Нижегородский государственный университет им. Н.И. Лобачевского
Институт информационных технологий, математики и механики

Об опыте оптимизации базовых алгоритмов работы с ленточными матрицами в библиотеке OpenBLAS

Ковалев К.И., Пирова А.Ю.,
Воденеева А.А., Устинов А.В.,
Козинов Е.А., Линев А.В.,
Волокитин В.В., Мееров И.Б.

«Суперкомпьютерные дни в России»
Москва, 29-30 сентября 2025 г.

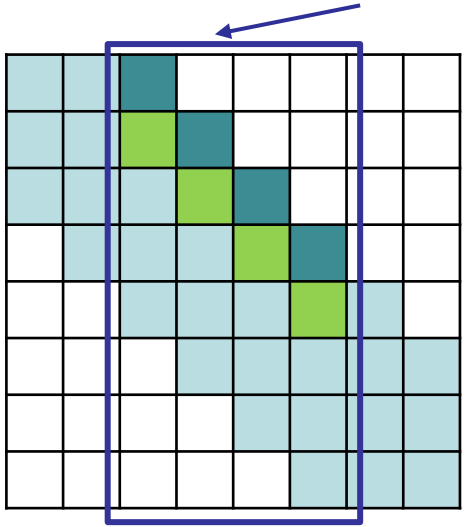
Цель работы

- Цель работы - оптимизация реализаций некоторых матрично-векторных операций с ленточными матрицами в библиотеке OpenBLAS для процессоров архитектуры RISC-V.
- По результатам тестирования выбраны четыре функции BLAS-2:
  - **GBMV** – умножение ленточной матрицы общего вида на вектор;
  - **SBMV** – умножение симметричной ленточной матрицы на вектор;
  - **TBMV** – умножение треугольной ленточной матрицы на вектор;
  - **TBSV** – решение СЛАУ с треугольной ленточной матрицей.

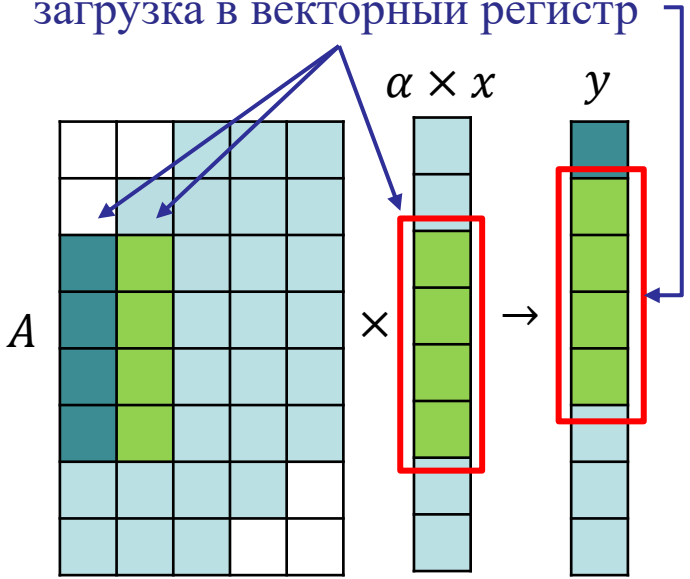
Алгоритм умножения ленточной матрицы общего вида на вектор (GBMV) для матриц с малым числом диагоналей

- **Идея**: разделить матрицу на вертикальные полосы, обход выполнять по диагоналям матрицы

