

Международная научная конференция  
Параллельные вычислительные технологии (ПавТ'2023)  
Санкт-Петербург, 28–30 марта 2023

---

# Автоматизированная разметка больших временных рядов на кластере с GPU узлами

А.И. Гоглачев, М.Л. Цымблер  
Южно-Уральский государственный университет (Челябинск)

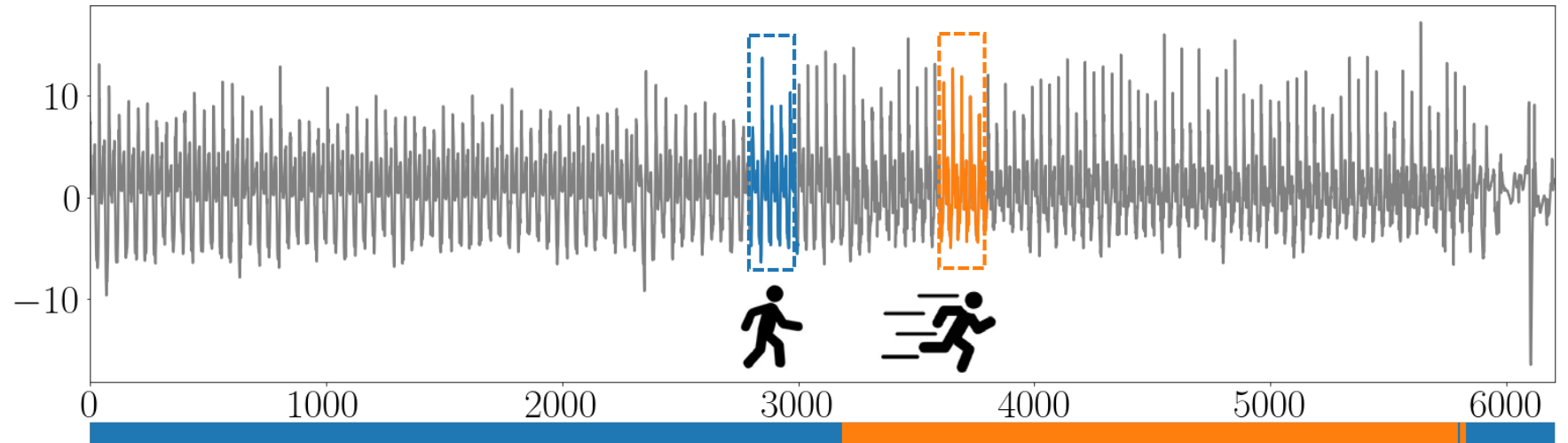
# Автоматизированная разметка временных рядов

