Летняя школа 2014 по параллельному программированию

Реализация алгоритма Штрассена умножения матриц на ускорителе Intel Xeon Phi

Студент: Яндуганов А.Е. НГТУ ФМПИ 3 курс

Руководитель: Киреев Сергей Евгеньевич

Цель работы

- Исследование возможностей реализации параллельного алгоритма Винограда-Штрассена для реализации операции DGEMM из библиотеки BLAS на ускорителе **Intel Xeon Phi**

Алгоритм Винограда-Штрассена

Особенности:

- Рекурсивный
- Требует дополнительную память
- 7 умножений вместо 8

$$\mathbf{S}_1 := (\mathbf{A}_{2,1} + \mathbf{A}_{2,2})$$
$$\mathbf{S}_2 := (\mathbf{S}_1 - \mathbf{A}_{1,1})$$
$$\mathbf{S}_3 := (\mathbf{A}_{1,1} - \mathbf{A}_{2,1})$$
$$\mathbf{S}_4 := (\mathbf{A}_{1,2} - \mathbf{S}_2)$$
$$\mathbf{S}_5 := (\mathbf{B}_{1,2} - \mathbf{B}_{1,1})$$
$$\mathbf{S}_6 := (\mathbf{B}_{2,2} - \mathbf{S}_5)$$
$$\mathbf{S}_7 := (\mathbf{B}_{2,2} - \mathbf{B}_{1,2})$$
$$\mathbf{S}_8 := (\mathbf{S}_6 - \mathbf{B}_{2,1})$$

$$\mathbf{P}_1 := \mathbf{S}_2\mathbf{S}_6$$
$$\mathbf{P}_2 := \mathbf{A}_{1,1}\mathbf{B}_{1,1}$$
$$\mathbf{P}_3 := \mathbf{A}_{1,2}\mathbf{B}_{2,1}$$
$$\mathbf{P}_4 := \mathbf{S}_3\mathbf{S}_7$$
$$\mathbf{P}_5 := \mathbf{S}_1\mathbf{S}_5$$
$$\mathbf{P}_6 := \mathbf{S}_4\mathbf{B}_{2,2}$$
$$\mathbf{P}_7 := \mathbf{A}_{2,2}\mathbf{S}_8$$

$$\mathbf{T}_1 := \mathbf{P}_1 + \mathbf{P}_2$$
$$\mathbf{T}_2 := \mathbf{T}_1 + \mathbf{P}_4$$

$$\mathbf{C}_{1,1} := \mathbf{P}_2 + \mathbf{P}_3$$
$$\mathbf{C}_{1,2} := \mathbf{T}_1 + \mathbf{P}_5 + \mathbf{P}_6$$
$$\mathbf{C}_{2,1} := \mathbf{T}_2 - \mathbf{P}_7$$
$$\mathbf{C}_{2,2} := \mathbf{T}_2 + \mathbf{P}_5$$