ширина полосы = длине векторного регистра

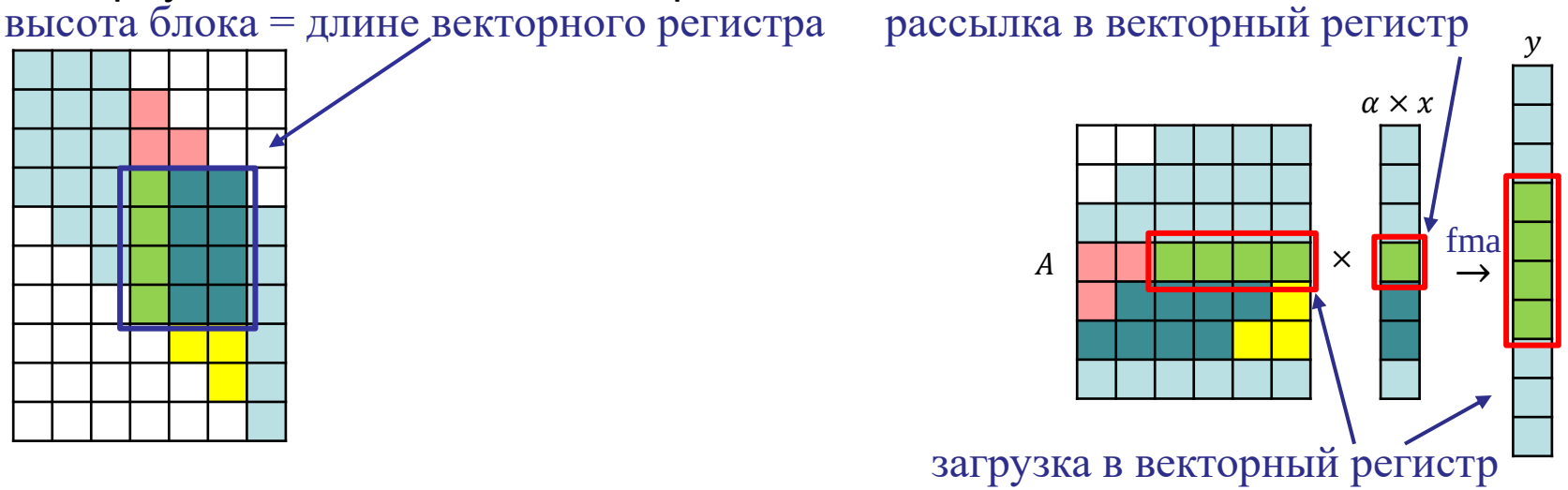


загрузка в векторный регистр



Алгоритм умножения ленточной матрицы общего вида на вектор (GBMV) для матриц с большим числом диагоналей

- **Идея**: Матрица разделяется на полосы ширины k, в каждой полосе выделяется плотный прямоугольный блок и два треугольных блока над и под ним. Вычисления для прямоугольного блока выполняются векторно, для треугольных блоков - скалярно.

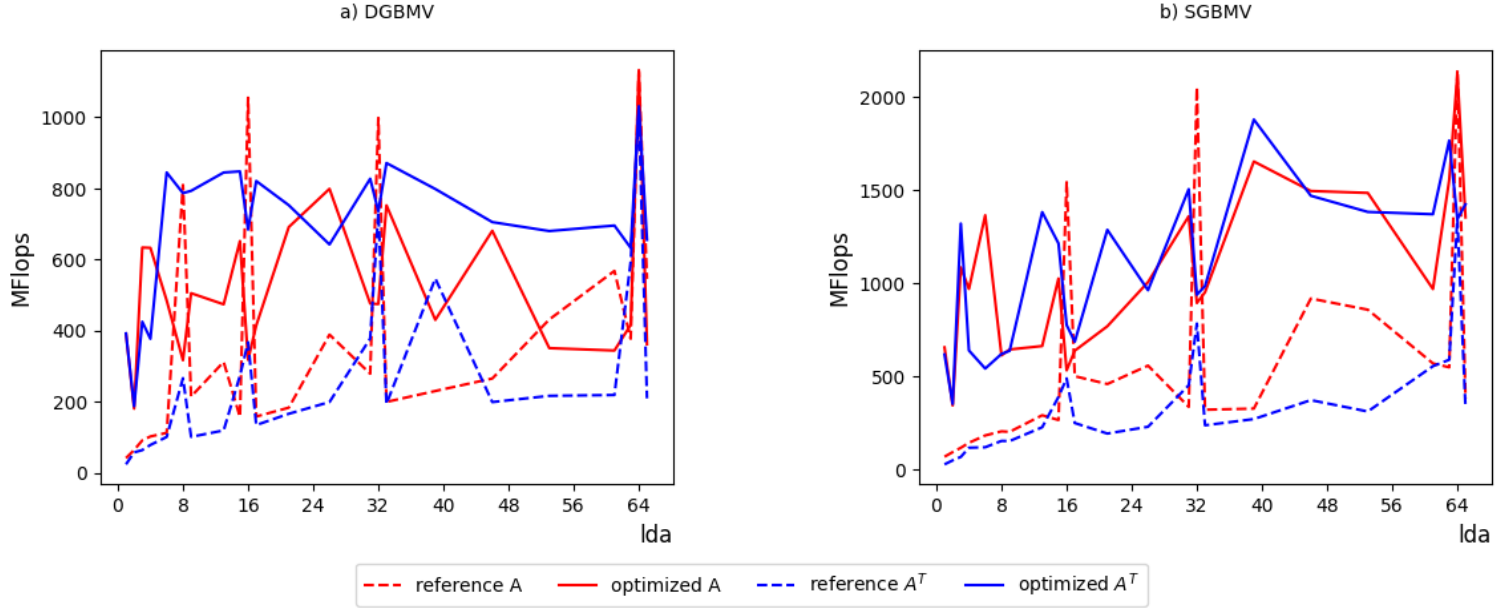


Программная реализация

- Базовая версия – библиотека OpenBLAS.
- В библиотеку интегрированы реализации рассмотренных алгоритмов для матриц с малым и большим числом диагоналей, векторизованные для процессоров RISC-V.
  - наборы инструкций RVV 0.7.1, RVV 1.0;
  - точность одинарная и двойная;
  - последовательная версия.
- Итоговый алгоритм для функций GBMV, SBMV, TBMV:
  из общего интерфейса, реализованного в OpenBLAS, выбирается один из оптимизированных алгоритмов или базовый алгоритм, в зависимости числа диагоналей матрицы. Пороги переключения подобраны экспериментально.
- Код доступен: https://github.com/UNN-ITMM-Software/OpenBLAS/tree/band_matrix_RVV_improving