- Персональная медицина

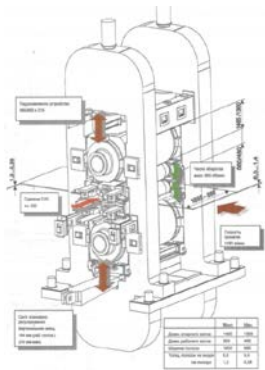


Акселерометр

Ускорение,  $m/c^2$

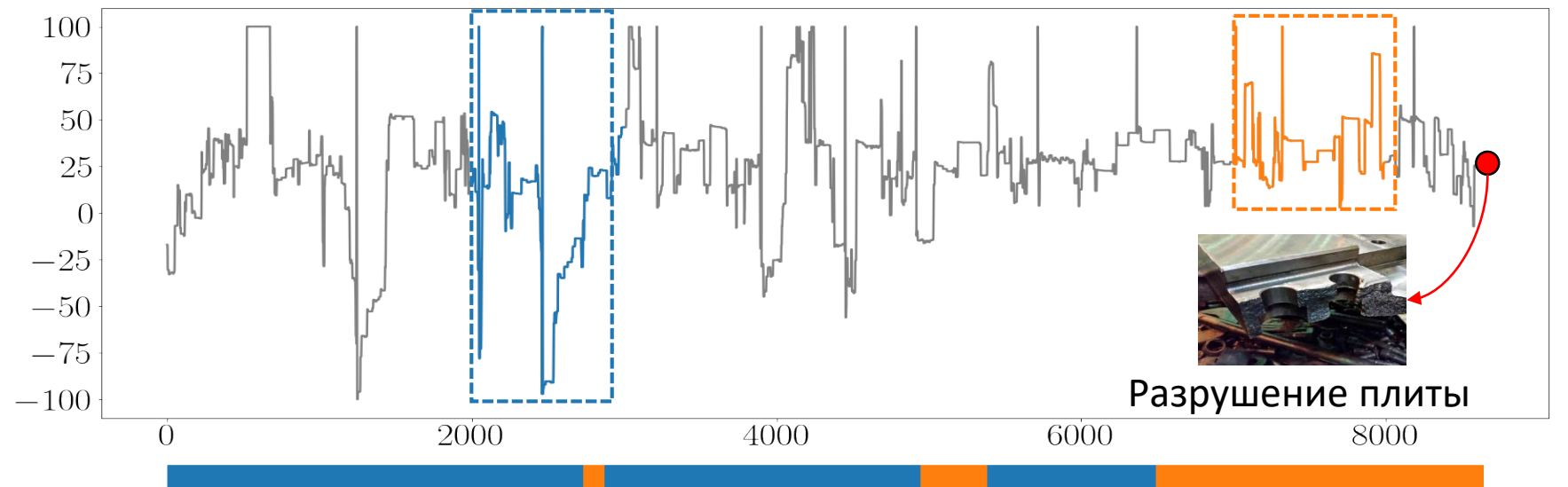


- Цифровая индустрия



Клеть стана

Перемещения гидроцилиндра, мм



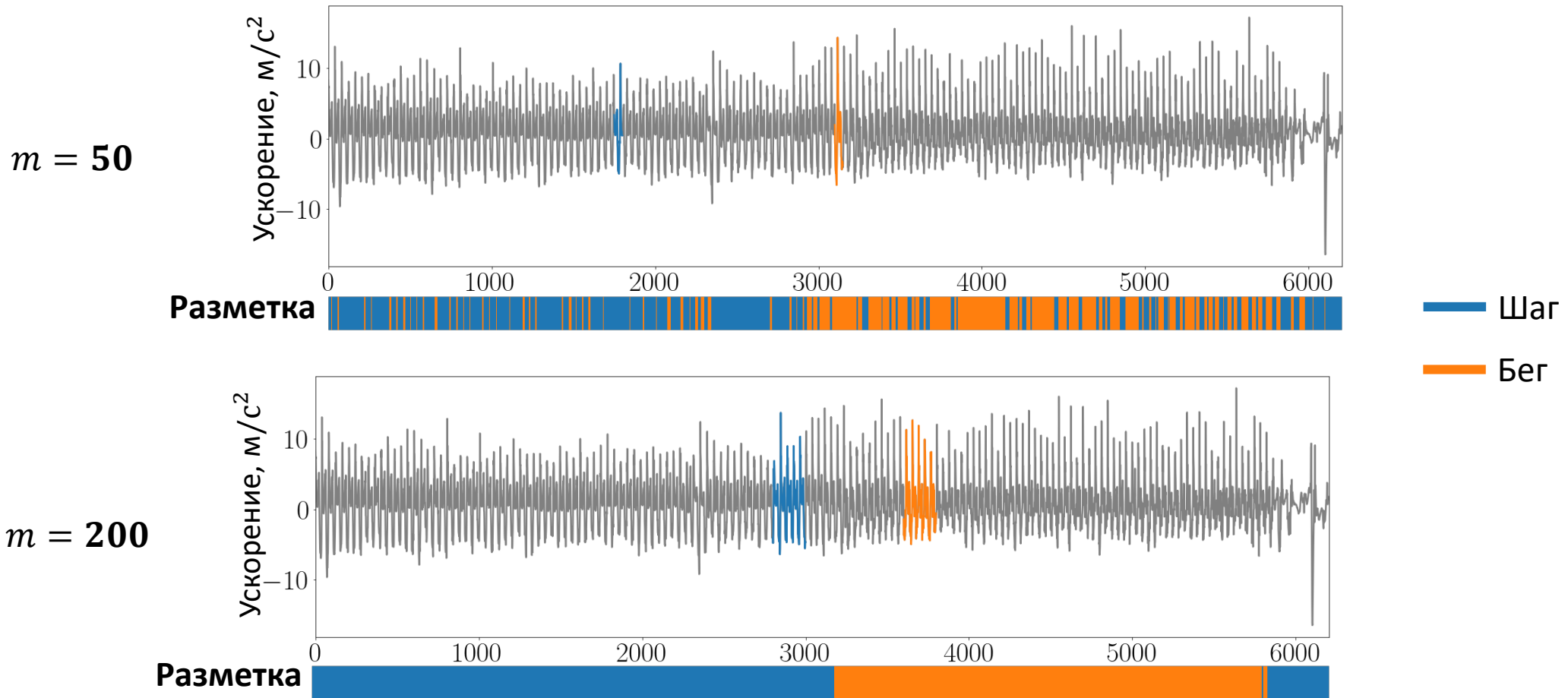
Разрушение плиты

# Обзор работ

Базовая концепция	Алгоритм	Недостаток
<b>Последовательные алгоритмы</b>		
Шейплеты	Yeh, C.M., et. al.: Matrix profile I: all pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. IEEE 16th ICDM, pp. 1317–1322.	Находят лишь типичные подпоследовательности (не находят степень схожести с остальными подпоследовательностями)
Лейтмотивы	Mueen, A., et. al. Exact discovery of time series motifs. Proceedings of the SIAM International Conference on Data Mining. pp. 473–484	