Шаг 1: последовательный алгоритм

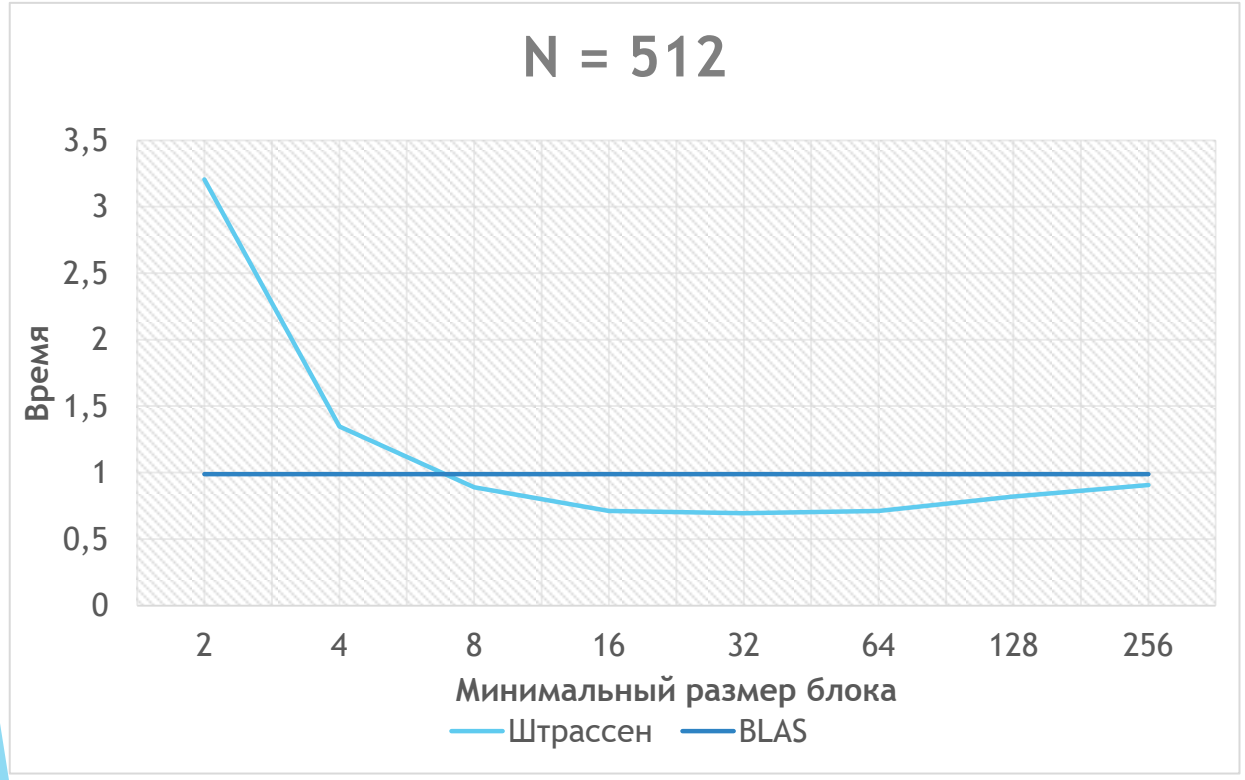
- Реализована операция DGEMM из библиотеки BLAS

