https://doi.org/10.25559/SITITO.020.202403.563-572
УДК 004.42, 004.051
Оригинальная статья

Векторизация циклов с условными операциями с помощью комбинирования векторных масок

А. А. Рыбаков[1,2]
[1] ФГБУ «Национальный исследовательский центр «Курчатовский институт»», г. Москва, Российская Федерация
Адрес: 123182, Российская Федерация, г. Москва, пл. Академика Курчатова, д. 1
[2] ФГУ «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Москва, Российская Федерация
Адрес: 117218, Российская Федерация, г. Москва, Нахимовский пр., д. 36, корп. 1
rybakov.aax@gmail.com

Аннотация
Статья посвящена проблеме повышения эффективности векторизации вычислений на вещественных числах. При выполнении вычислений несколько одинаковых скалярных операций могут быть объединены в единую векторную команду, существенно повышая скорость выполнения программы. Данная оптимизация является критически важной для расчетных задач суперкомпьютерного моделирования. Основным объектом, на который нацелена векторизация вычислений, является цикл с независимыми итерациями. При относительно простом виде тела рассматриваемого цикла проблем с векторизацией, как правило, не возникает. При появлении в теле цикла сложного управления, вложенных циклов и вызовов функций оптимизирующий компилятор зачастую не справляется с векторизацией. Однако особенности набора векторных инструкций AVX-512 с поддержкой выборочной обработки элементов данных векторов позволяют векторизовать циклы с телом практически произвольного вида. В настоящей статье рассматривается подход к векторизации цикла, содержащего условия. Подход основан на оптимизации слияния путей исполнения программы под соответствующими предикатами. Векторизованный предикат представляет собой маску обработки элементов вектора. Такие маски используются в векторных инструкциях AVX-512. При векторизации циклов, тело которых содержит сложное управление, основной проблемой является низкая плотность масок векторизованного кода, что приводит к снижению производительности. В статье рассмотрены методы, позволяющие повысить плотность векторных масок и эффективность выполнения векторного кода. Разработанные методы апробированы на программном контексте газодинамического решателя. Данные по эффективности векторизации были получены в режиме эмуляции векторных инструкций и на реальной машине (микропроцессор Intel Xeon Phi Knights Landing). После применения оптимизаций векторного кода были достигнуты показатели эффективности векторизации до 0.75 в режиме эмуляции и до 0.47 на реальной машине.

Ключевые слова: векторизация, AVX-512, плоский цикл, векторизация условий, векторная маска

Финансирование: работа выполнена в рамках государственного задания НИЦ «Курчатовский институт» по теме FNEF–2024-0016.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Рыбаков А. А. Векторизация циклов с условными операциями с помощью комбинирования векторных масок // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 3. С. 563-572. https://doi.org/10.25559/SITITO.020.202403.563-572

Original article

Vectorization of Loops with Conditional Operations by Combining Vector Masks

A. A. Rybakov[a,b]
[a] National Research center «Kurchatov Institute for Theoretical and Experimental physics", Moscow, Russian Federation
Address: 1 Akademicianika Kurchatova Sq., Moscow 123182, Russian Federation
[b] Scientific Research Institute for System Analysis of the Russian Academy of Sciences, Moscow, Russian Federation
Address: 36, build. 1, Nakhimov Ave., Moscow 117218, Russian Federation
rybakov.aax@gmail.com

Abstract
The article is devoted to the problem of increasing the efficiency of vectorization for calculations on real numbers. When performing calculations, several similar scalar operations can be combined into a single vector command, significantly increasing the speed of program execution. This optimization is crucial for computational tasks of supercomputer modeling. The main object that the vectorization of calculations is aimed at is a loop with independent iterations. With a relatively simple form of the body of the considered loop, problems with vectorization, as a rule, do not arise. When complex controls, nested loops, and function calls appear in the body of the loop, the optimizing compiler often fails to cope with vectorization. However, the features of the AVX-512 vector instruction set with support of selective processing of vector data elements make it possible to vectorize loops with a body of almost arbitrary structure. This article discusses an approach to vectorization of a loop which contains conditions. The approach is based on merging of program execution paths under the appropriate predicates. A vectorized predicate is a mask for processing vector elements. Such masks are used in AVX-512 vector instructions. When vectorizing loops whose body contains complex controls, the main problem is the low density of masks of the vectorized code, which leads to decrease in performance. The article discusses methods to increase the density of vector masks and the efficiency of vector code execution. The developed methods have been tested on the progamm context of a gas-dynamic solver. Data on vectorization efficiency were obtained in vector instruction emulation mode and on a real machine (Intel Xeon Phi Knights Landing microprocessor). After applying vector code optimizations, vectorization efficiency indicators were achieved up to 0.75 in emulation mode and up to 0.47 on a real machine.

Keywords: vectorization, AVX-512, flat loop, conditions vectorization, vector mask

Funding: The work was performed according to the Government research assignment for SIC "Kurchatov Institute", project FNEF–2024-0016.

Conflict of interests: The author declares no conflict of interests.

For citation: Rybakov A.A. Vectorization of Loops with Conditional Operations by Combining Vector Masks. *Modern Information Technologies and IT-Education.* 2024;20(3):563-572. https://doi.org/10.25559/SITITO.020.202403.563-572

Введение

Повышение эффективности приложений суперкомпьютерного моделирования является критически важной задачей. Выполнение одного расчета по моделированию физического процесса может занимать несколько часов или дней. В условиях экспоненциального повышения количества различных расчетов, которые необходимо выполнить для проведения исследований или проектирования новых изделий, время работы суперкомпьютера становится крайне дефицитным ресурсом. В настоящее время для численного решения задач газовой динамики уже возникают потребности в использовании суперкомпьютерных ресурсов экзафлопсного диапазона производительности [1]. Можно отметить, что как сама задача, так и вычислительное поле суперкомпьютера зачастую имеют гетерогенную природу. По отношению к задаче это выражено в изменяющихся адаптивных сетках [2], блочно-структурированных сетках [3], индуцированных сетках [4], перекрывающихся или химерных сетках [5] и других особенностям. По отношению к вычислительному полю это выражено в совокупности отличающихся двух от друга вычислительных узлов, микропроцессоров и графических карт [6]. В таких условиях для организации счета требуются специальные подходы, учитывающие балансировку вычислительной нагрузки [7]. При проведении суперкомпьютерных расчетов широко используется распараллеливание вычислений между узлами суперкомпьютерного кластера с помощью MPI [8, 9], распараллеливание внутри одного узла с помощью OpenMP [10, 11] и распараллеливание на уровне команд с помощью векторизации.