Сниппеты	Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020).	Кубическая временная сложность от длины временного ряда
<b>Параллельные алгоритмы</b>		
Лейтмотивы	Narang A., Bhattacharjee S. Parallel Exact Time Series Motif Discovery. Proceedings of the 16th International Euro-Par Conference. Springer, 2010. P. 304–315.	Находит лишь типичные подпоследовательности

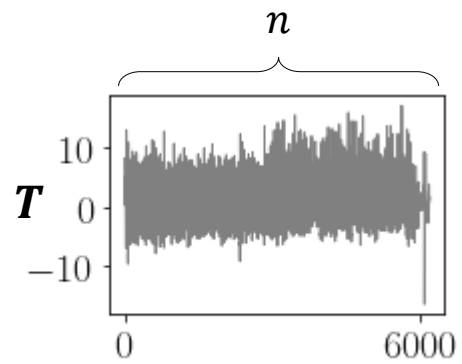
# Проблемы алгоритма PSF

## 1. PSF требует ручного подбора длины снippetа $m$

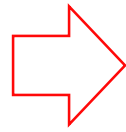


## 2. PSF не работает на GPU кластере

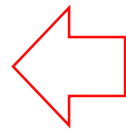
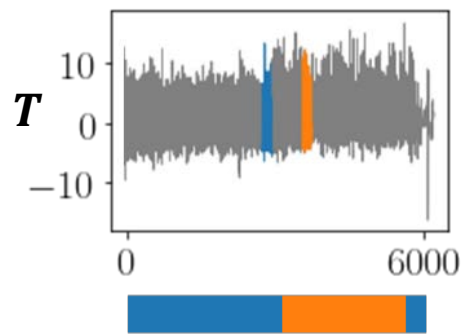
# Автоматизированная разметка на GPU кластере