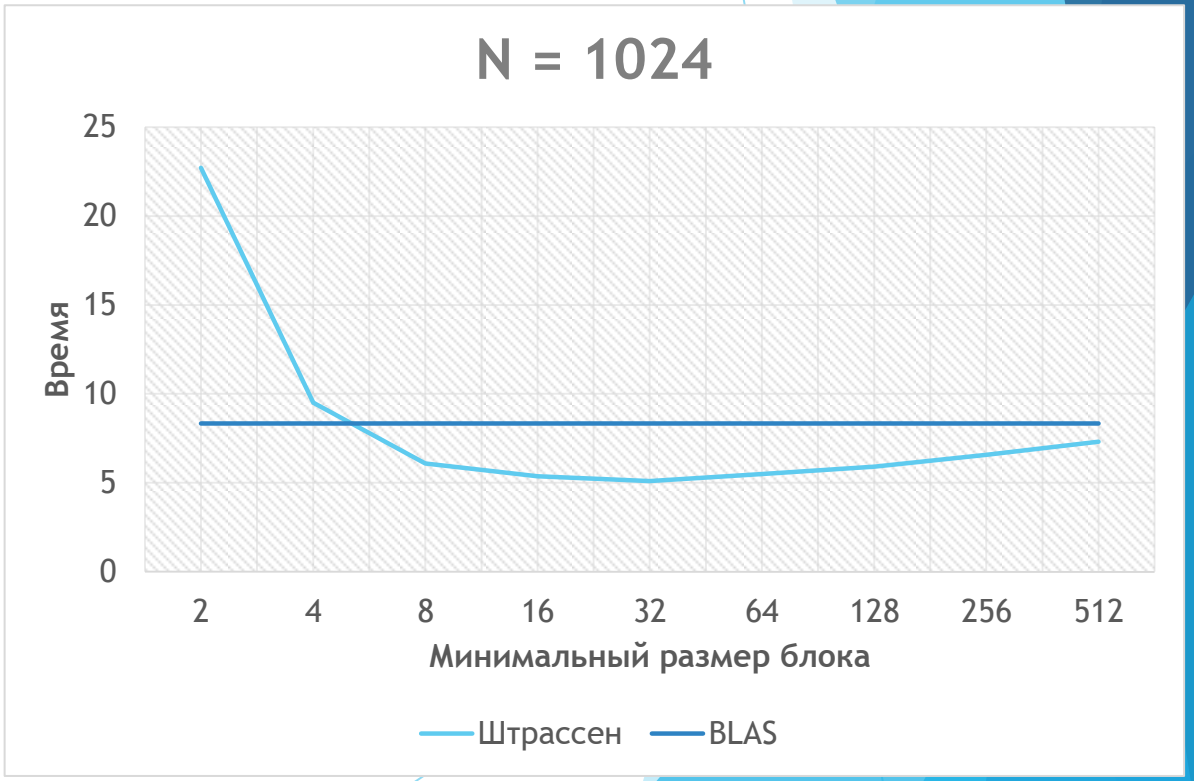
$$C = \alpha AB + \beta C$$

- Размер матриц $N = 2^k$
- 3 дополнительные блока матрицы размером 2^{k-1} на каждом уровне рекурсии
- Начиная с некоторого размера подматрицы используется библиотечная функция

Время работы

- Сравнение алгоритма Штрассена и обычного алгоритма на разных матрицах при разном размере минимального блока



