Болгак А.В.
Юго-Западный государственный университет
305040, г. Курск, ул. 50 лет Октября, 94

Алгоритмическая и высокоуровневая оптимизация в задаче умножения плотных квадратных вещественных матриц одинарной точности для однопоточной программной реализации с последующей оценкой реальной производительности

Высокопроизводительные вычисления используются во многих отраслях повседневной жизни. Областями применения таких вычислений являются оборонная промышленность, создание новейших лекарственных препаратов, высокочастотная торговля фьючерсами, моделирование различных деталей в машиностроении, проектирование роботизированных средств, томография, классификация бинарных отношений [1], прогнозирование смены климата, решение задач линейной алгебры и дифференциальных уравнений и так далее.

Умножение матриц – это одна из фундаментальных задач, решение которой позволяет эффективно задействовать основные вычислительные ресурсы современных процессоров и графических ускорителей, тем самым, увеличивая реальную производительность и уменьшая временные затраты на решение поставленной задачи за счет алгоритмической и высокоуровневой оптимизации соответствующей программной реализации.

Актуальность данной проблематики заключается в том, что в настоящее время многие алгоритмы решения прикладных задач сводятся именно к матричному умножению, при этом необходимо минимизировать дополнительные расходы, связанные с подготовкой данных, получаемых из оперативной памяти [2]. Время выполнения таких программных реализаций напрямую влияет на время решения той или иной прикладной задачи. Исходя из этого, разработано множество подходов, оптимизирующих выполняемые операции различными способами.

Целью данной работы является алгоритмическая и высокоуровневая оптимизация программной реализации задачи умножения квадратных матриц для однопоточной CPU-ориентированной программной реализации с последующей оценкой её производительности.

В работе рассмотрены основные аспекты анализа эффективности различных подходов к реализации решения задачи умножения квадратных матриц A и B размера N × N для однопоточной высокоуровневой программной реализации на современном поколении процессоров. Данная реализация ориентирована на использование процессоров семейства «x86» с оптимизацией работы кэш-памяти процессора.

Для решения поставленной задачи и оценки производительности каждого из алгоритмов были разработаны следующие программные реализации: «классическое» умножение, умножение с буферизацией столбца, блочное умножение. Исследование проводилось на базе процессора 11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz, используемый компилятор – Microsoft Visual Studio 2017.

Анализ достигнутой реальной производительности вышеперечисленных программных реализаций позволяет сделать вывод о том, что умножение матриц с использованием блочного подхода является наиболее оптимальным и в каждом конкретном случае (размерность задачи, аппаратная конфигурация) имеет место вполне определенный оптимальный размер блока S, при котором достигается максимальная реальная производительность, а время умножения матриц становится минимальным.

Вероятнее всего, данный эффект связан с микроархитектурными особенностями реализации кэш-памяти для данного процессора (параметры оперативной памяти (латентность, пропускная способность), особенности размещения данных в КЭШах процессора различного уровня, их латентность, ассоциативность, пропускная способность связывающих их шин и так далее).

Также стоит отметить, что путем раскрутки внутреннего цикла (англ. loop unrolling) [3], формально являющейся высокоуровневой оптимизацией, удалось добиться дополнительного снижения времени выполнения умножения матриц. При этом значительно сокращается время обработки, а также изменяется оптимальный размер блока S (см. рис. 1). Раскрутка на разумное число итераций позволяет повысить параллелизм на уровне инструкций (сокр. ILP), тем самым

сокращая время выполнения умножения матриц. Чрезмерная раскрутка цикла, напротив, может привести к исчерпанию емкости кэша команд L1i, что существенно увеличивает время обработки алгоритма.