Векторизация вычислений это низкоуровневая оптимизация, правильное применение которой способно кратно повысить производительность наиболее горячих участков программного кода и сократить энергопотребление [12]. В мире постоянно проводятся исследования, направленные на повышение производительности векторного кода. В работе [13] путем векторизации безусловных операций с помощью функций-интринсиков было продемонстрировано повышение производительности газодинамического решателя на 200% при исполнении на микропроцессорах Intel Xeon Phi KNL и Intel Xeon Scalable. В работе [14] описывается сравнение реализации римановских решателей в применении к теории мелкой воды, одним из результатов работы является ускорение решателя с помощью инструкций AVX-512 в 16.7 раз при работе с вещественными числами одинарной точности. В исследовании [15] было достигнуто ускорение в 3.27 газодинамического решателя ADflow, работающего на структурированных расчетных сетках. Ускорение достигнуто путем декомпозиции сетки на вычислительные блоки, которые могут быть эффективно обработаны с точки зрения использования кэш и применения векторных инструкций AVX-512. В работе [16] рассмотрена реализация расчета гравитационного взаимодействия между N телами, из результатов видно, что использование набора инструкций AVX-512 позволило ускорить работу приложения в 2 и более раз. Также использование векторизации позволяет ускорить работу не только газодинамических решателей. В работе [17] описано успешное применение векторных инструкций в задаче поиска сходных участков в белковых последовательностях. Работы [18, 19] посвящены ускорению алгоритмов, связанных с шифрованием. Однако, в данной работе мы будем больше фокусироваться на программном контексте, работающим с вещественными числами и наиболее характерно представленном в реализации решателей для расчета физических процессов.

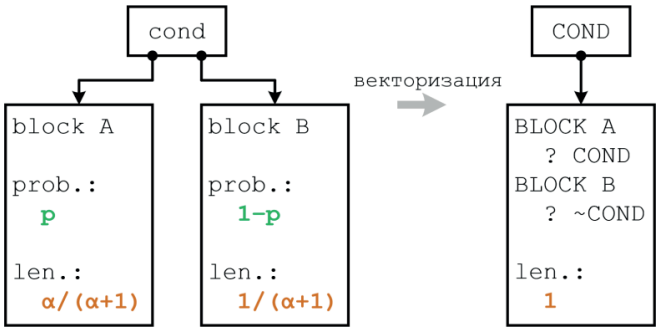
В качестве объекта исследования будем рассматривать тип программного контекста, который наилучшим образом подходит для векторизации – плоский цикл. Под плоским циклом будем подразумевать обычный цикл for (int i = 0; i < n; ++i) который удовлетворяет следующим требованиям. Во-первых, все итерации цикла являются независимыми друг от друга, что позволяет выполнять их в любом порядке. Во-вторых, на i-ой итерации цикла все обращения в массивы данных имеют вид a[i]. И, в-третьих, все используемые массивы данных не пересекаются и выровнены в памяти для использования 512-битного чтения их блоков. При выполнении этих простых требований использование инструкций AVX-512 позволяет векторизовать плоский цикл с телом практически произвольного вида (включая сложное управление, вложенные циклы и вызовы функций) [20]. В данной работе будем рассматривать подходы к повышению эффективности векторизации плоского цикла, содержащего условия. Векторизация плоского цикла с шириной векторизации w будет представлять собой объединение w последовательных итераций цикла в одну широкую итерацию, внутри которой все скалярные операции заменяются на векторные аналоги, а предикаты заменяются на векторные маски. В дальнейшем при рассмотрении будем опускать сам плоский цикл и рассматривать только его тело.

Слияние путей исполнения по условию

Универсальным способом векторизации программного кода, содержащего условия, является слияние всех ветвей исполнения под соответствующими предикатами. Рассмотрим это действие на примере простого условия cond, по результату которого выполняется переход на один из блоков block A и block B. Пусть известны вероятности перехода на эти блоки – они равны p и $1-p$ соответственно. Длины рассматриваемых блоков подберем таким образом, чтобы в сумме они давали единицу, а отношение их длин задавалось параметром α. Таким образом, длины блоков будут равны $\frac{\alpha}{\alpha+1}$ и $\frac{1}{\alpha+1}$ соответственно (см. рис. 1). При этом условимся считать, что длина блока и время его исполнения это по сути одно и то же (то есть, время исполнения блока исчисляется количеством содержащихся в нем операций).

Согласно представленной схеме программного контекста математическое ожидание времени исполнения рассматриваемых блоков в зависимости от условия, будет равно $T_1 = \frac{p\alpha}{\alpha+1} + \frac{1-p}{\alpha+1} = p\left(\frac{\alpha-1}{\alpha+1}\right) + \left(\frac{1}{\alpha+1}\right)$. Время выполнения w таких участков кода в невекторизованном виде будет равно wT_1. При векторизации кода необходимо избавиться от операций перехода, вместо этого все операции блоков block A и block B должны быть поставлены под предикаты cond и ~cond соответственно. Далее выполняется объединение w участков кода, при котором скалярные операции под предикатами cond/~cond заменяются на векторные аналоги, выполняющиеся с использованием векторных масок COND/~COND. Так как длины блоков выбирались таким образом, чтобы в сумме они давали единицу, то время исполнения векторной версии кода

в точности равно $T_w = 1$. Таким образом, эффективность векторизации рассмотренного фрагмента кода равна $e = \frac{T_1}{T_w} = p\left(\frac{\alpha - 1}{\alpha + 1}\right) + \left(\frac{1}{\alpha + 1}\right)$. На рис. 2 представлены графики зависимостей эффективности векторизации для разных значений параметра α.

