Международная научная конференция
«Параллельные вычислительные технологии 2010»
(29 марта – 2 апреля 2010 г., Уфа)

Решение задачи анализа рыночной корзины на процессорах Cell

К.С. Пан, М.Л. Цымблер
Южно-Уральский государственный университет (Челябинск)

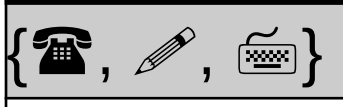
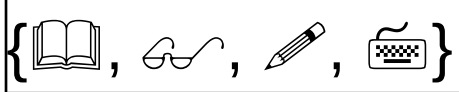
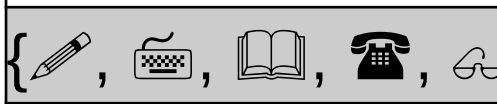
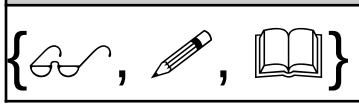
Работа выполнена при финансовой поддержке РФФИ (проект 09-07-00241-а).

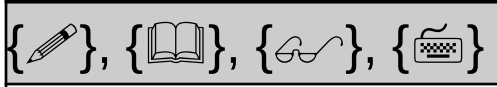
Задача анализа рыночной корзины

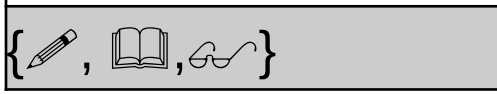
- Нахождение всех наборов товаров, которые часто приобретаются совместно.
- *Корзина* – набор товаров, приобретенных совместно.
  - B – множество анализируемых корзин
  - I – множество всех товаров $I = \bigcup_{b \in B} b$
- *Опорное число* – количество корзин, содержащих данный набор товаров.
  - $support(c, B) = card\{b \in B : c \subset b\}$
  - s_{min} – минимальное опорное число
- Найти множество

$$L = \{l \subset I : support(l, B) \geq s_{min}\}$$

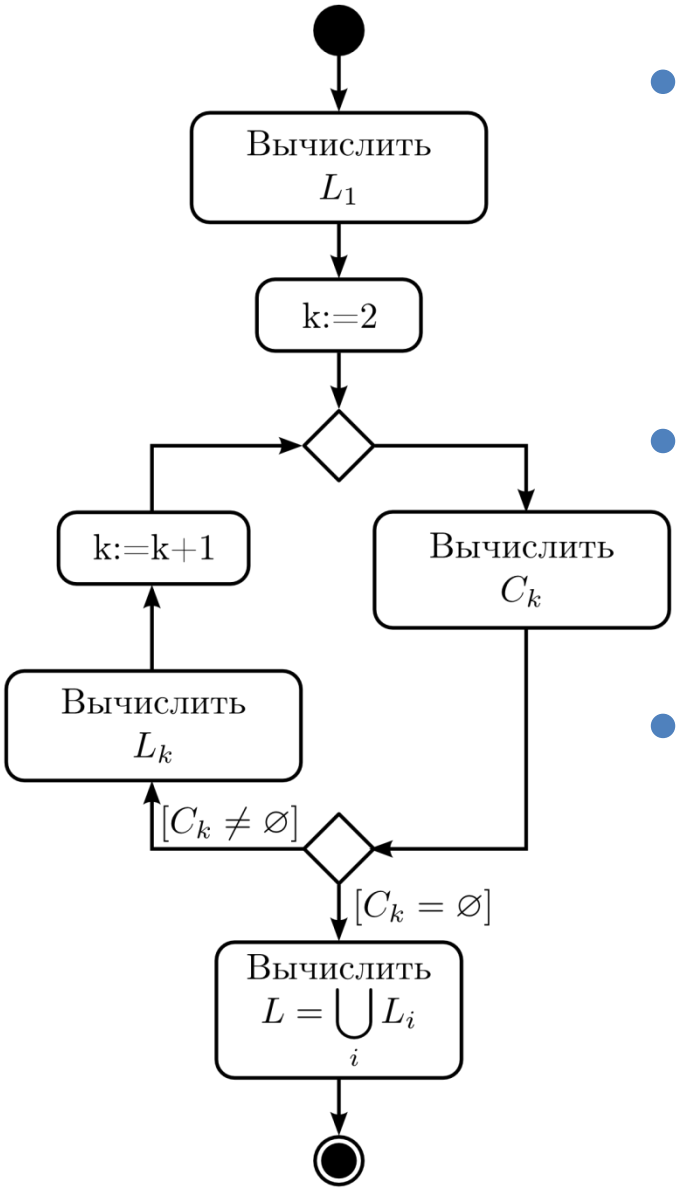
Товары


Корзины





Частые наборы (s_{min}=3)




Последовательный алгоритм apriori



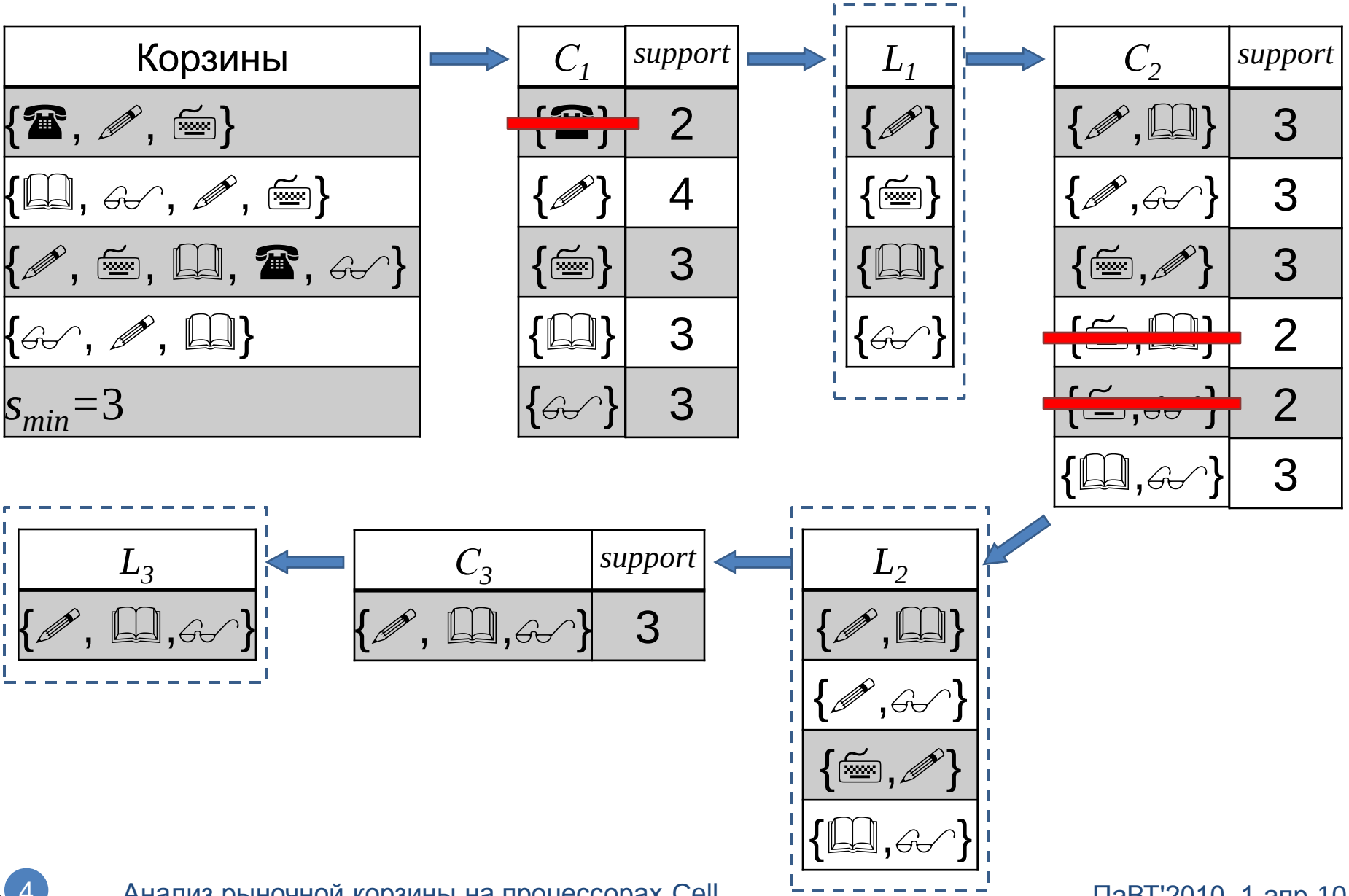
- *Антимонотонность* опорного числа:

$$\forall \gamma \subset c \; support(\gamma, B) \geq support(c, B)$$

$support(\{📖\}) \geq support(\{✍, ✏, 📖\})$

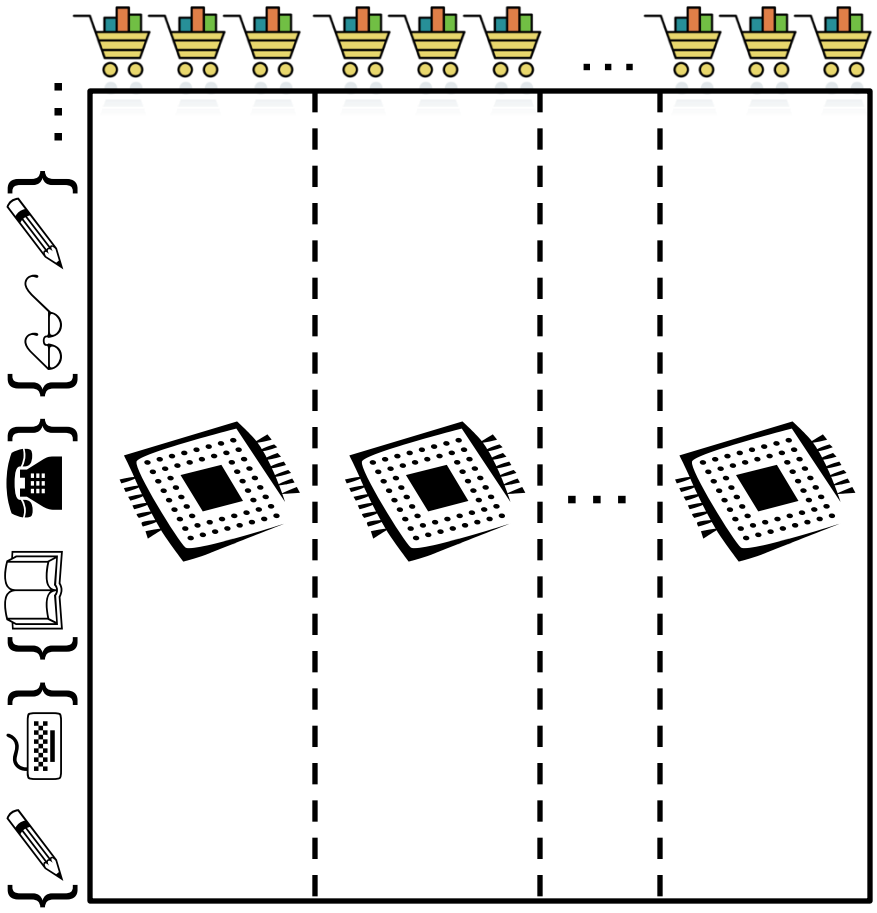
- *Кандидат* – набор c, для которого проверяется условие $c \in L$.
- Формирование C_i (множества кандидатов)
  - на *1* шаге – из I (множества товаров)
  - на *k* шаге – из L_{k-1} (множество часто встречающихся наборов длины *k*-1)

Последовательный алгоритм apriori

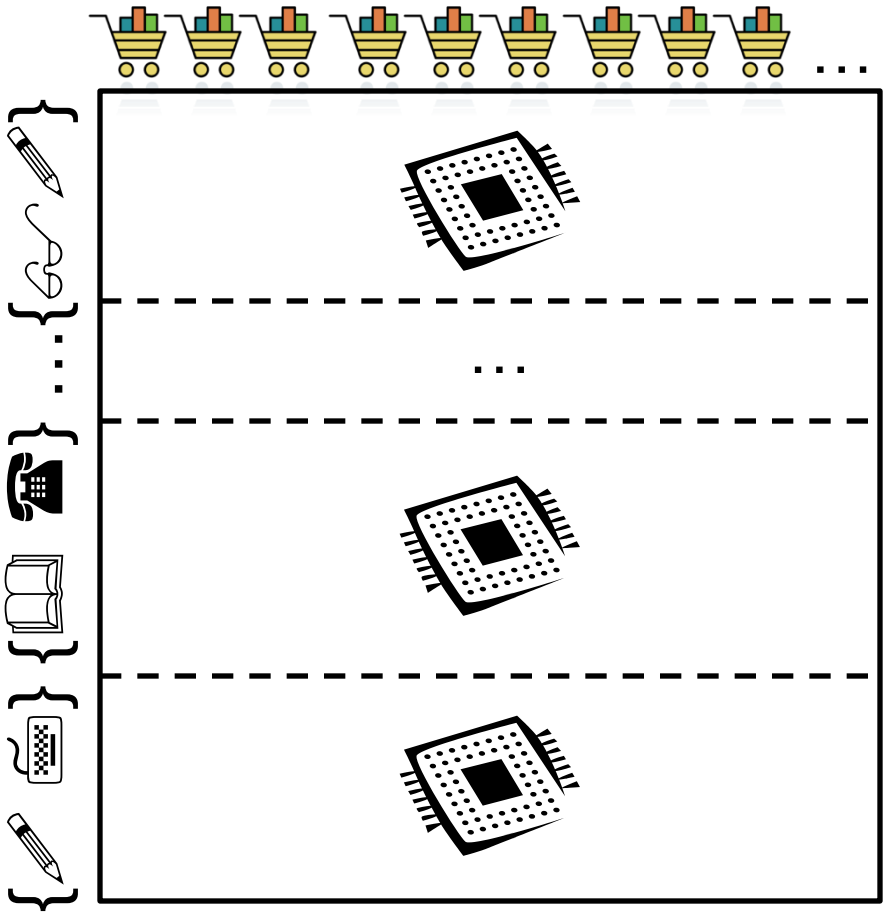


Распараллеливание алгоритма apriori

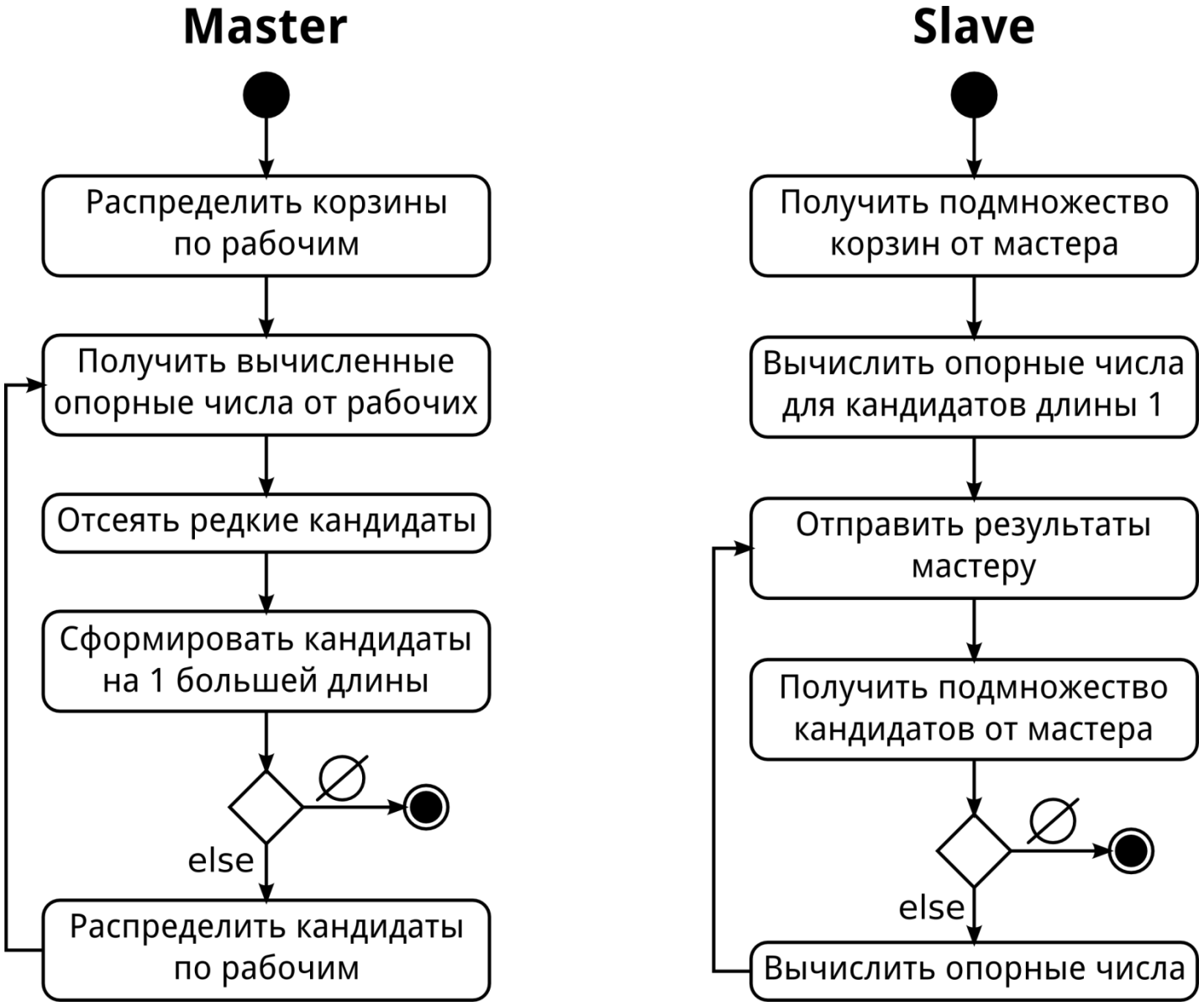
Count Distribution (CDapriori)