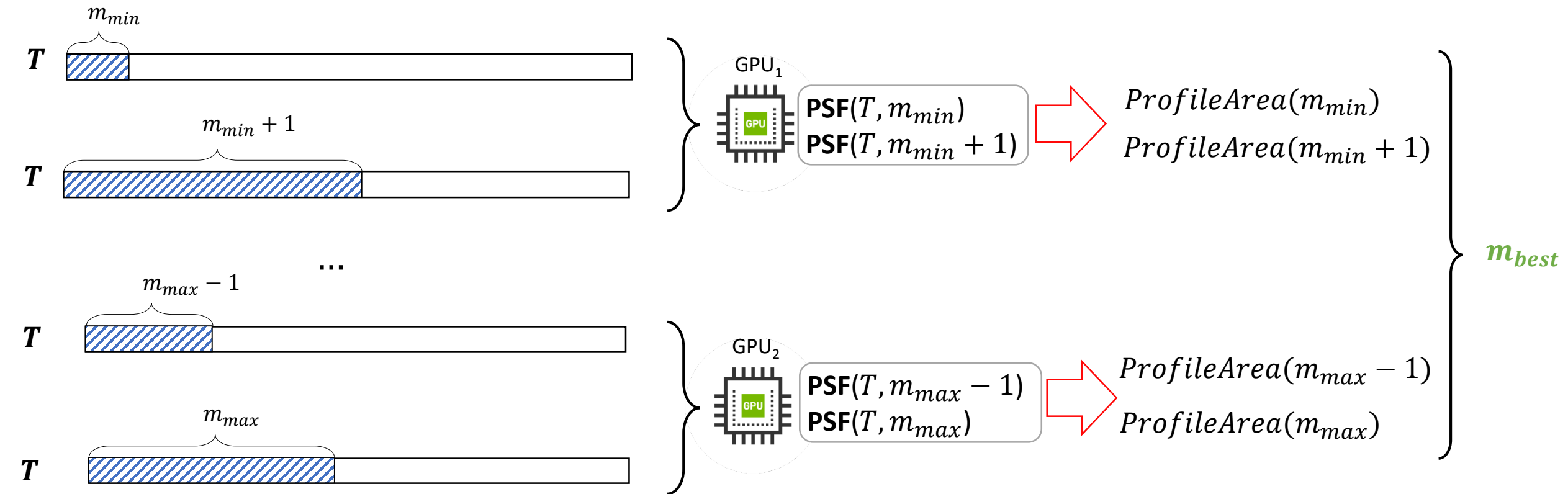
$m_{min} \dots m_{max}$



$m_{best} = 200$

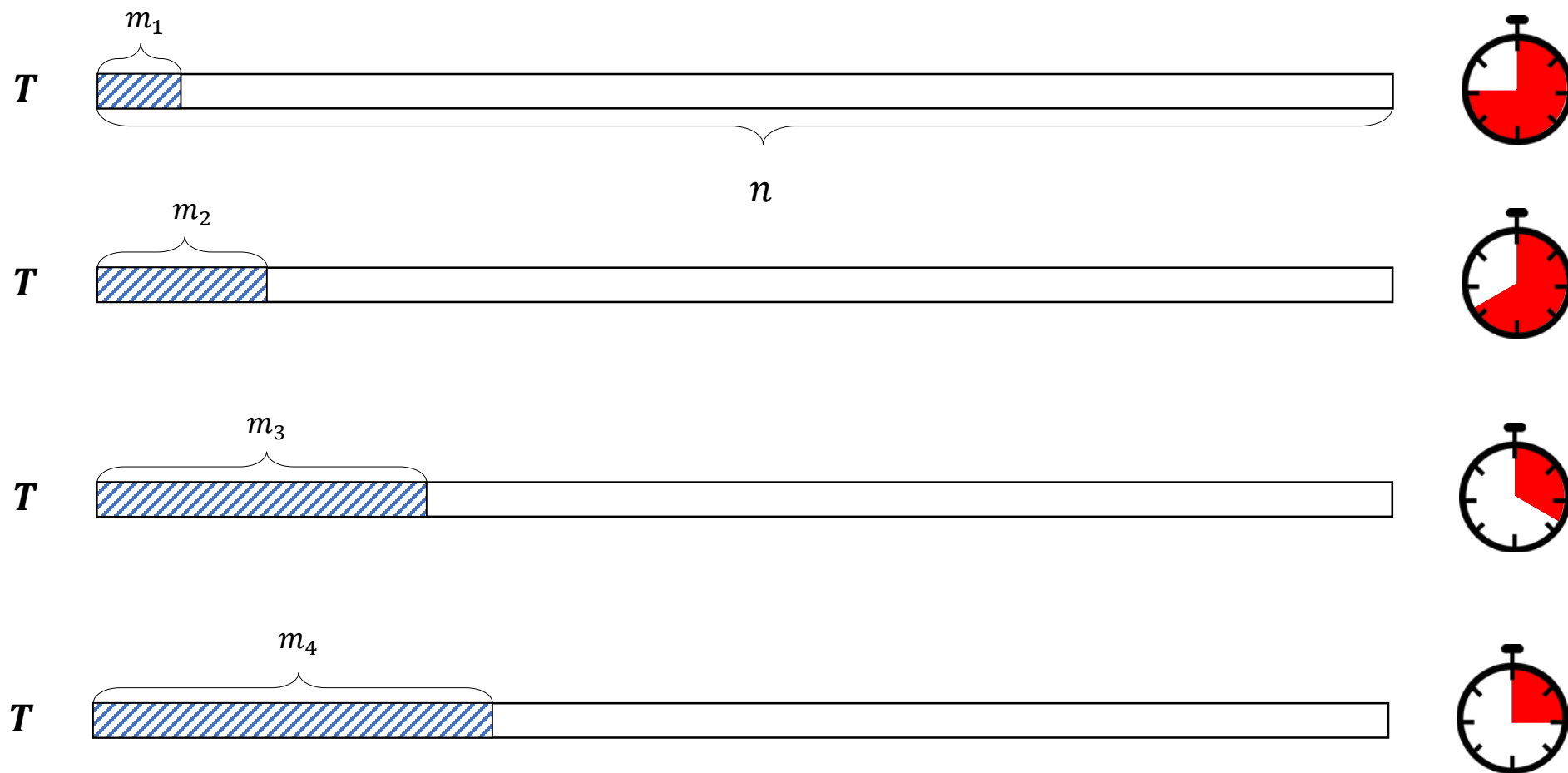


# Параллельный подбор $m_{best}$