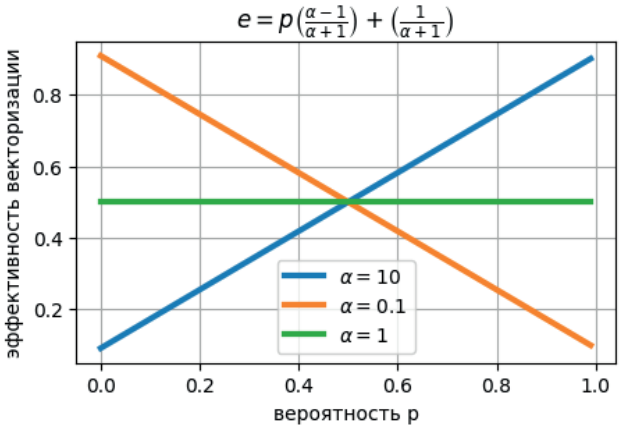


Р и с. 1. Схема векторизации участка программного кода, состоящего из одного условия и двух блоков, переход на которые осуществляется в соответствии с этим условием

F i g. 1. A vectorization scheme for a section of program code consisting of one condition and two blocks, the transition to which is carried out in accordance with this condition

Источник: здесь и далее в статье все таблицы и рисунки составлены автором.

Source: Hereinafter in this article all tables and figures were made by the author.



Р и с. 2. Графики зависимостей эффективности векторизации от вероятности перехода на block A при значениях отношения длин блоков block A и block B, равных 10.0, 0.1, 1.0 и при использовании простого слияния ветвей исполнения

F i g. 2. The dependence graphs of vectorization efficiency on the probability of transition to block A for values of the ratio of the lengths of blocks block A and block B equal to 10.0, 0.1, 1.0 and when using simple merging of execution branches

Из рис. 2 видно, что при $\alpha = 1$ (то есть при одинаковых длинах блоков block A и block B) эффективность векторизации постоянна и равна 0.5. В тех же случаях, когда длины блоков отличаются, эффективность векторизации возрастает, если вероятность перехода на более длинный блок выше, чем на более короткий блок. В любом случае, можно констатировать, что такой подход прямого слияния ветвей исполнения под соответствующими предикатами в единый линейный участок является крайне неэффективным. При возрастании количества условий эффективность векторизации таким способом падает экспоненциально. Это связано с появлением в программном коде большого количества векторных инструкций с практически пустыми масками. Для повышения эффективности векторизации контекста с условиями требуется рассмотрение других подходов, позволяющих повысить плотность масок исполняемых векторных команд.

В качестве примера, на котором мы будем проводить анализ методов повышения эффективности векторизации кода с условиями, возьмем одну из функций реализации газодинамического решателя – функцию prefun[1]. Для удобства будем пользоваться реализацией данной функции на языке программирования C, как это представлено на листинге 1 (исходный код функции доступен в открытом репозитории[2]).

```
void scase_prefun_1(float& f,
                    float& fd,
                    float p,
                    float dk,
                    float pk,
                    float ck)
{
    if (p <= pk)
    {
        // Rarefaction wave.

        float prat = p / pk;

        f = riemann::sg4 * ck * (pow(prat, rie-
mann::sg1) - 1.0f);
        fd = (1.0f / (dk * ck)) * pow(prat, -rie-
mann::sg2);
    }
    else
    {
        // Shock wave.

        float ak = riemann::sg5 / dk;
        float bk = riemann::sg6 * pk;
        float qrt = sqrt(ak / (bk + p));

        f = (p - pk) * qrt;
        fd = (1.0f - 0.5f * (p - pk) / (bk + p))
* qrt;
    }
}
```

Л и с т и н г 1. Код функции prefun, обрабатывающей один набор скалярных данных

L i s t i n g 1. The prefun function code, which processes a single set of scalar data

[1] Toro E. F. The Riemann Problem for the Euler Equations // Riemann Solvers and Numerical Methods for Fluid Dynamics. Berlin, Heidelberg : Springer, 2009. P. 155. https://doi.org/10.1007/b79761_4

[2] Rybakov A. flatvec [Electronic resource] // Github, 2024. URL: https://github.com/r-aax/flatvec (дата обращения: 10.06.2024). Исходный файл fvcases/cases.cpp

```
void vcase_prefun_1(__m512& f,
                    __m512& fd,
                    __m512& p,
                    __m512& dk,
                    __m512& pk,
                    __m512& ck,
                    __mmask16 m)
{
    // Conditions.
    __mmask16 cond = _mm512_kand(_mm512_cmple_ps_
mask(p, pk), m);
    __mmask16  ncond  =  _mm512_kand(_mm512_
knot(cond), m);

    // The first branch.
    {
        __m512 prat = _mm512_mask_div_ps(zero,
cond, p, pk);

        f = _mm512_mask_mul_ps(f, cond,
                  _mm512_mask_mul_ps(zero, cond,
riemann::g4, ck),
                _mm512_mask_sub_ps(zero, cond,
                  _mm512_mask_pow_ps(zero, cond,
prat, riemann::g1),
                    one));

        fd = _mm512_mask_mul_ps(fd, cond,
                _mm512_mask_div_ps(zero, cond,
                    one,
                  _mm512_mask_mul_ps(zero, cond,
dk, ck)),
                _mm512_mask_pow_ps(zero, cond,
                    prat,
                  _mm512_mask_sub_ps(zero, cond,
zero, riemann::g2)));
    }

    // The second branch.
    {
        __m512 ak = _mm512_mask_div_ps(zero, ncond,
riemann::g5, dk);
        __m512 bk = _mm512_mask_mul_ps(zero, ncond,
riemann::g6, pk);
        __m512 qrt = _mm512_mask_sqrt_ps(zero,
ncond,
                        _mm512_mask_div_ps(zero,
ncond,
                            ak,
                          _mm512_mask_add_ps(ze-
ro, ncond, bk, p)));

        f = _mm512_mask_mul_ps(f, ncond,
                _mm512_mask_sub_ps(zero, ncond,
p, pk),
              qrt);

        fd = _mm512_mask_mul_ps(fd, ncond,
                _mm512_mask_sub_ps(zero, ncond,
                    one,
                  _mm512_mask_mul_ps(zero, ncond,
                        half,
                        _mm512_mask_div_ps(zero,
ncond,
                          _mm512_mask_sub_ps(ze-
ro, ncond, p, pk),
                          _mm512_mask_add_ps(ze-
ro, ncond, bk, p)))),
                qrt);
    }
}
```