Data Distribution (DDapriori)

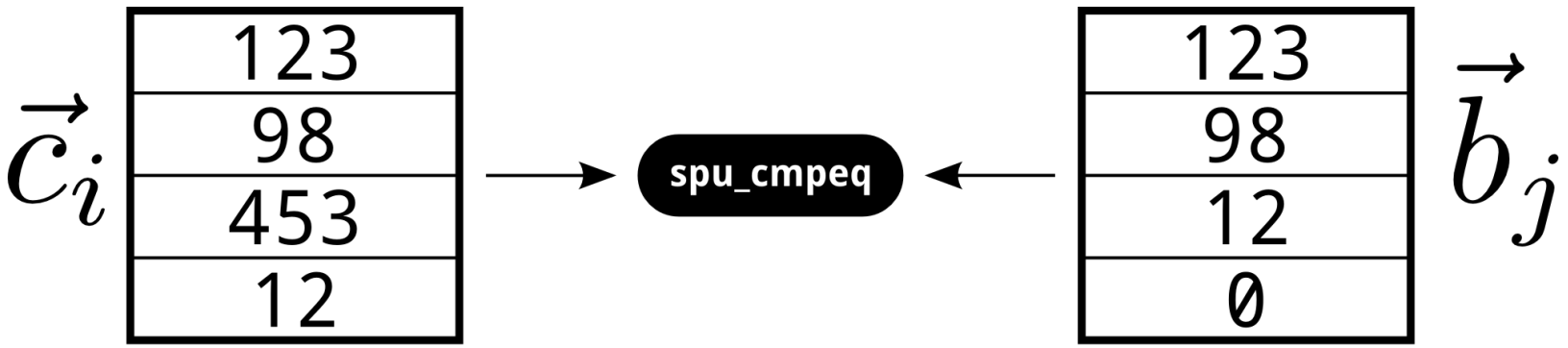


Параллельный алгоритм DDC-apriori

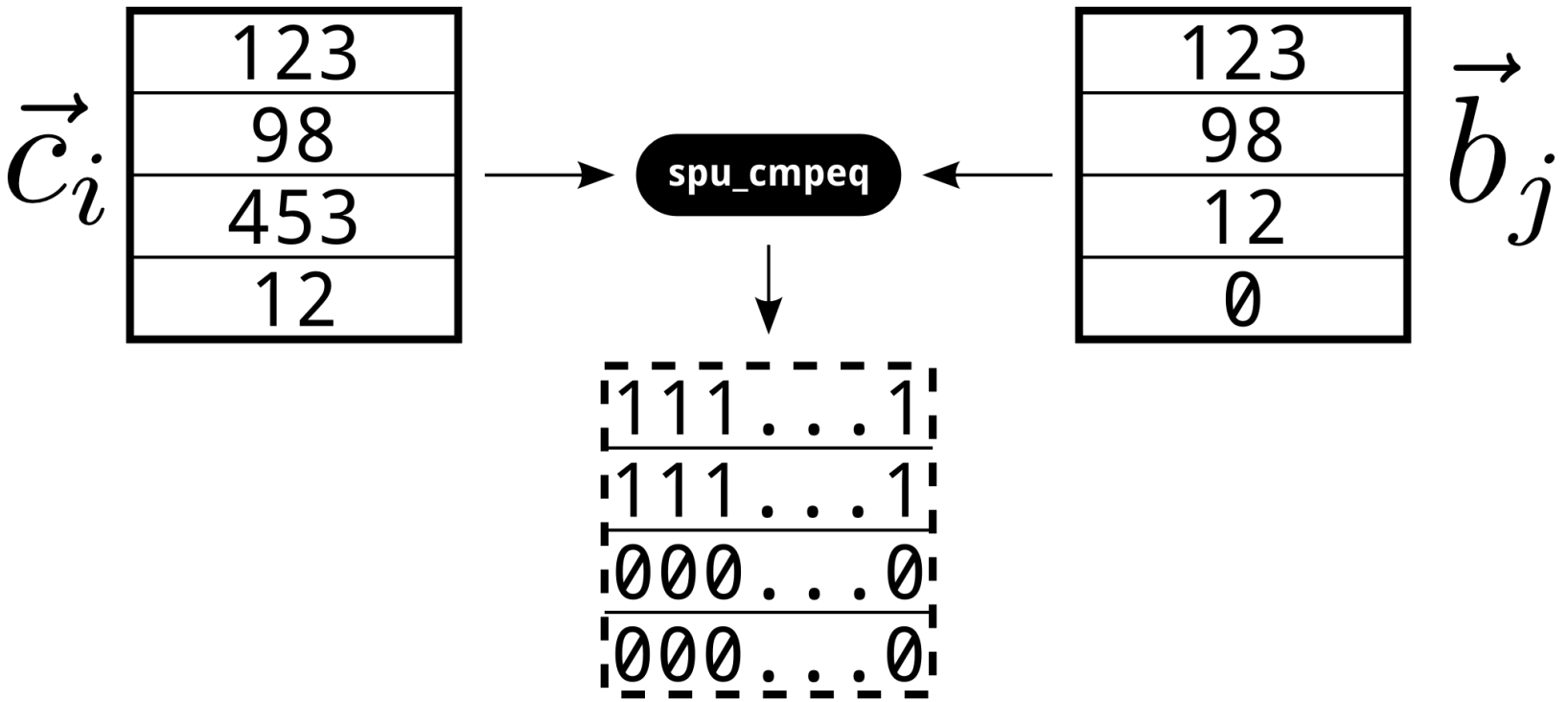


Использование SIMD-функций

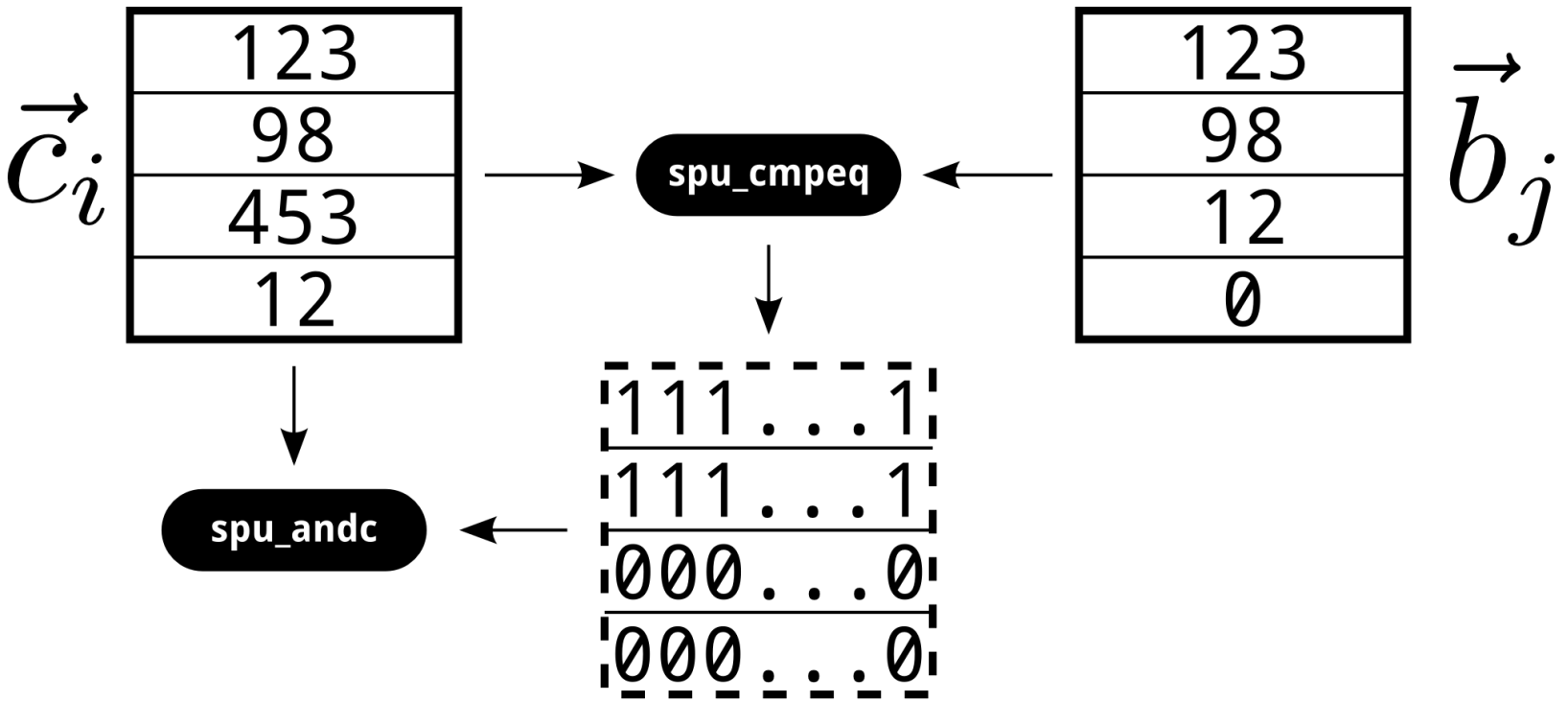
Использование SIMD-функций



