.

ЛИТЕРАТУРА:

1. И.И. Левин Реконфигурируемые вычислительные системы с открытой масштабируемой архитектурой // Труды Пятой Международной конференции «Параллельные вычисления и задачи управления» PACO'2010. - М.: Учреждение Российской академии наук Институт проблем управления им. В.А. Трапезникова РАН, 2010. - С.83-95.
2. А.В. Каляев, И.И. Левин Модульно-наращиваемые многопроцессорные системы со структурно-процедурной организацией вычислений. - М.: Янус-К, 2003. – 380 с.
3. И.А. Каляев, И.И. Левин, Е.А. Семерников, В.И. Шмойлов Реконфигурируемые мультиконвейерные вычислительные структуры /Изд. 2-е, перераб. и доп. / Под общ. ред. И.А. Каляева. - Ростов-на-Дону: Изд-во ЮНЦ РАН, 2009. – 344 с.
4. И.А. Каляев, И.И. Левин Семейство реконфигурируемых вычислительных системы с высокой реальной производительностью // Труды международной научной конференции «Параллельные вычислительные технологии» (ПАВТ'2009). – Нижний Новгород: электронное издание НГУ имени Н.И. Лобачевского, 2009. – С.186-196.
5. Н.Н. Дмитренко, И.А. Каляев, И.И. Левин, Е.А. Семерников Развитие аппаратной платформы реконфигурируемых вычислительных систем // Труды Международной суперкомпьютерной конференции «Научный сервис в сети Интернет: суперкомпьютерные центры и задачи. – М.: Изд-во МГУ, 2010. – С. 315-320.
6. Коваленко, В.Б. Организация многоуровневого программирования реконфигурируемых вычислительных систем [Текст] / Е.А. Семерников, В.Б. Коваленко // Вестник компьютерных и информационных технологий. – М.: Машиностроение, 2011. - № 9. – С. 3-10