Л и с т и н г 2. Код функции prefun, обрабатывающей один набор векторных данных, представляющий собой объединение 16 наборов скалярных данных

L i s t i n g 2. The prefun function code, which processes one set of vector data, which is the union of 16 sets of scalar data

В листинге 1 мы видим реализацию функции prefun, обрабатывающей один набор скалярных данных. Функция принимает на входные аргументы p, dk, pk, ck и вычисляет выходные аргументы f, fd. Функция содержит одно условие в зависимости от которого выходные аргументы вычисляются с помощью той или иной последовательности операций. Все задействованные в реализации функции операции имеют векторные аналоги в наборе инструкций AVX-512 (точнее в наборе функций-интринсиков), поэтому приведенная функция может быть векторизована путем замены скалярных операций векторными аналогами и слияния ветвей исполнения под соответствующими предикатами, как это показано на листинге 2. В процессе векторизации умышленно не применялись никакие локальные оптимизации, все скалярные операции были строго заменены на векторные аналоги с сохранением порядка вычислений с точности до ассоциативности умножения. Для удобства код, относящийся к блокам block A и block B, заключен в фигурные скобки.

Сравнивая листинги 1 и 2, содержащие скалярный и векторный код, можно установить соответствие между участвующими в них объектами, как это показано в таблице 1.

Т а б л и ц а 1. **Соответствие между скалярными объектами реализации scase_prefun_1 и векторными объектами реализации vcase_prefun_1**

T a b l e 1. **Correspondence between scalar implementation objects scase_prefun_1 and vector implementation objects vcase_prefun_1**

Объект скалярной версии scase_prefun_1	**Объект векторной версии vcase_prefun_1**
скалярные аргументы float f, fd, p, dk и т.д.	векторные аргументы __m512 f, fd, p, dk и т.д.
скалярные глобальные данные riemann::sg1, riemann::sg2 и т.д.	векторные глобальные данные riemann::g1, riemann::g2
скалярные операции +, -, *, /, pow	векторные команды, заданные функциями-интринсиками _mm512_mask_add_ps, _mm512_mask_sub_ps, _mm512_mask_mul_ps, _mm512_mask_div_ps, _mm512_mask_pow_ps
скалярная операция сравнения <=	векторные операции получения масок _mm512_cmple_ps_mask, _mm512_knot, _mm512_kand

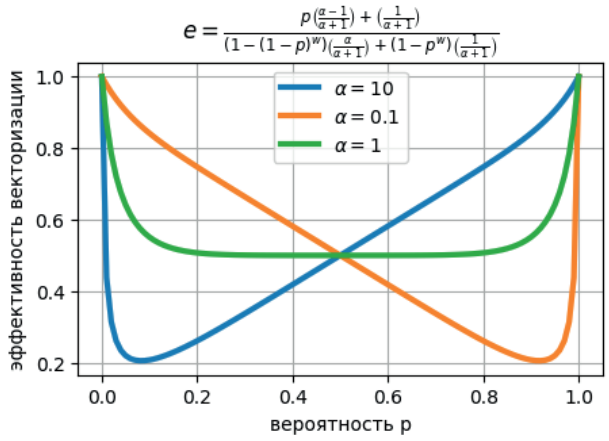
Из приведенного листинга 2 векторного кода видно, что часть команд выполняется с использованием векторной маски cond, тогда как другая часть команд использует векторную маску ncond. При этом понятно, что если маска cond окажется нулевой, то нет необходимости выполнять инструкции, использующие эту маску. То же касается маски ncond. Для того, чтобы учесть это изменение достаточно перед выполнением блока операций, относящихся к block A, проверить маску cond на пустоту, в противном случае вообще не выполнять соответствующие инструкции. Аналогично следует поступить с маской ncond. Скорректируем зависимости эффективности вектори-

зации для рассматриваемого примера с учетом этого условия. Величина T_1 остается неизменной, а вот T_w несколько изменится. Если считать, что в каждом наборе обрабатываемых данных переход на тот или иной блок является независимым событием, то вероятность пустой маски cond будет равна $(1-p)^w$, тогда как вероятность пустой маски ncond равна p^w. Тогда общая длина векторизованного кода может быть вычислена как $T_w = \left(1-(1-p)^w\right)\left(\frac{\alpha}{\alpha+1}\right)+\left(1-p^w\right)\left(\frac{1}{\alpha+1}\right)$, а эффективность векторизации примет следующий вид:

На рис. 3 представлены зависимости эффективности векторизации при разных значениям параметра α с учетом проверок

$$e = \frac{p\left(\frac{\alpha-1}{\alpha+1}\right)+\left(\frac{1}{\alpha+1}\right)}{\left(1-(1-p)^w\right)\left(\frac{\alpha}{\alpha+1}\right)+\left(1-p^w\right)\left(\frac{1}{\alpha+1}\right)}$$

масок на пустоту для ширины векторизации 16, что соответствует использованию вещественного формата данных одинарной точности.