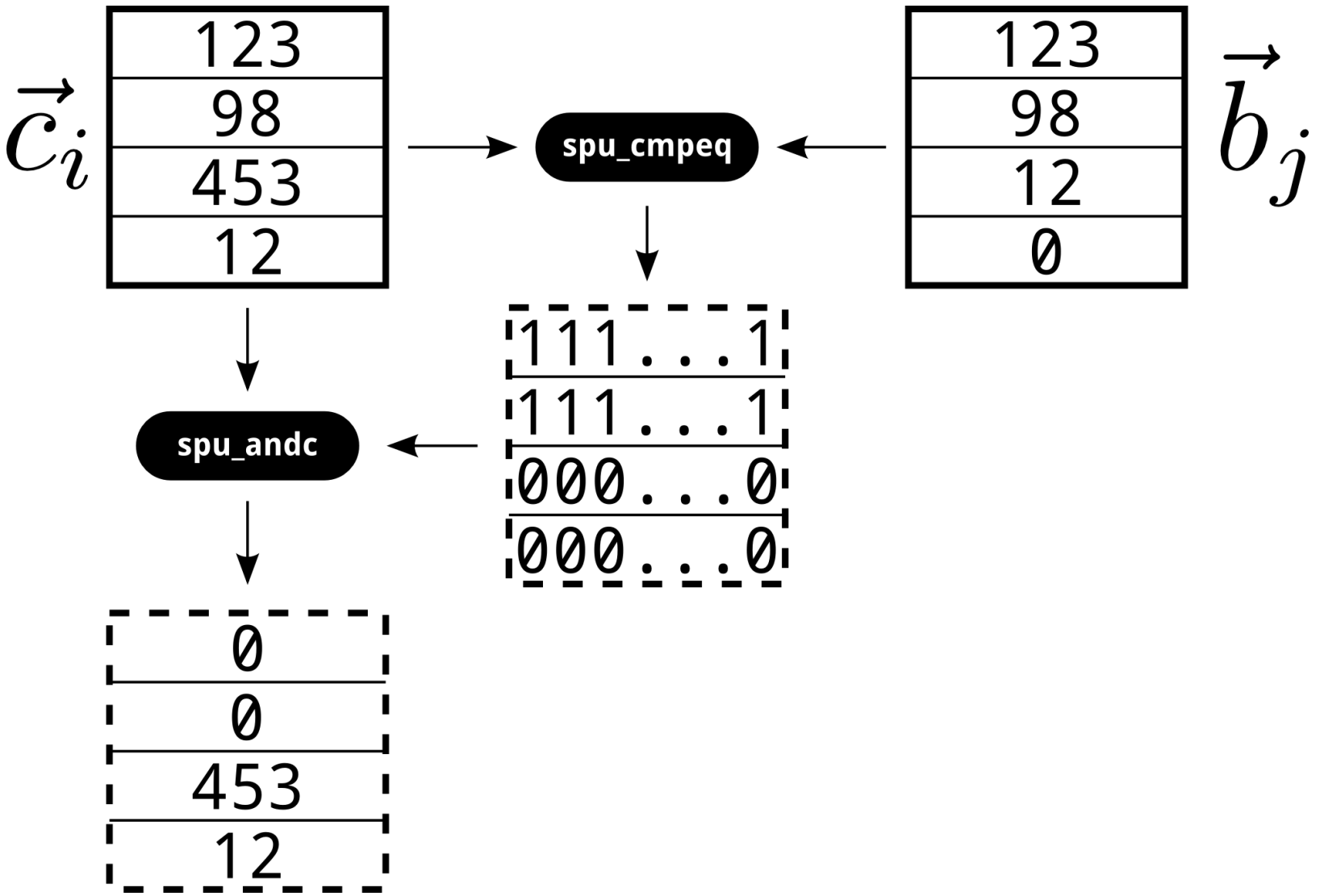
Использование SIMD-функций



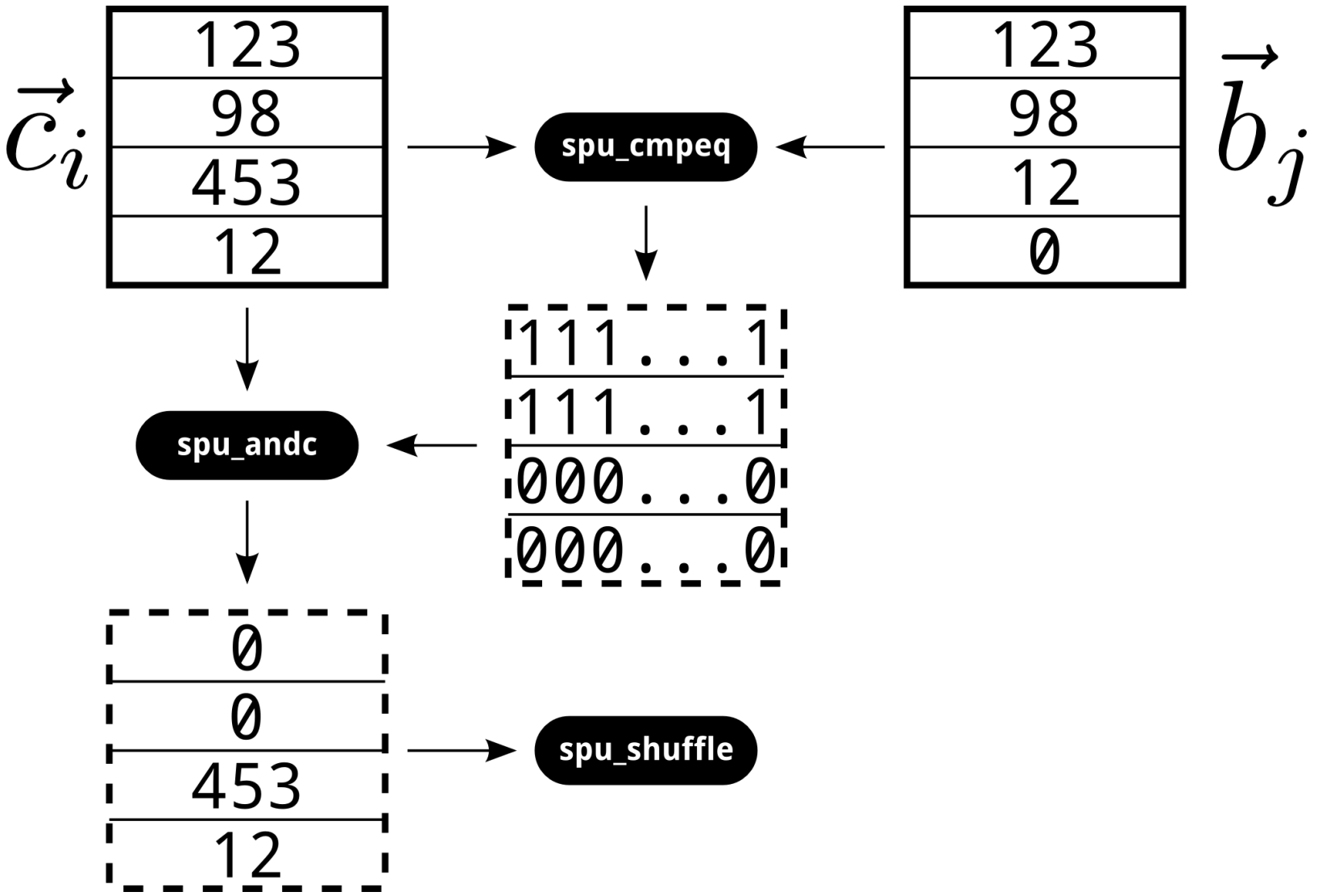
Использование SIMD-функций



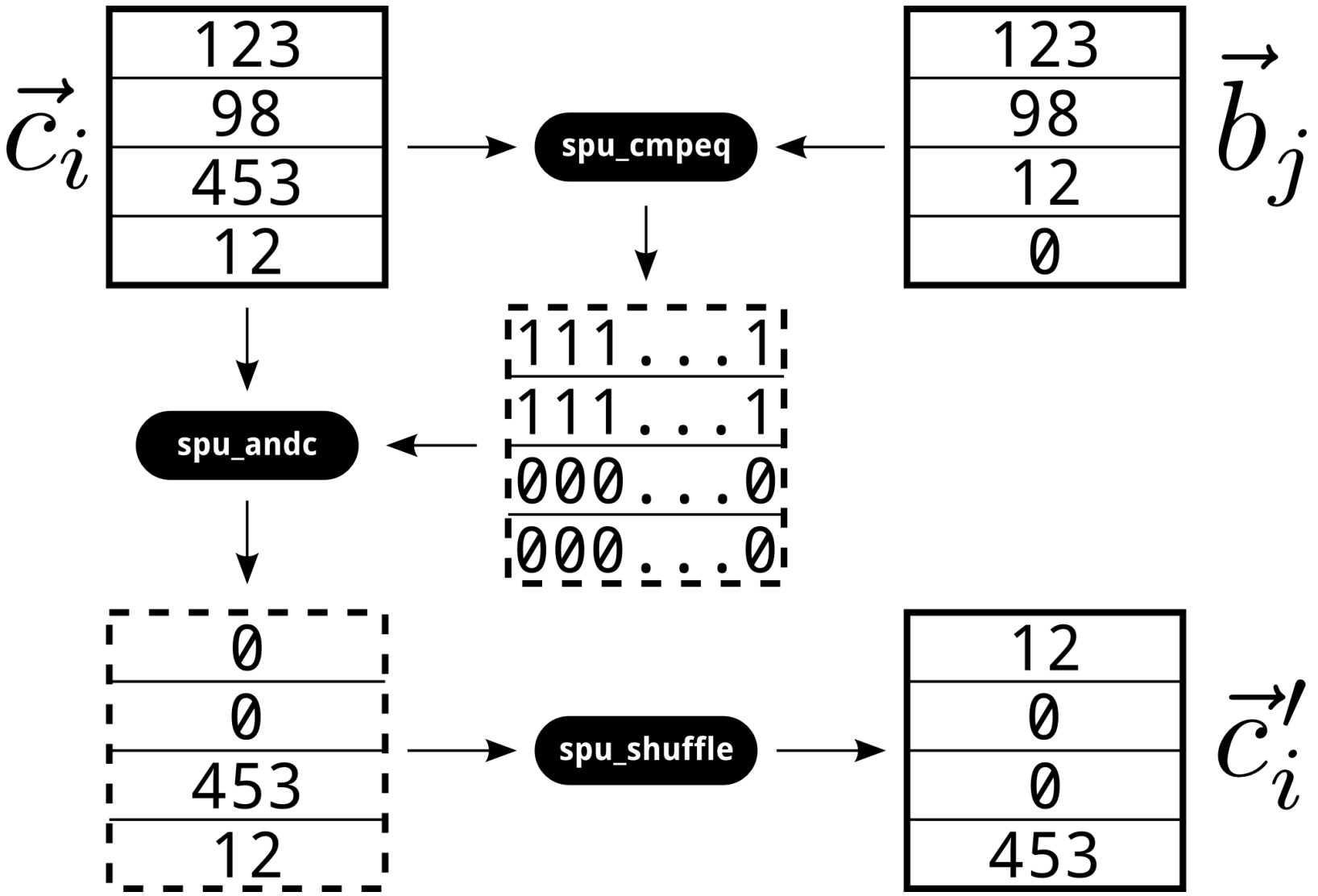
Использование SIMD-функций



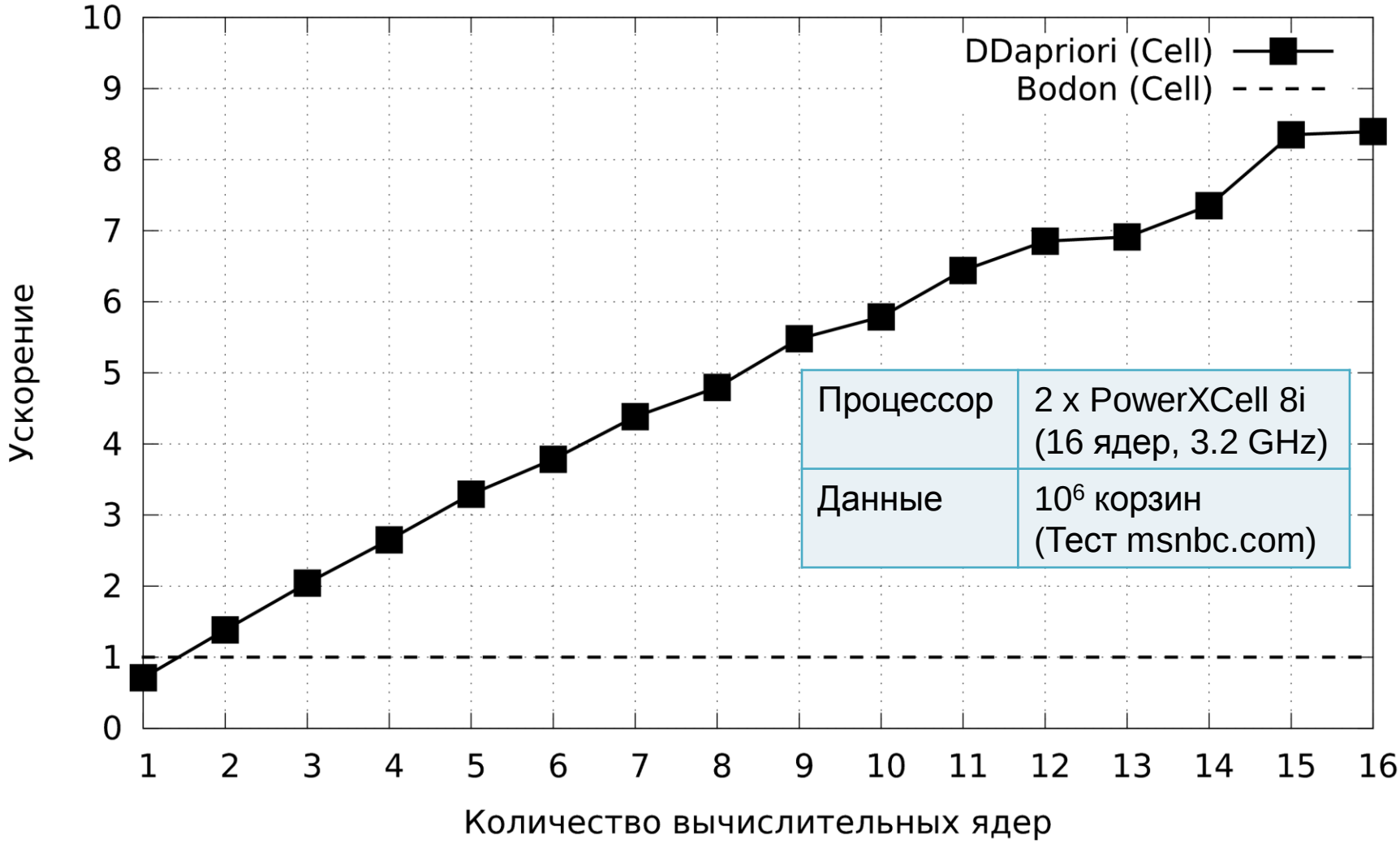