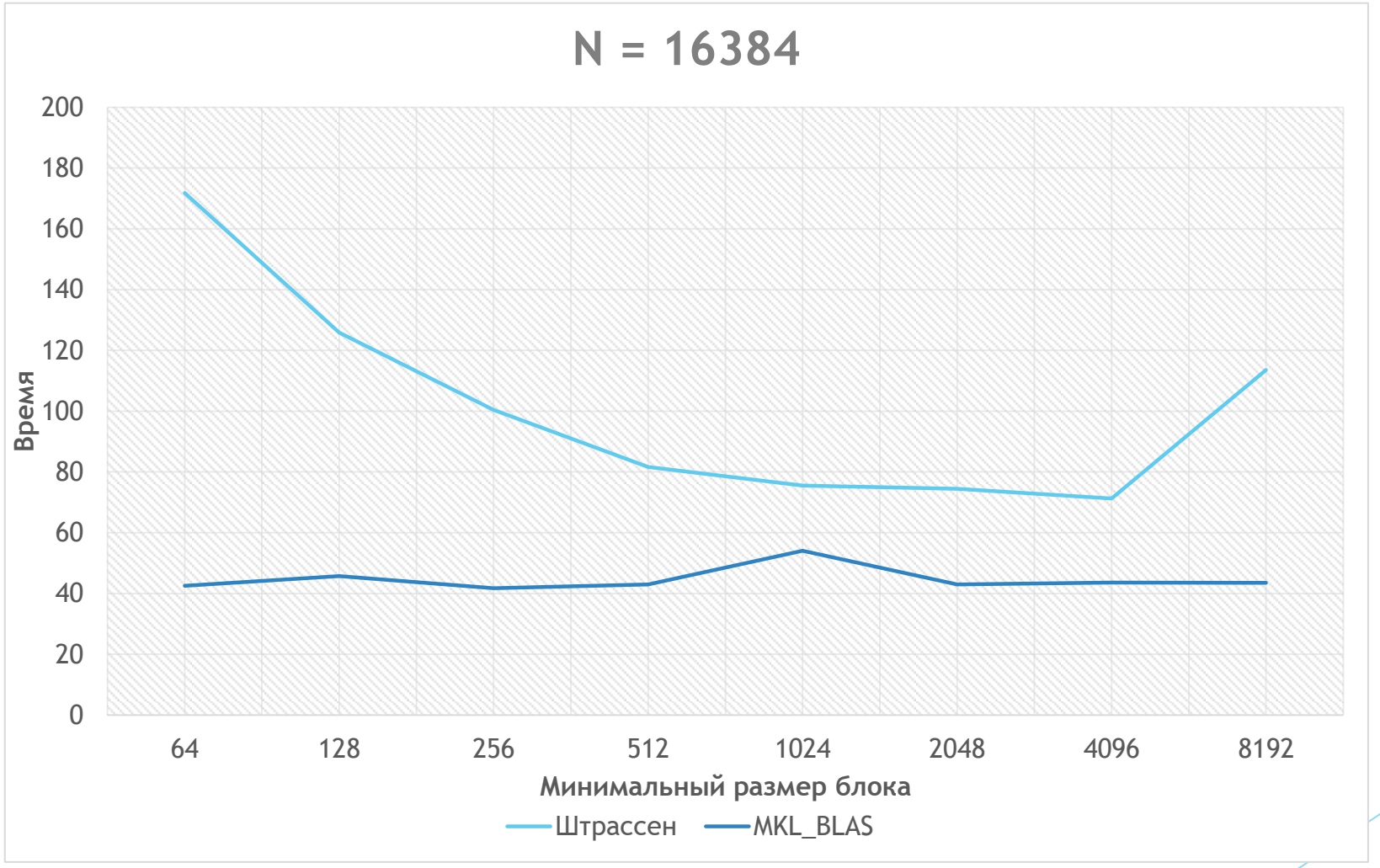


Шаг 2: параллельный алгоритм