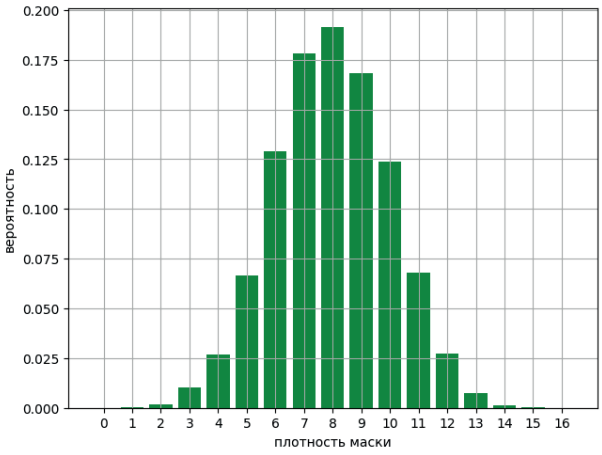


Р и с. 3. Графики зависимостей эффективности векторизации от вероятности перехода на block A при значениях отношения длин блоков block A и block B, равных 10.0, 0.1, 1.0 и при использовании слияния ветвей исполнения с проверкой масок на пустоту

F i g. 3. The dependence graphs of vectorization efficiency on the probability of transition to block A for the values of the ratio of the lengths of blocks block A and block B equal to 10.0, 0.1, 1.0 and when using branch merging with mask checking for emptiness

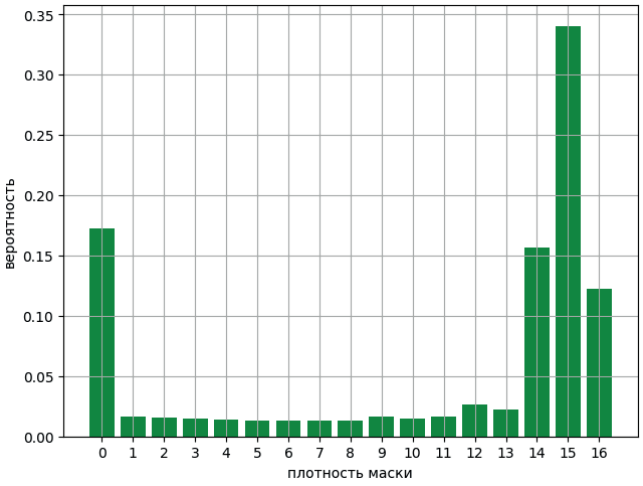
Из рис. 3 видно, что эффективность векторизации возрастает, если значение вероятности перехода на один из блоков близко к единице, однако в среднем вероятность векторизации остается невысокой. Когда мы вычисляли вероятность появления пустой маски, то принимали условное соглашение, что выполнение условий для разных наборов скалярных данных являются независимыми событиями. На самом деле это не так и существенным образом зависит от локальности размещения данных, участвующих в расчетах [21]. Рассмотрим более подробно наше условие p <= pk. Элементы данных p и pk, свои для каждой расчетной ячейки. Если речь идет о физических расчетах (а функция prefun относится к газодинамическому решателю), то значение элемента данных изменяется не слишком сильно при переходе от одной ячейки к соседней ячейке. Из этого следует, что и значение условия p <= pk при переходе от одной ячейке к соседней будет изменяться также не слишком быстро. Но условие это дискретная величина, а это значит, что часто значение условия будет сохраняться при переходе к соседней ячейке. Рассмотри теоретический краевой случай, когда во время обработки n наборов скалярных данных условие выполняется для первых np из них и не выполняется для оставшихся $n(1-p)$. Для простоты будем считать, что все числа np, $n(1-p)$ кратны ширине векторизации w. В этом случае очевидно, что во время выполнения векторной версии кода первые $\frac{np}{w}$ масок cond будут полные, а остальные $\frac{n(1-p)}{w}$ масок cond будут пустыми (с масками ncond ситуация будет обратной). В таком случае, время исполнения векторизованной версии кода с проверкой обеих масок cond и ncond на пустоту в точности совпадет со временем T_1. Таким образом, эффективность векторизации в этом теоретически идеальном случае будет ровно единица с поправкой на дополнительные операции проверки масок на пустоту.

Для рассматриваемой функции prefun были собраны расчетные данные распределения плотности маски условия p <= pk, чтобы оценить вероятность появления пустых масок cond и ncond. На рис. 4 представлено распределение плотности масок в случае независимости условий для разных наборов скалярных данных (нетрудно видеть, что распределение является нормальным). Результаты распределения плотности маски cond, собранные по настоящему профилю исполнения векторного кода на реальных данных, представлены на рис. 5.



Р и с. 4. Гистограмма распределения плотностей маски cond при условии, что все условия p <= pk для наборов скалярных данных являются независимыми

F i g. 4. Histogram of the density distribution of the cond mask, given that all terms p <= pk for scalar data sets are independent



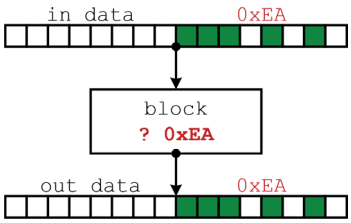
Р и с. 5. Гистограмма распределения плотностей маски cond на реальном профиле исполнения

F i g. 5. Histogram of the distribution of cond mask densities on a real performance profile

Из рис. 4 и рис. 5 видно, что распределение плотностей масок на реальных данных совершенно не похоже на распределение, вычисленное в предположении о независимости условий переходов. Можно заметить, что в реальном коде более четверти всех масок cond являются либо пустыми, либо полными (в этом случае пустой является маска ncond), а значит использование проверок масок на пустоту обосновано.

Комбинирование масок при исполнении векторного кода

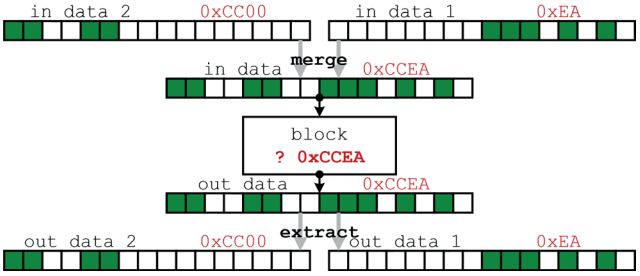
В общем случае можно считать, что в результате слияния под соответствующими предикатами ветвей исполнения внутри тела плоского цикла мы получим совокупность векторных блоков, обрабатывающихся сходим образом: загрузка входных данных in_data под маской векторного блока, выполнение вычислений block под маской блока, сохранение результатов out_data под маской блока (см. рис. 6).