Использование SIMD-функций



Использование SIMD-функций

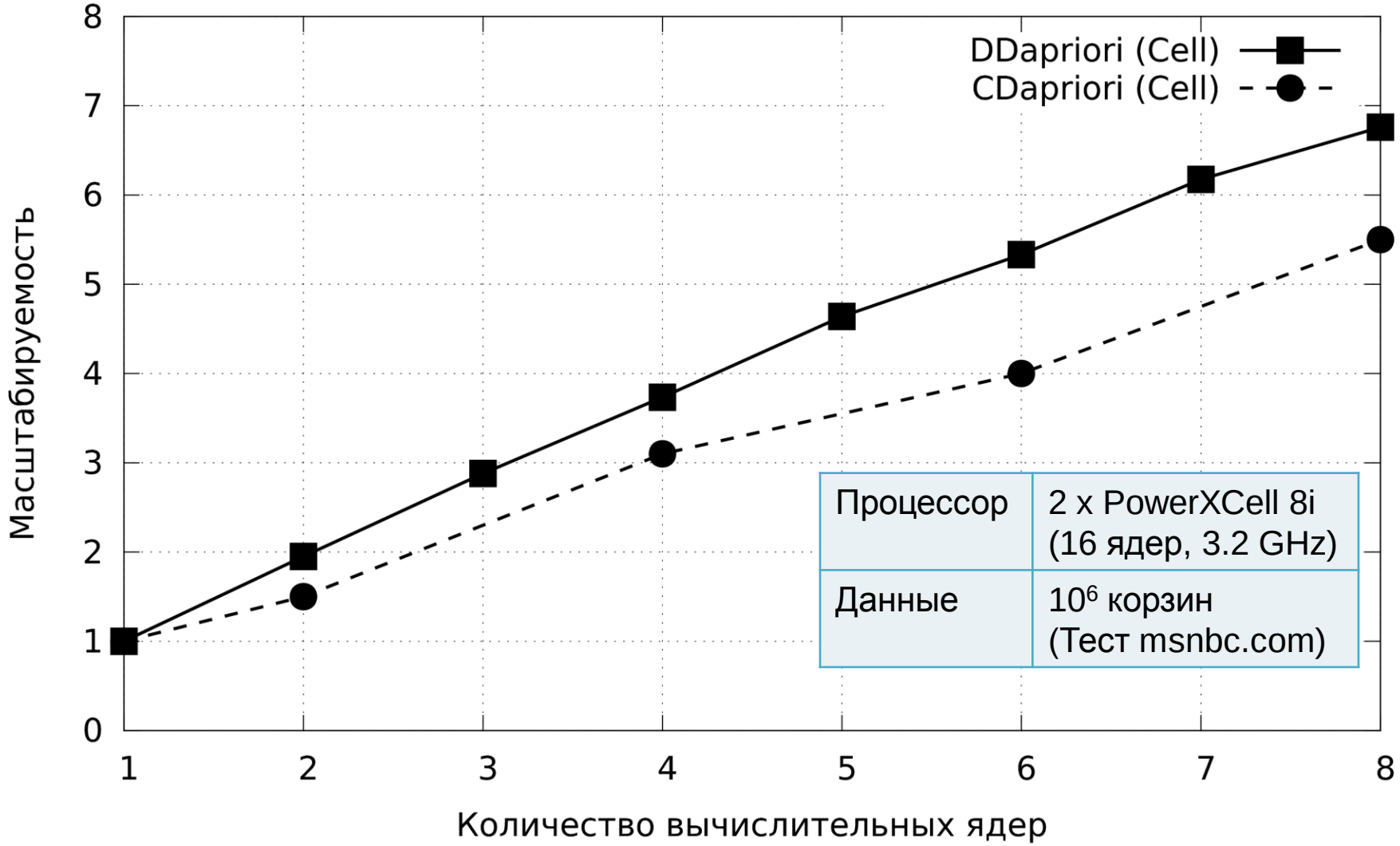


Эксперименты: ускорение

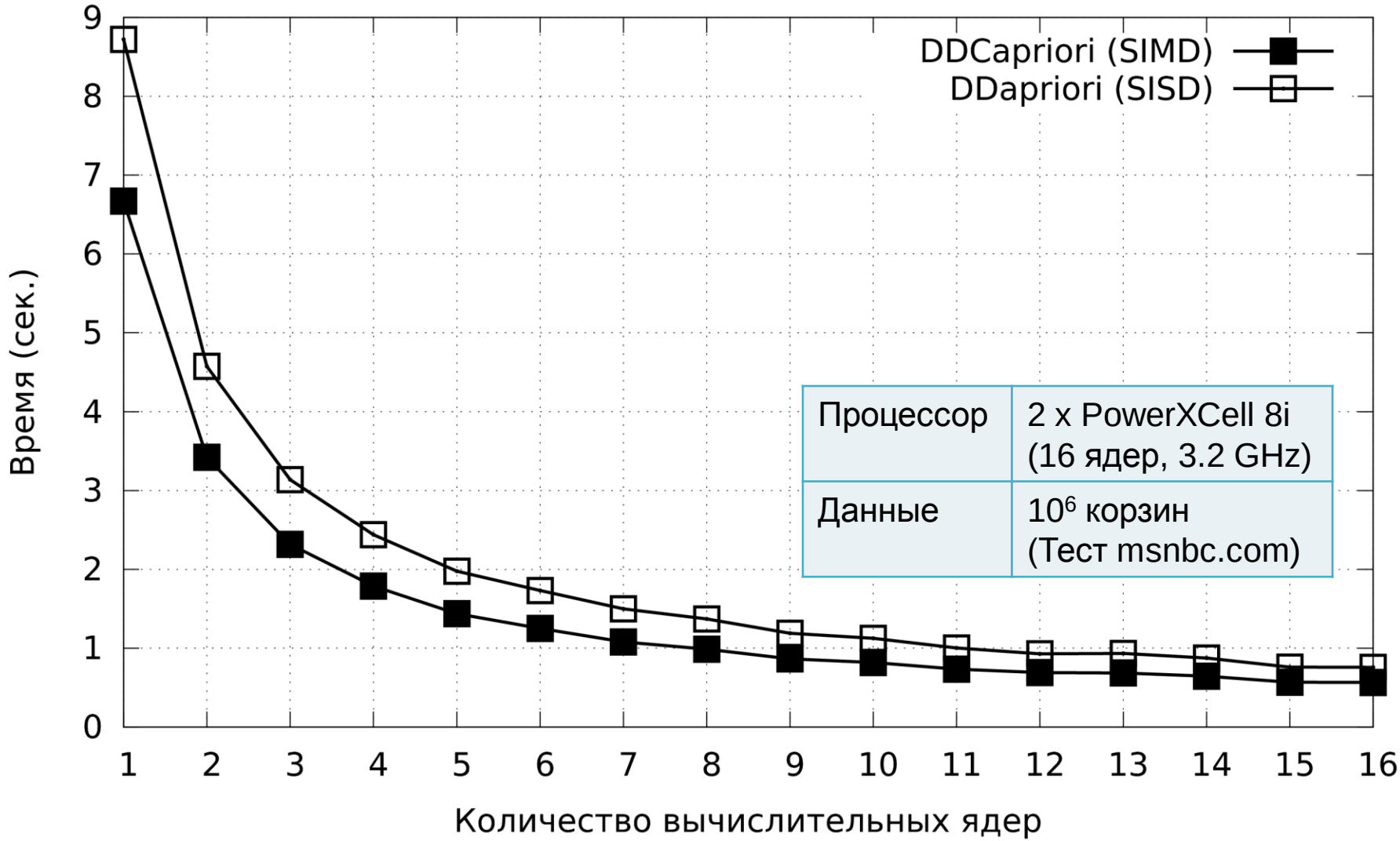


Процессор	2 x PowerXCell 8i (16 ядер, 3.2 GHz)
Данные	10^6 корзин (Тест msnbc.com)