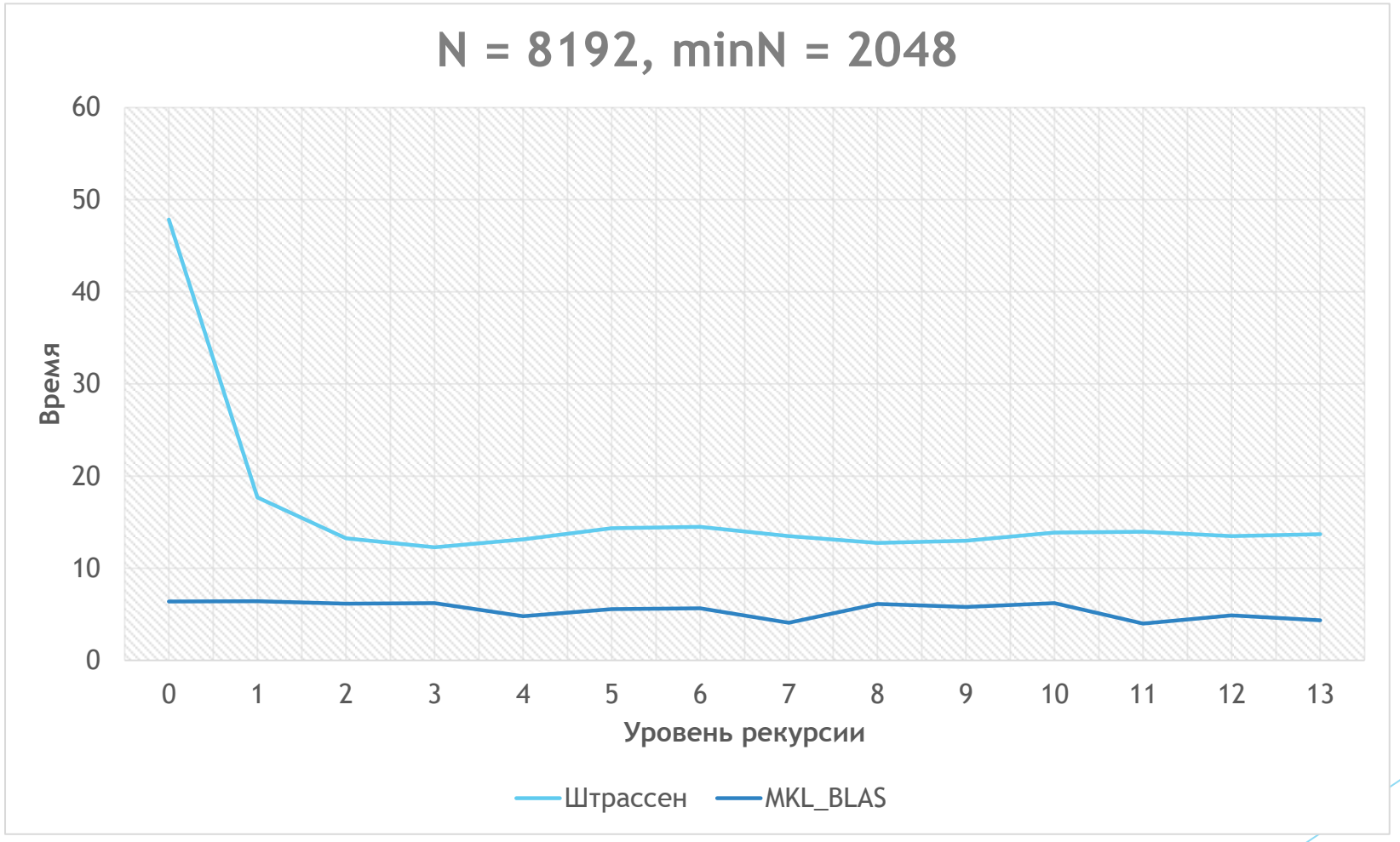
Особенности:

- Максимальный параллелизм (одновременно 7 умножений подматриц на каждом уровне рекурсии)
- Не экономим память (15 дополнительных подматриц на каждом уровне рекурсии)

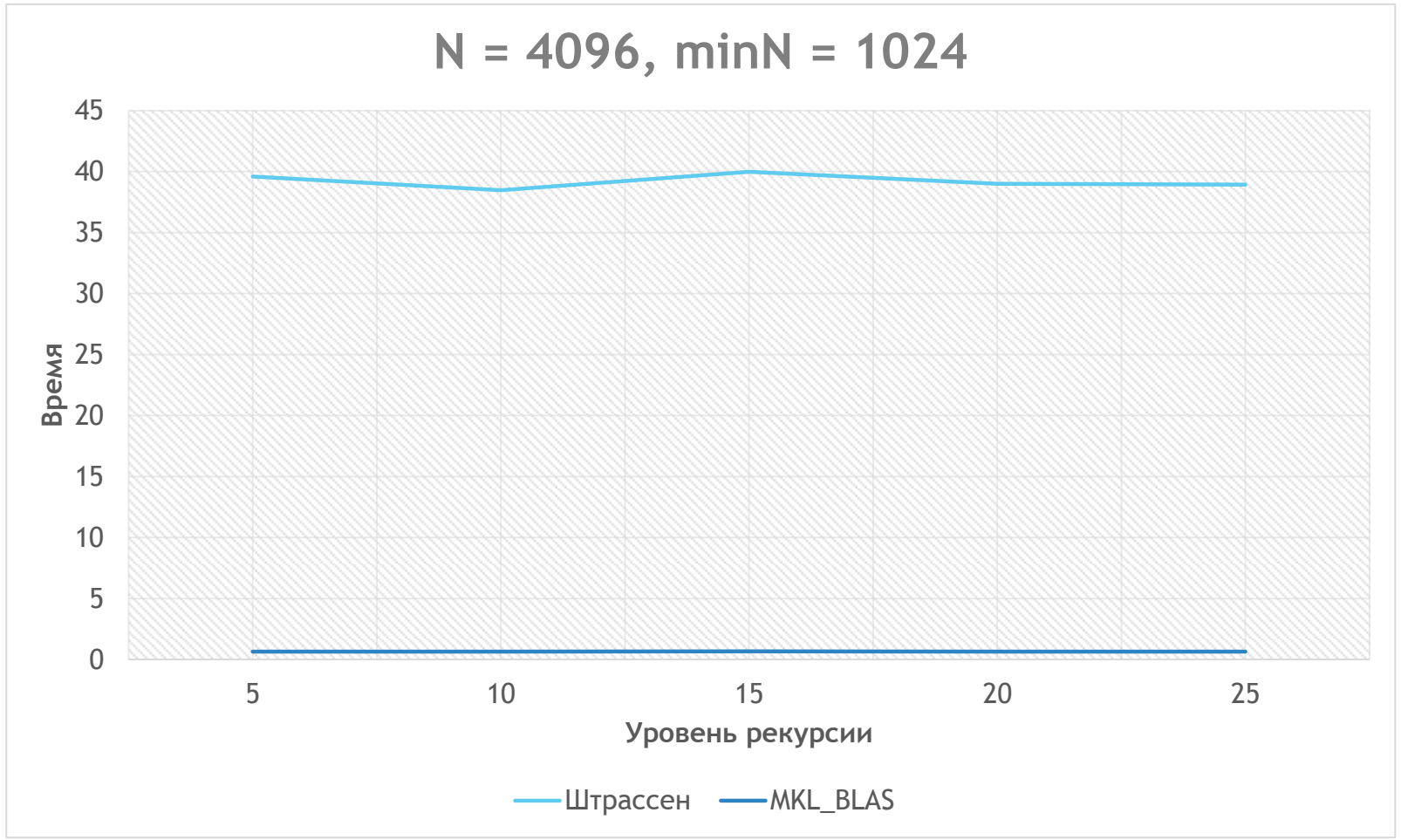
Большая матрица, разные размеры блока



Оптимальный размер блока, разное число параллельных блоков

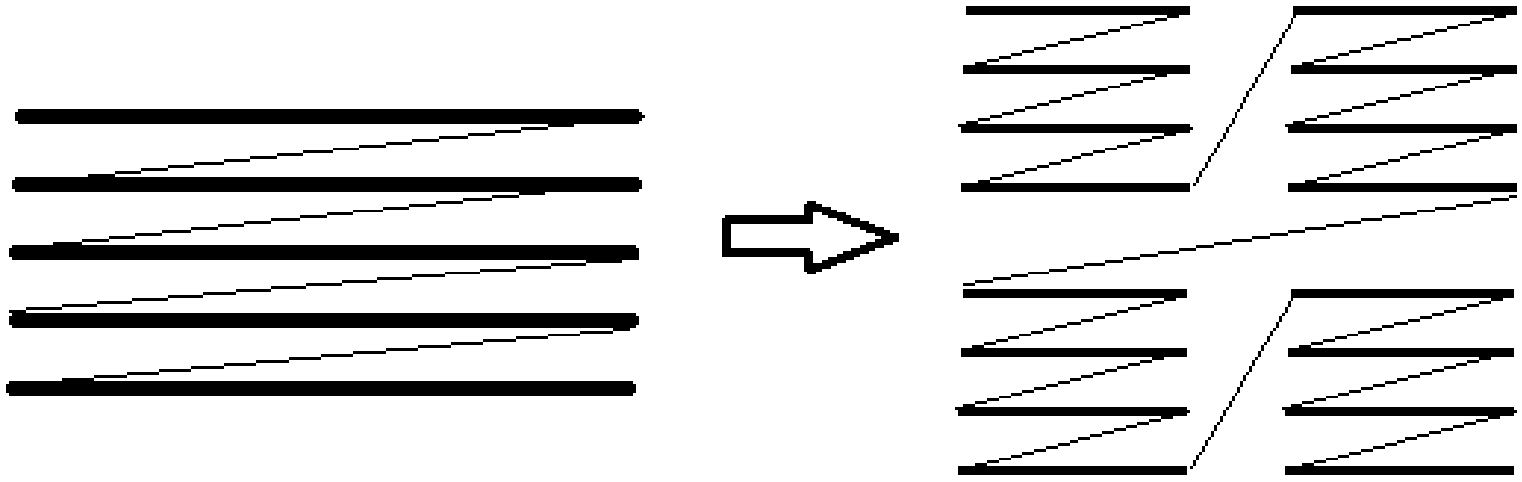


Запуск на Intel Xeon Phi



Пути дальнейшей оптимизации

- Уменьшение требований к памяти за счет меньшей степени параллелизма
- Переупорядочивание матриц для более эффективного доступа к памяти



Спасибо за внимание!

- Вопросы?