Р и с. 6. Схема вычислений векторизованного блока команд с входными данными in_data, выходными данные out_data и маской исполнения 0xEA

F i g. 6. The computation schematic diagram of a vectorized command block with input data in_data, output data out_data and execution mask 0xEA

Проверка маски блока на пустоту может повысить эффективность кода, если маски часто оказываются пустыми. Однако этот никак не поможет в том случае, если в маске выставлено несколько битов. В некоторых случаях достичь повышения производительности можно путем объединения двух соседних векторных блоков. Рассмотрим простейший случай такого объединения. Если у нас есть два соседних векторных блока in_data_1 → block → out_data_1 и in_data_2 → block → out_data_2, которые должны выполняться под разными векторными масками mask_1 и mask_2, и в дополнение к этому для этих масок выполнено условие (mask_1 & mask_2) == 0x0, то вычисление этих двух соседних блоков можно объединить. Вместо последовательного выполнения двух векторных блоков можно объединить входные данные с помощью слияния векторов по условию in_data = _mm512_mask_blend_ps(mask_1, in_data_2, in_data_1), после чего выполнить тот же блок вычислений под маской mask_1 | mask_2. Ввиду отсутствия пересечения векторных масок в результирующем выходном векторе out_data будут содержаться как необходимые элементы вектора out_data_1, так и необходимые элементы вектора out_data_2. Последним действием, которое нужно выполнить является извлечение из объединенного результата out_data векторов out_data_1 и out_data_2 (см. рис. 7). В результате такого преобразования в случае отсутствия пересечения векторных масок количество вычислений рассматриваемого блока block сокращается вдвое, а плотность векторных масок внутри блока повышается. Однако вместе с этим появляются накладные расходы, связанные с проверками масок, а также операции слияния векторов до вычислений блока и выделения нужных данных после вычислений. Заметим, что эту технику можно применять для объединения трех и более соседних блоков, однако это связано с еще большим возрастанием накладных расходов.

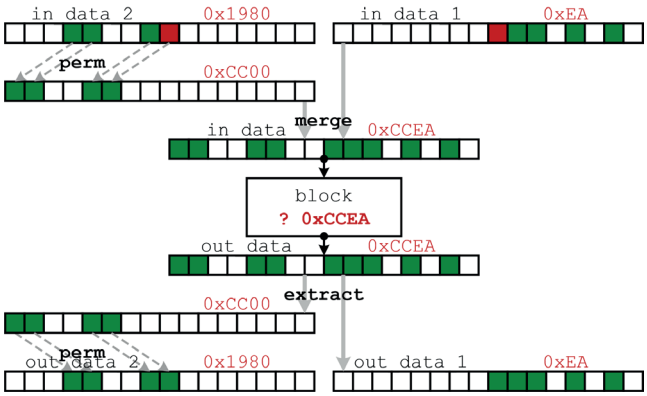


Р и с. 7. Схема вычислений с объединением двух векторизованных блоков in_data_1 → block → out_data_1, in_data_2 → block → out_data_2. Объединение допустимо, так как векторные маски 0xCC00 и 0xEA не пересекаются, объединенный блок выполняется под маской 0xCCEA

F i g. 7. Computational scheme with merging two vectorized blocks with in_data_1 → block → out_data_1, in_data_2 → block → out_data_2. The merge is allowed because the vector masks 0xCC00 and 0xEA do not intersect, the merged block is executed under the mask 0xCCEA

Еще один подход, о котором стоит упомянуть, но который не тестировался в рамках данной работы, связан с объединением соседних блоков, для которых (mask_1 & mask_2) != 0x0. Если мы имеем дело с двумя масками низкой плотности, которые пересекаются, но для которых выполнено условие popcnt(mask_1) + popcnt(mask_2) <= w, то такие блоки также можно объединить. Для этого необходимо найти такое преобразование одной из масок (например, mask_1) perm_to, что оно будет иметь обратное преобразование perm_from(perm_to(mask_1)) == mask_1 и будет выполнено условие (perm_to(mask_1) & mask_2) == 0x0. В этом случае элементы входных данных переставля-

ются местами в соответствии с преобразованием perm_to, применяется описанная выше техника объединения блоков, а для выходных данных выполняется перестановка элементов в соответствии с преобразованием perfm_from (см. рис. 8).



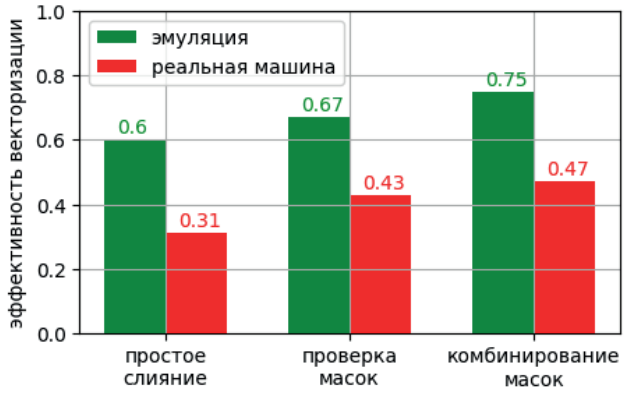
Р и с. 8. Схема вычислений с объединением двух векторизованных блоков при условии пересечения их масок. Для объединения применяется техника изменения порядка элементов в данных одного из блоков

F i g. 8. The computational scheme with the merging of two vectorized blocks under the condition of their masks intersecting. The merging uses a technique of changing the order of elements in the data of one of the blocks

В описанном подходе следует отметить следующие моменты. Во-первых, объединять можно не только два соседние блока, но также три и более, хотя это существенно усложняет программный код и увеличивает количество накладных расходов. Во вторых, следует принимать во внимание доступные операции по изменению порядка расположения элементов векторов, так как таких операций достаточно много и они отличаются по времени выполнения (SHUF, UNPCK, VPERM, VPERMIL)[3].