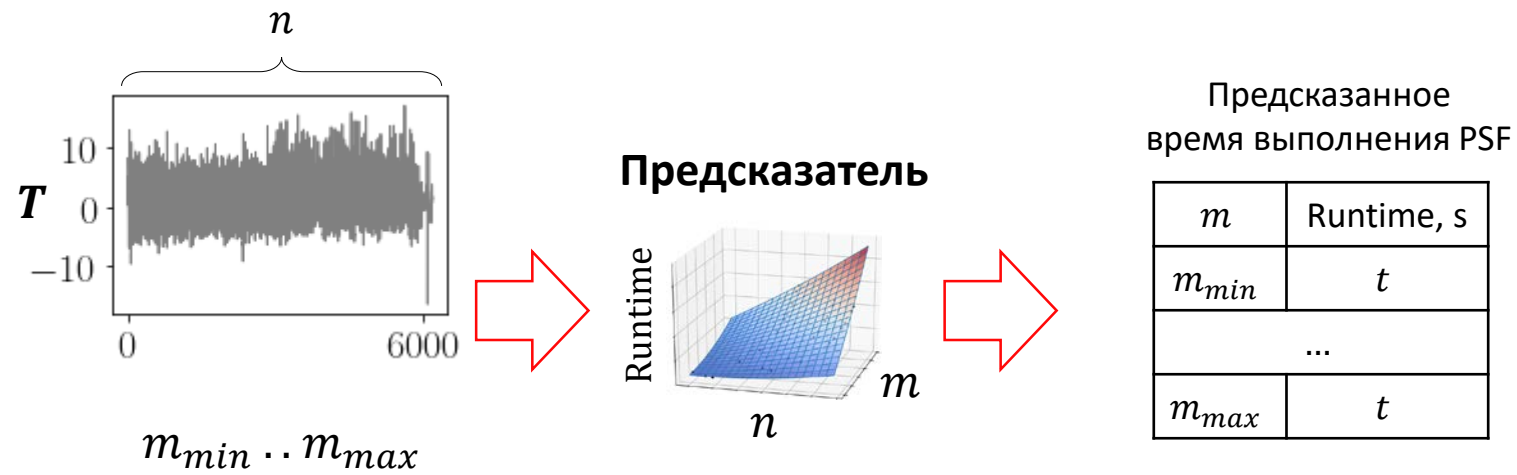


# Время выполнения PSF

Время выполнения PSF



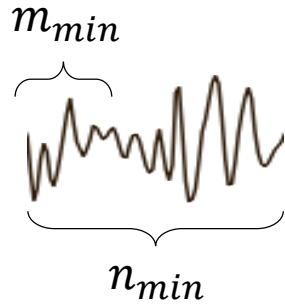
# Автоматизированная разметка на GPU кластере