Вычислительные эксперименты. GBMV. Lichee Pi 4A, RVV 0.7.1

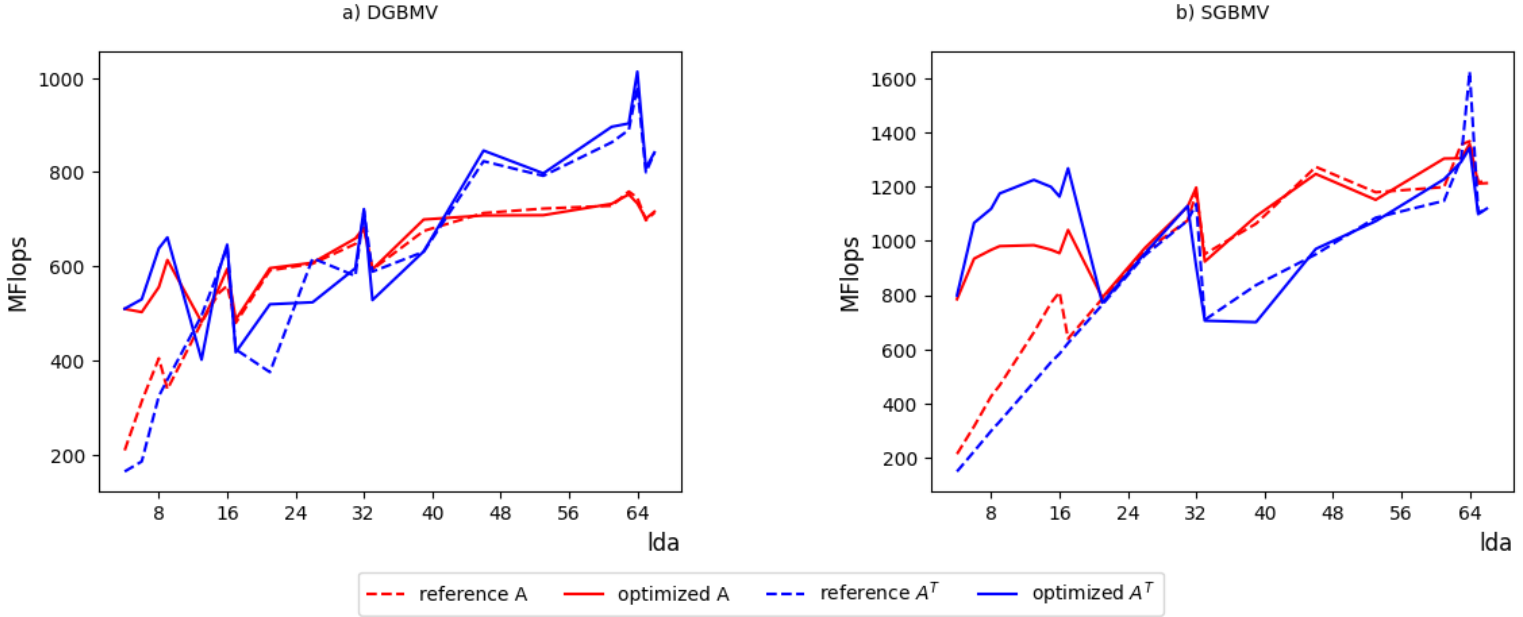
Матрица $n = m = 2.5$ млн.



- Оптимизированный алгоритм **лучше** базовой реализации **в 2–7 раз** при **lda < 64**, кроме lda, кратных 8. Оптимизированный алгоритм 1 используется при lda < 25 для A^T и lda < 3 для A, иначе – алгоритм 2.

Вычислительные эксперименты. GBMV. Banana Pi BPI-F3, RVV 1.0

Матрица $n = m = 2.5$ млн.



- Оптимизированный алгоритм **лучше** базовой реализации при **lda < 9** для double и **lda < 17** для single precision. Среднее опережение – 3 раза.
- Используется оптимизированный алгоритм 1.

Заключение

- Был предложен подход к оптимизации алгоритмов умножения ленточных матриц на вектор, в основе которого – изменение порядка обхода матриц.
- Эффективность предложенных алгоритмов зависит от числа диагоналей матрицы.
- Использование векторных инструкций позволяет значительно ускорить вычисления. Однако реализация с использованием интринсиков не переносима на другие архитектуры, должна разрабатываться отдельно под каждый новый набор векторных команд.
- Предложенные реализации работают до 7 раз быстрее базовой реализации.