Результаты исследования

Для анализа полученных результатов были рассмотрены следующие три подхода к векторизации плоского цикла с условием. В качестве базового метода векторизация принималось простое слияние путей исполнения под соответствующими предикатами с последующим объединением w последовательных скалярных итераций в одну векторную (простое слияние). Данный базовый метод сравнивался с двумя рассмотренными выше улучшениями: проверка масок блоков на пустоту (проверка масок) и слияние двух соседних блоков при условии отсутствия пересечения их масок (комбинирование масок). Анализ эффективности применения преобразований рассматривался на приведенной в листинге 1 функции prefun из реализации газодинамического римановского решателя. Профиль исполнения функции собирался на задачах моделирования распада разрыва при различных начальных условиях [22, 23]. Собранный профиль исполнения функций римановского решателя доступен в открытом репозитории[4]. Эффективность векторизации при выбранных подходах измерялась двумя способами. В качестве первого способа использовался режим эмуляции векторных инструкций. В настоящее время используются различные эмуляторы AVX-512 с помощью которых можно оценить эффективность векторного кода [24]. В данной работе мы ограничились инструментом, позволяющим отследить плотность используемых в коде масок и общее количество скалярных и векторных операций [25]. Вторым способом сравнения был замер производительности результирующего векторного кода на микропроцессоре Intel Xeon Phi Knights Landing 7290[5]. Результаты сравнения представлены на рис. 9.



Р и с. 9. Результаты сравнения эффективности векторизации при простом слиянии, с проверкой масок и с комбинированием масок в режимах эмуляции и на микропроцессоре Intel Xeon Phi Knights Landing 7290

F i g. 9. Comparison results of vectorization efficiency with simple merging, with mask checking and with mask combining in emulation and on the Intel Xeon Phi Knights Landing 7290 microprocessor modes

Эксперименты показали, что в режиме эмуляции слияние ветвей исполнения привело к эффективности векторизации 0.6. Использование дополнительных методов повышения плотности векторных масок в коде – проверки масок и комбинирования масок – привело к повышению эффективности векторизации до показателей 0.67 и 0.75 соответственно. Эффективность векторизации при проведении замеров на реальной машине оказалась скромнее. Простое слияние путей исполнения позволило достичь эффективности 0.31, а использование проверок масок и комбинирования масок позволило повысить эффективность векторизации до 0.43 и 0.47 соответственно.

Обсуждение и заключение

Исследования, проведенные в рамках данной работы, показали, что векторизация плоских циклов с условиями может быть

[3] Intel 64 and IA-32 Architectures Software Developer's Manual. Combined Volumes: 1, 2A, 2B, 2C, 2D, 3A, 3B, 3C, 3D, and 4.

[4] Rybakov A. riemann_vec [Electronic resource] // Github, 2024. URL: https://github.com/r-aax/riemann_vec (дата обращения: 10.06.2024).

[5] Jeffers J., Reinders J., Sodani A. Intel Xeon Phi Processor High Performance Programming. Knights Landing edition. San Francisco, CA, USA: Morgan Kaufmann, 2016. 662 p.

выполнена и при этом могут достигаться высокие показатели эффективности векторизации. Так на рассмотренном отдельном примере из реализации газодинамического решателя простое слияние путей исполнения привело к эффективности векторизации с показателем 0.31 на реальной машине, а использование дополнительных стратегий повышения плотности векторных масок в результирующем коде позволило повысить этот показатель до значения 0.47. При этом в работе были рассмотрены только самые простые подходы к проверке и комбинированию масок. Не затрагивались методы комбинирования масок для трех и более векторных блоков, не рассматривались методы комбинирования для пересекающихся масок. Так как возможности применения стратегий повышения плотности масок связаны с динамическими проверками масок, то это сопряжено с возникновением серьезных накладных расходов. Интересным видится подход к выбору стратегии повышения плотности масок, основанный на предварительном анализе профиля исполнения того или иного участка кода (анализ масок и вероятностей переходов). Такие эксперименты планируется выполнить в последующих исследованиях.

References

[1] Trebotich D. Exascale CFD in heterogenous systems. *Journal of Fluids Engineering.* 2024:146(4):1-19. https://doi.org/10.1115/1.4064534

[2] Menshov I., Pavlukhin P. GPU-native gas dynamic solver on octree-based AMR grids. *Journal of Physics Conference Series.* 2020:1640(1):012017. https://doi.org/10.1088/1742-6596/1640/1/012017

[3] Bosnjak D., Pepe A., Schussnig R., et al. Higher-order block-structured hex meshing of tubular structures. *Engineering with Computers.* 2023:40(2):931-951. https://doi.org/10.1007/s00366-023-01834-7

[4] Carr G.E., Biocca N., Urquiza S.A. A biologically-inspired mesh moving method for cyclic motions mesh fatigue. *Computational Mechanics.* 2024;75:475-486. https://doi.org/10.1007/s00466-024-02514-z

[5] Park Y.M., Jee S. Numerical study on interactional aerodynamics of a quadcopter in hover with overset mesh in OpenFOAM. *Physics of Fluids.* 2023:35(8):085138. https://10.1063/5.0160689

[6] Wan Y., Zhao Z., Liu J., et al. Large-scale homo- and heterogeneous parallel paradigm design based on CFD application PHengLEI. *Concurrency and Computation Practice and Experiment.* 2024:36:e7933. https://doi.org/10.1002/cpe.7933

[7] Jakobs T., Klöckner O., Rünger G. Parallelization with load balancing of the weather scheme WSM7 for heterogeneous CPU-GPU platforms. *The Journal of Supercomputing.* 2024:80:14645-14665. https://doi.org/10.1007/s11227-024-06009-9

[8] Lai J., Yu H., Tian Z., et al. Hybrid MPI and CUDA parallelization for CFD applications on multi-GPU HPC clusters. *Scientific Programming.* 2020:1-15. https://doi.org/10.1155/2020/8862123

[9] Bashir S., Usman A., Mumtaz Y., et al. Parallelization of lattice Bolzmann method for CFD using message passing interface. *Thermal Science.* 2022:26(spec. issue 1):211-218. https://doi.org/10.2298/TSCI22S1211B