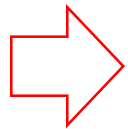


# Предсказатель (полиномиальная регрессия)

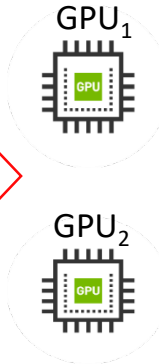
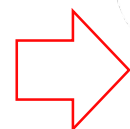
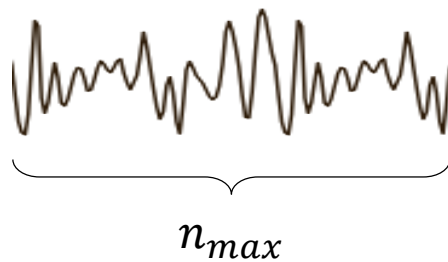
Синтетические данные



Генератор запусков PSF



...

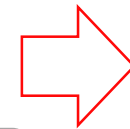


Запуски PSF

PSF( $n_{min}, m_{min}$ )

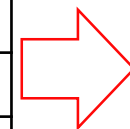
...

PSF( $n_{max}, m_{max}$ )

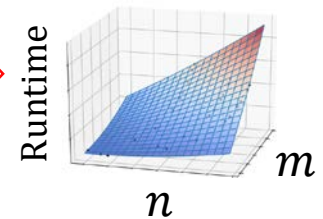


Обучающая выборка

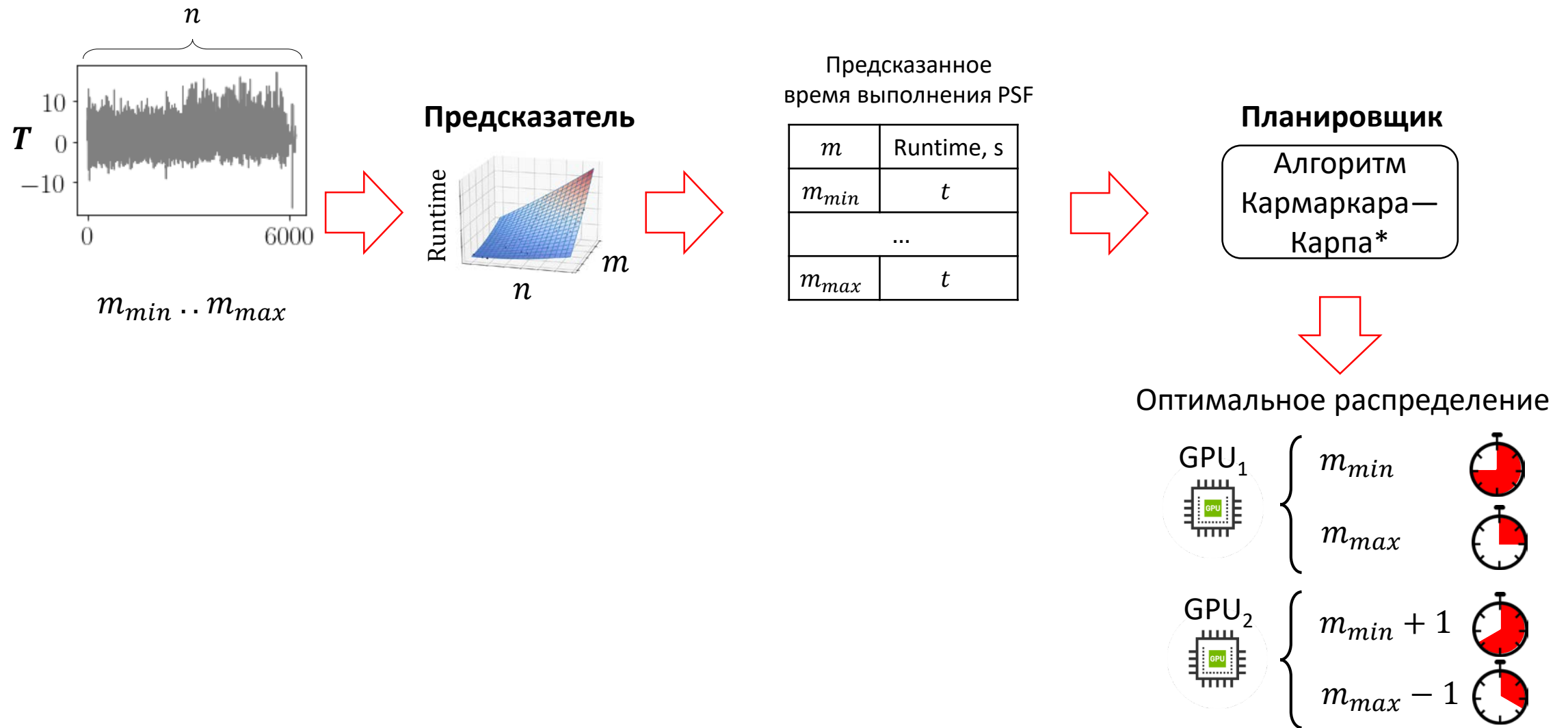
$n$	$m$	Runtime, s
$n_{min}$	$m_{min}$	$t$
...		
$n_{max}$	$m_{max}$	$t$