Эксперименты: DDapriori vs CDapriori

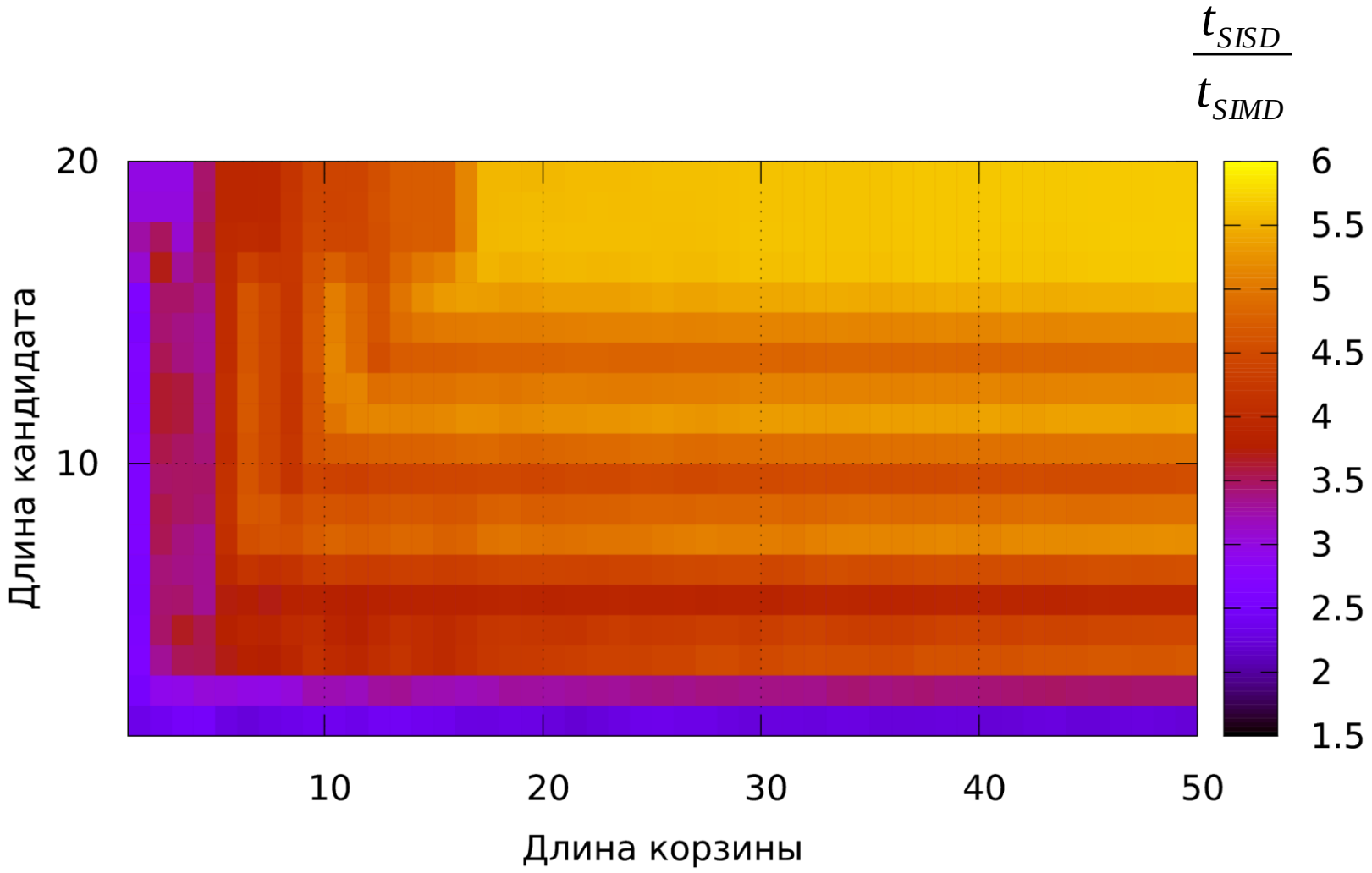


Эксперименты: SIMD vs SISD

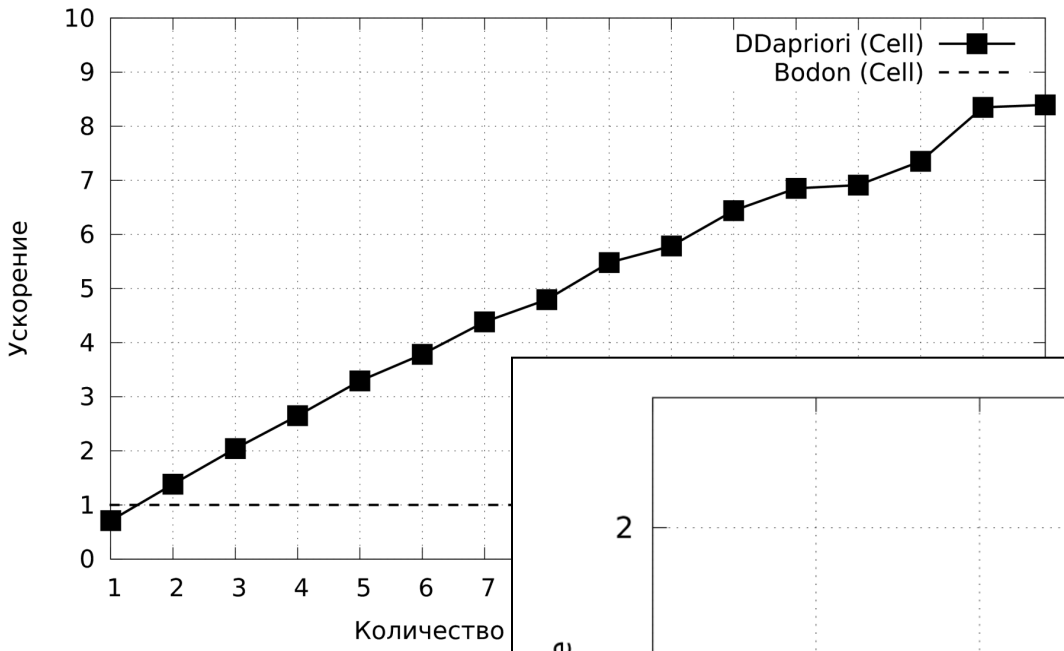


Процессор	2 x PowerXCell 8i (16 ядер, 3.2 GHz)
Данные	10^6 корзин (Тест msnbc.com)