[10] Lin H., Yan L., Chang Q. et al. O2ath: an OpenMP offloading toolkit for the sunway heterogeneous manycore platform. *CCF Transactions on High Performance Computing.* 2024:6(7). https://doi.org/10.1007/s42514024001911

[11] He X., Wang K., Feng Y., et al. An implementation of MPI and hybrid OpenMP/MPI parallelization strategies for an implicit 3D DDG solver. *Computers & Fluids.* 2022:241(4):105455. https://doi.org/10.1016/j.compfluid.2022.105455

[12] Cebrian J.M., Natvig L., Jahre M. Scalability analysis of AVX-512 extensions. *The Journal of Supercomputing.* 2020:76:2082-2097. https://doi.org/10.1007/s11227-019-02840-7

[13] Kulikov I., Chernykh I., Tutukov A. A new hydrodynamic code with explicit vectorization instructions optimizations that is dedicated to the numerical simulation of astrophysical gas flow. I. Numerical method, tests, and model problem. *The Astrophysical Journal. Supplement Series.* 2019;243(1):4. https://doi.org/10.3847/1538-4365/ab2237

[14] Glinting B.M., Mundani R.-P. Comparison of shallow water solvers: applications for dam-break and tsunami cases with reordering strategy for efficient vectorization on modern hardware. *Water.* 2019:11(4):639. https://doi.org/10.3390/w11040639

[15] Yildirim A., Mader C., Martins J.R.R.A. Accelerating parallel CFD codes on modern vector processors using blockettes. In: Proceedings of the Platform for Advanced Scientific Computing Conference (PASC '21). New York, NY, USA: Association for Computing Machinery; 2021. Article number: 11. https://doi.org/10.1145/3468267.3470615

[16] Rucci E., Moreno E., Pousa A., et al. Optimization of the N-body simulation on Intel's architectures based on AVX-512 instruction set. In: Pesado P., Arroyo M. (eds.) Computer Science – CACIC 2019. CACIC 2019. *Communications in Computer and Information Science.* Vol. 1184. Cham: Springer; 2020. p. 37-52. https://doi.org/10.1007/978-3-030-48325-8_3

[17] Rucci E. Garcia C., Botella G., et al. SWIMM 2.0: Enhanced Smith-Waterman on Intel's multicore and manycore architectures based on AVX-512 vector extensions. *International Journal of Parallel Programming.* 2019;47(17). https://doi.org/10.1007/s10766-018-0585-7

[18] Choi Y., Choi H., Chung S. AVX512Crypto: parallel implementations of Korean block ciphers using AVX512. *IEEE Access.* 2023;11:55094-55106. https://doi.org/10.1109/ACCESS.2023.3278993

[19] Cheng H., Fotiadis G., Großschädl J., et al. Batching CSIDH group actions using AVX512. *IACR Transactions on Cryptographic Hardware and Embedded Systems.* 2021:4:618-649. https://doi.org/10.46586/tches.v2021.i4.618-649

[20] Savin G.I., Shabanov B.M., Rybakov A.A., Shumilin S.S. Vectorization of flat loops of arbitrary structure using instructions AVX-512. *Lobachevskii Journal of Mathematics.* 2020:41(12):2575-2592. https://doi.org/10.1134/S1995080220120331

[21] Rybakov A., Chopornyak A. Improving vector code performance by monitoring masks density in vector instructions. *Trudy NIISI RAN.* 2020:10(4):40-47. (In Russ., abstract in Eng.) https://doi.org/10.25682/NIISI.2020.4.0006

[22] Toh Y.H. Efficient non-iterative multi-point method for solving the Riemann problem. *Nonlinear Dynnamics.* 2024:112(1):1-13. https://doi.org/10.1007/s11071-023-09229-5

[23] Zeng Z., Feng C., Yu C., et al. Linearized double-shock approximate Riemann solver for augmented linear elastic solid. *Numerical Mathematics Theory Methods and Applications.* 2021:15(1). https://doi.org/10.4208/nmtma.OA-2021-0021

[24] Lee S., Kim Y., Nam D., et al. Gem5-AVX: Extension of the Gem5 simulator to support AVX instruction sets. *IEEE Access.* 2024;12:20767-20778. https://doi.org/10.1109/ACCESS.2024.3359296

[25] Rybakov A.A., Shwindt A.N. Tools for vectorizing a flat loop body using AVX-512 vector instructions. *Software & Systems.* 2023:36(4):561-572. (In Russ., abstract in Eng.) https://doi.org/10.15827/0236-235X.142.561-572

Поступила 10.06.2024; одобрена после рецензирования 22.08.2024; принята к публикации 20.09.2024.
Submitted 10.06.2024; approved after reviewing 22.08.2024; accepted for publication 20.09.2024.

Об авторе:

Рыбаков Алексей Анатольевич, начальник отдела суперкомпьютерных технологий и систем, отделения суперкомпьютерных систем и параллельных вычислений, ФГБУ «Национальный исследовательский центр «Курчатовский институт»» (123182, Российская Федерация, г. Москва, пл. Академика Курчатова, д. 1); ведущий научный сотрудник Межведомственного суперкомпьютерного центра Российской академии наук – филиала ФГУ «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук» (119334, Российская Федерация, г. Москва, Ленинский пр., д. 32а), кандидат физико-математических наук, **ORCID: https://orcid.org/0000-0002-9755-8830,** rybakov.aax@gmail.com

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Alexey A. Rybakov, Head of the Department of Supercomputer Technologies and Systems, Division of Supercomputer Systems and Parallel Calculations, National Research center “Kurchatov Institute for Theoretical and Experimental physics” (1 Akademicianika Kurchatova Sq., Moscow 123182, Russian Federation); Lead Researcher in Joint Supercomputer Center of the Russian Academy of Sciences – Branch of the Federal State Institution “Scientific Research Institute for System Analysis of the Russian Academy of Sciences” (32a Leninsky Ave., Moscow 119334, Russian Federation), Cand. Sci. (Phys.-Math.), **ORCID: https://orcid.org/0000-0002-9755-8830,** rybakov.aax@gmail.com

The author has read and approved the final manuscript.