Предсказатель



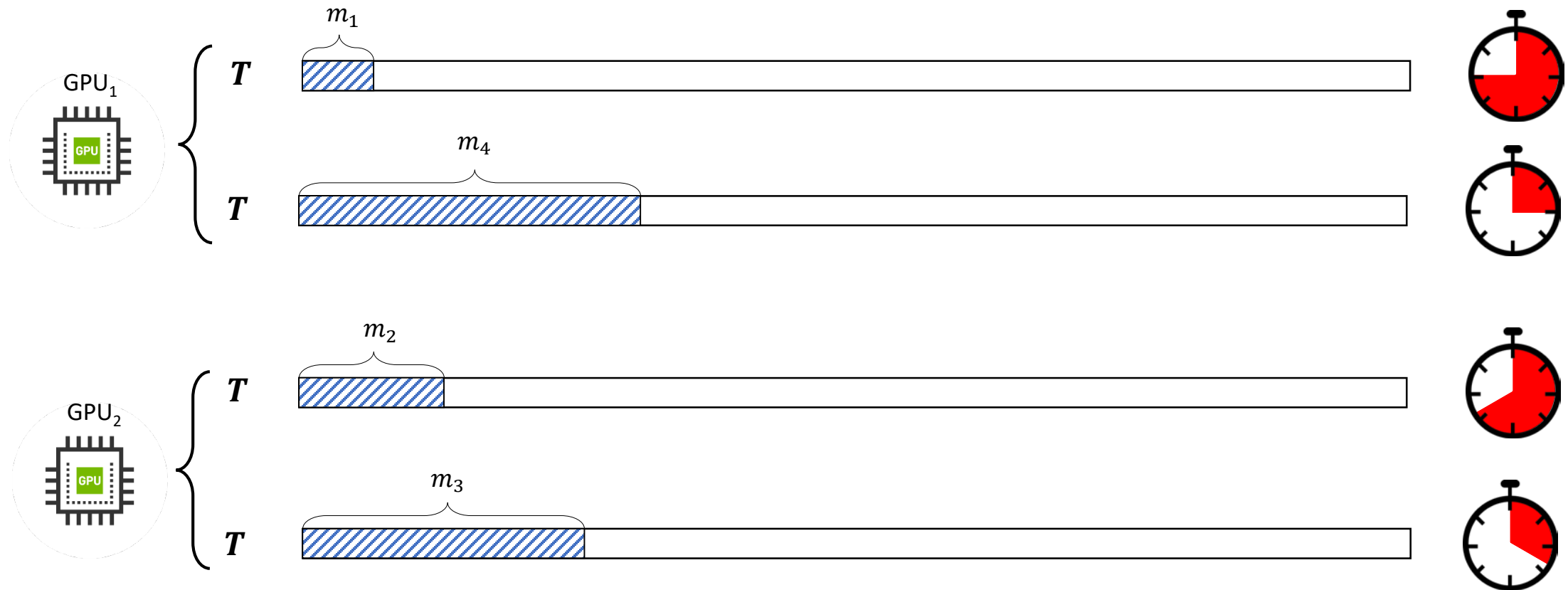
# Автоматизированная разметка на GPU кластере