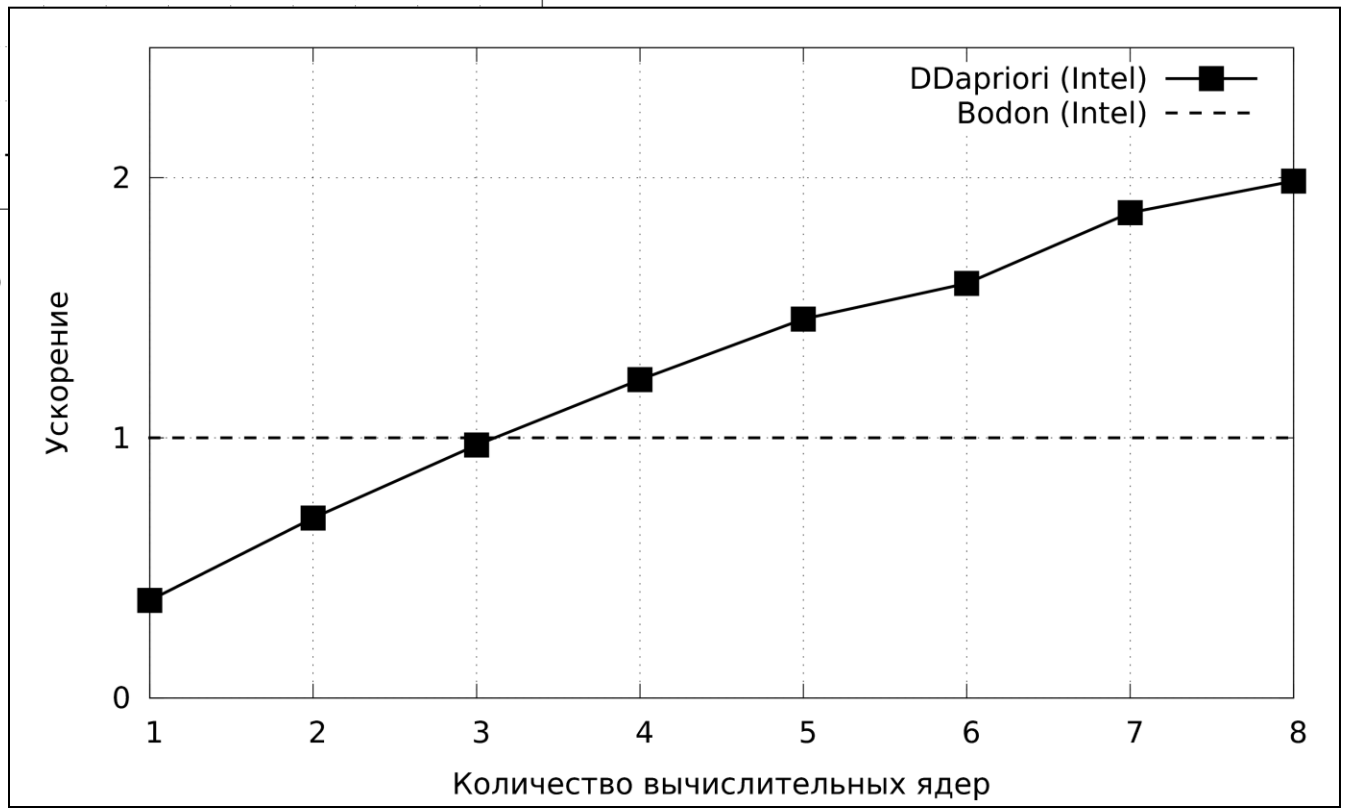
Эксперименты: влияние длины корзины



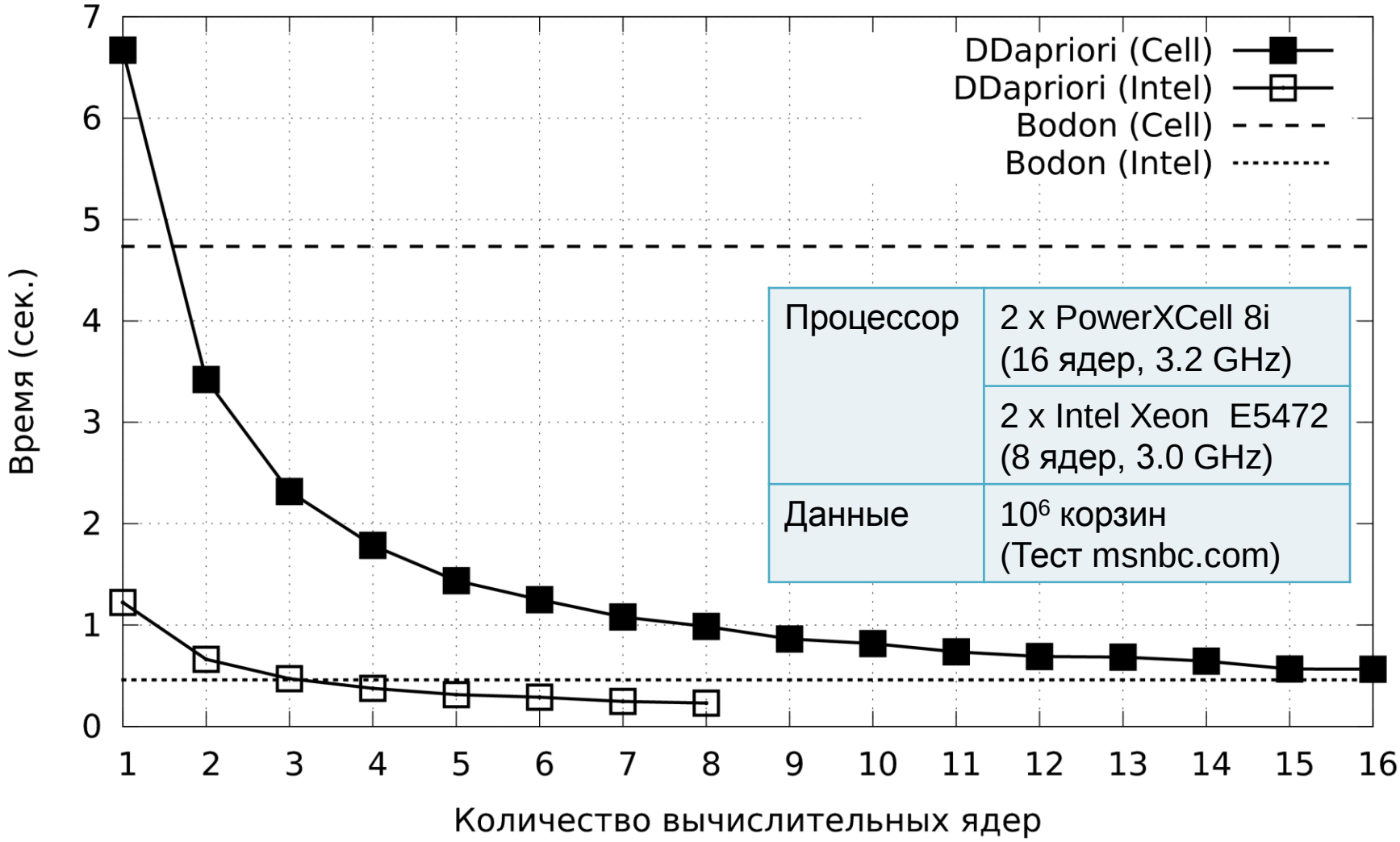
Эксперименты: Cell vs Intel, ускорение



Процессор	2 x PowerXCell 8i (16 ядер, 3.2 GHz)
	2 x Intel Xeon E5472 (8 ядер, 3.0 GHz)
Данные	10^6 корзин (Тест msnbc.com)



Эксперименты: Cell vs Intel, время



Процессор	2 x PowerXCell 8i (16 ядер, 3.2 GHz)
	2 x Intel Xeon E5472 (8 ядер, 3.0 GHz)
Данные	10^6 корзин (Тест msnbc.com)

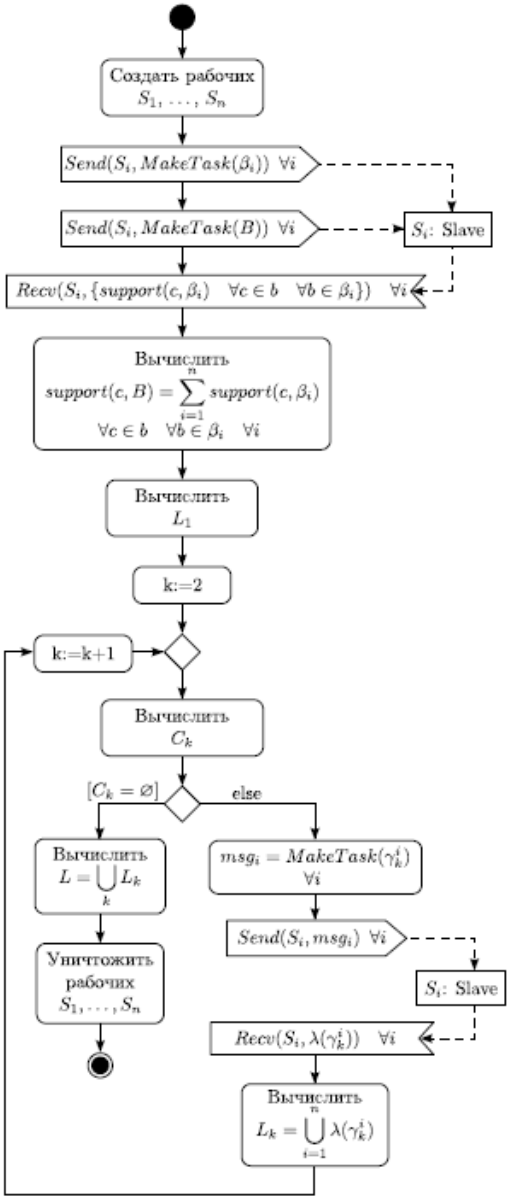