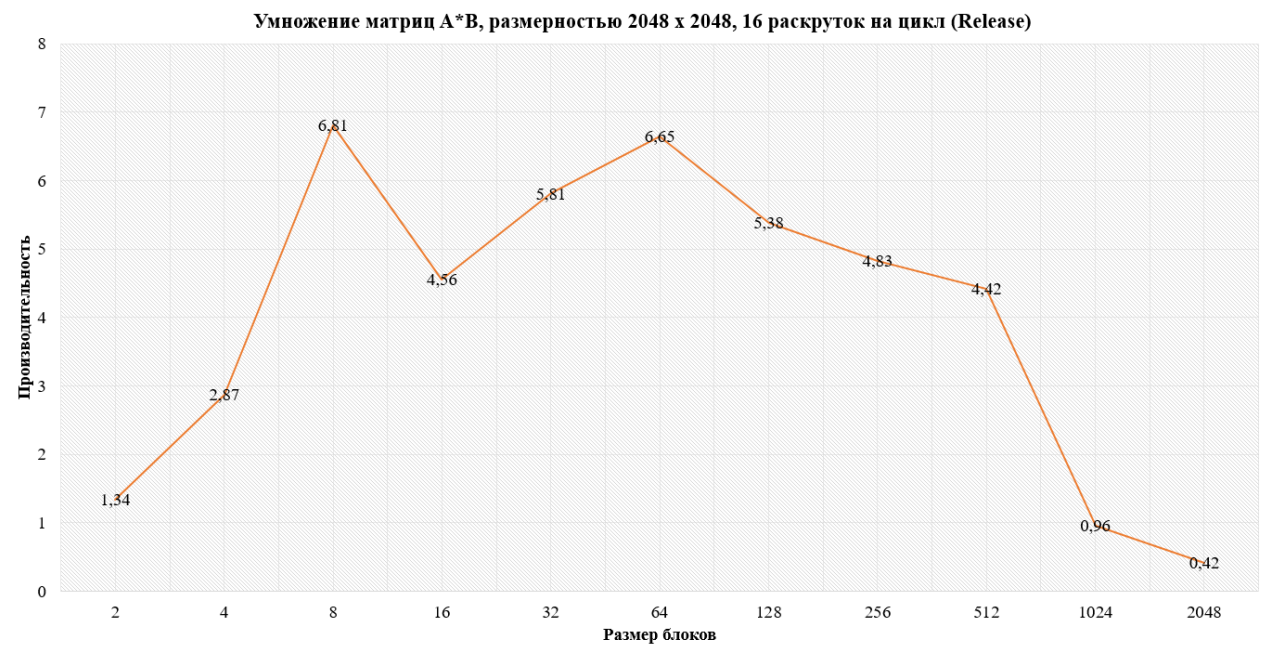

Рис. 1. График зависимости реальной производительности умножения вещественных матриц одинарной точности размером 2048 x 2048 от размера блока S (раскрутка внутреннего цикла – 16 итераций)

Сравнение блочного умножения и умножения с буферизацией столбца позволяет сделать вывод о том, что для процессора 11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz оба варианта приводят к достижению сопоставимой производительности.

Таким образом, результаты исследования показали, что процессор 11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz демонстрирует реальную производительность на уровне 2,1 – 6,8 GFLOP/s при однопоточной скалярной обработке данных, что в десятки раз превосходит производительность «классического» варианта умножения без оптимизации работы с памятью и сопоставимо с полученными ранее цифрами для процессоров предыдущих поколений [4]. В перспективе дальнейших исследований планируется анализ реальной производительности для матриц различного типа и разработка программных реализаций, ориентированных на выполнение на вычислительных средствах с параллельной архитектурой.

Литература

1. Ватутин Э.И., Зотов И.В. Построение матрицы отношений в задаче оптимального разбиения параллельных управляющих алгоритмов // Известия Курского государственного технического университета. Курск, 2004. № 2. С. 85–89.

2. Штейнберг Б.Я. Блочно-рекурсивное параллельное перемножение матриц // Известия высших учебных заведений. Приборостроение. 2009. Т. 52. № 10. С. 33–41.

3. Intel 64 and IA-32 Architectures Software Developer’s Manual. Volume1: Basic Architecture. Order number 253665-021.

4. Ватутин Э.И., Мартынов И.А., Титов В.С. Оценка реальной производительности современных процессоров в задаче умножения матриц для однопоточной программной реализации // Известия Юго-Западного государственного университета. Серия: Управление, вычислительная техника, информатика. Медицинское приборостроение. 2013. № 4. С. 11–20.