Спасибо за внимание

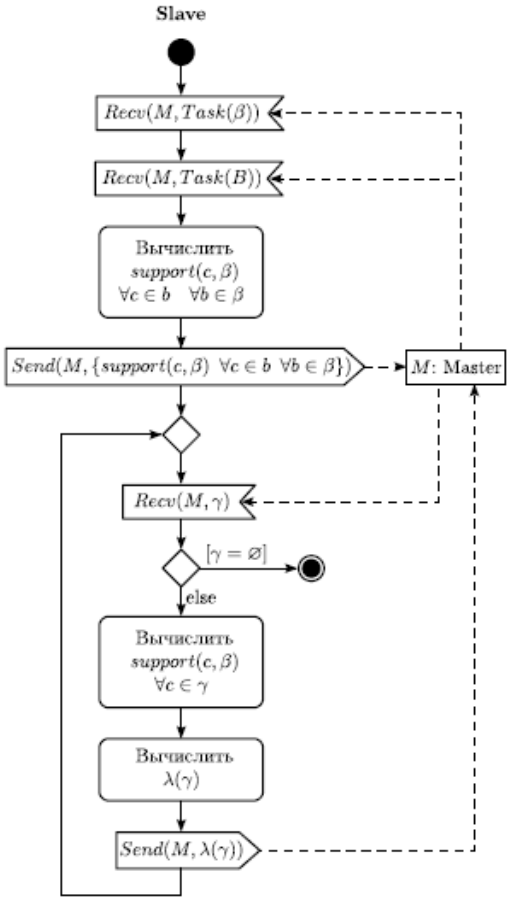
- Вопросы?
  - Константин Сергеевич Пан
    kpan@mail.ru
  - Михаил Леонидович Цымблер
    mzym@susu.ru

Дополнительные слайды

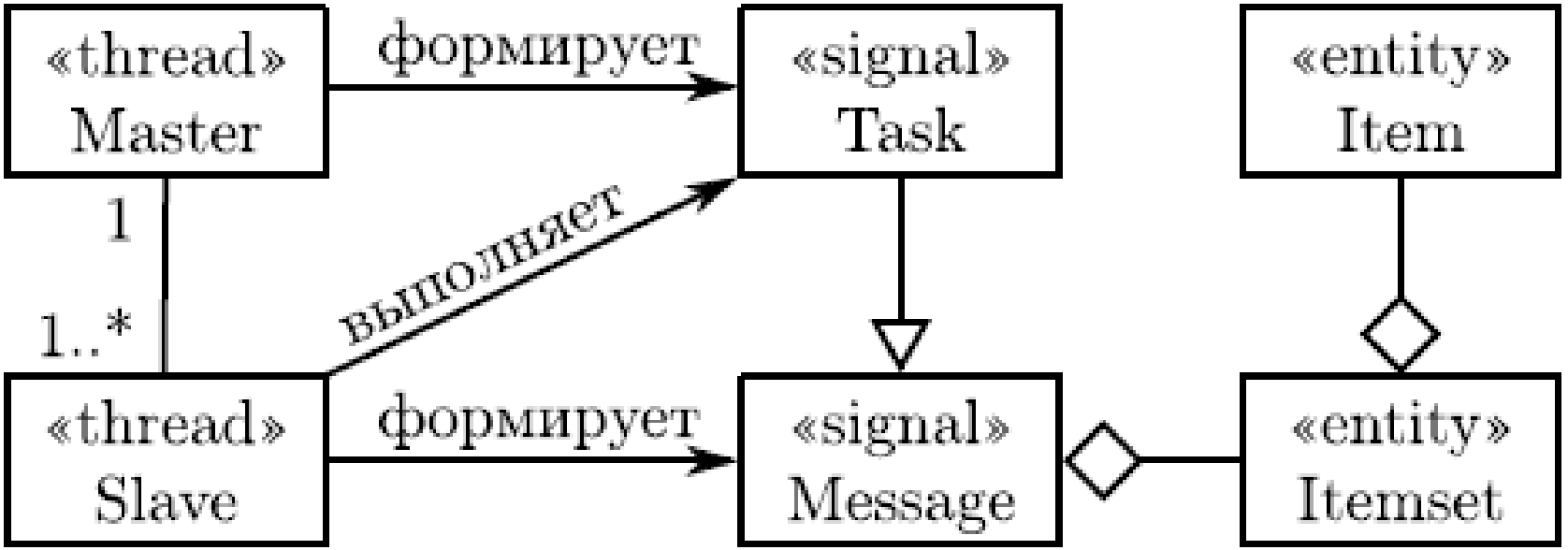
- DDC-apriori (без упрощений)
- DDC-apriori (детали реализации)

Параллельный алгоритм DDC-apriori





DDC-apriori: детали реализации



- Язык программирования: C
- Библиотеки: IBM Cell Broadband Engine SDK
- Объем исходных текстов: 1 тыс. строк