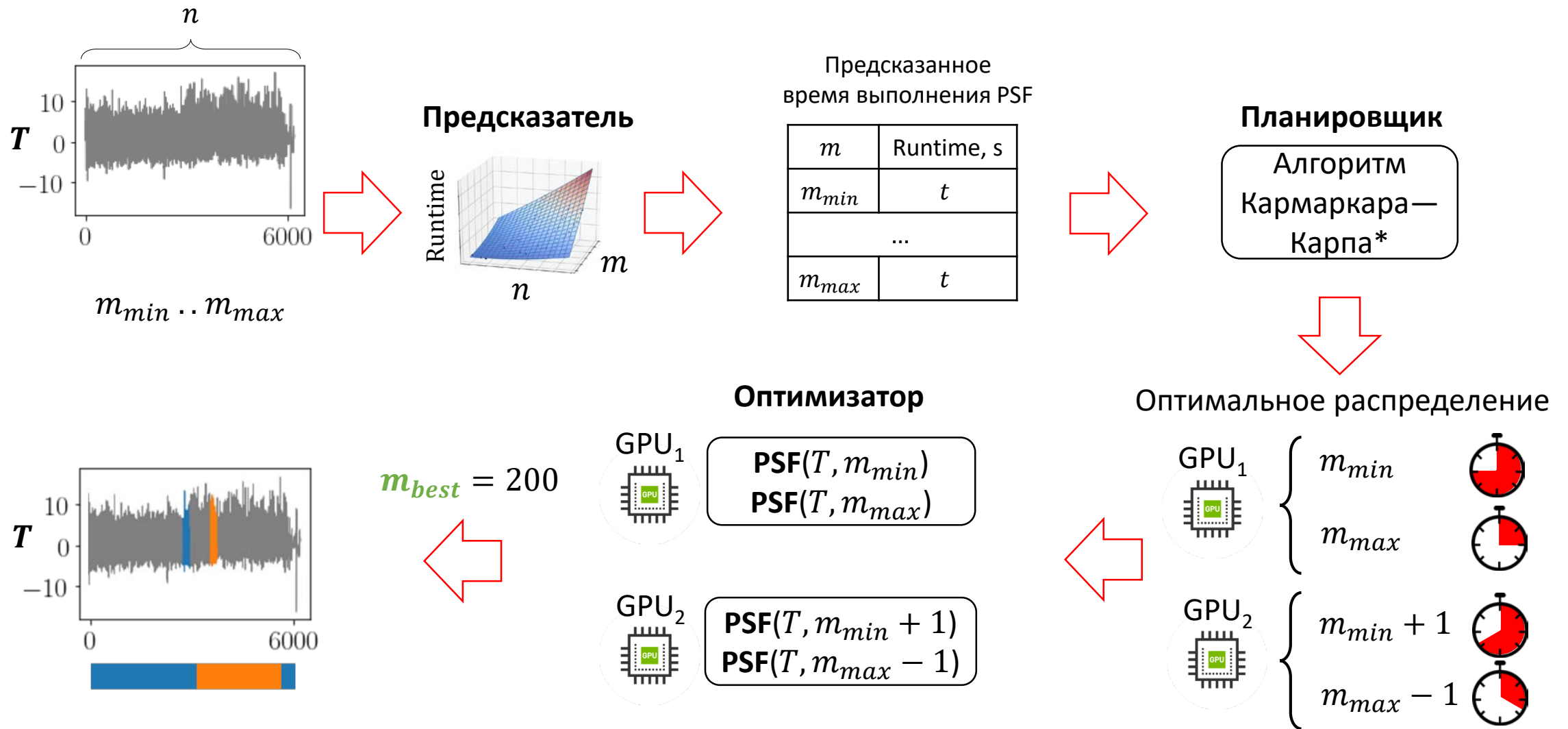
\*Narendra Karmarkar and Richard M. Karp, "The differencing method of set partitioning", Tech report UCB/CSD 82/113, Computer science division, University of California, Berkeley, 1982

# Оптимальное распределение длин снимков

Время выполнения PSF

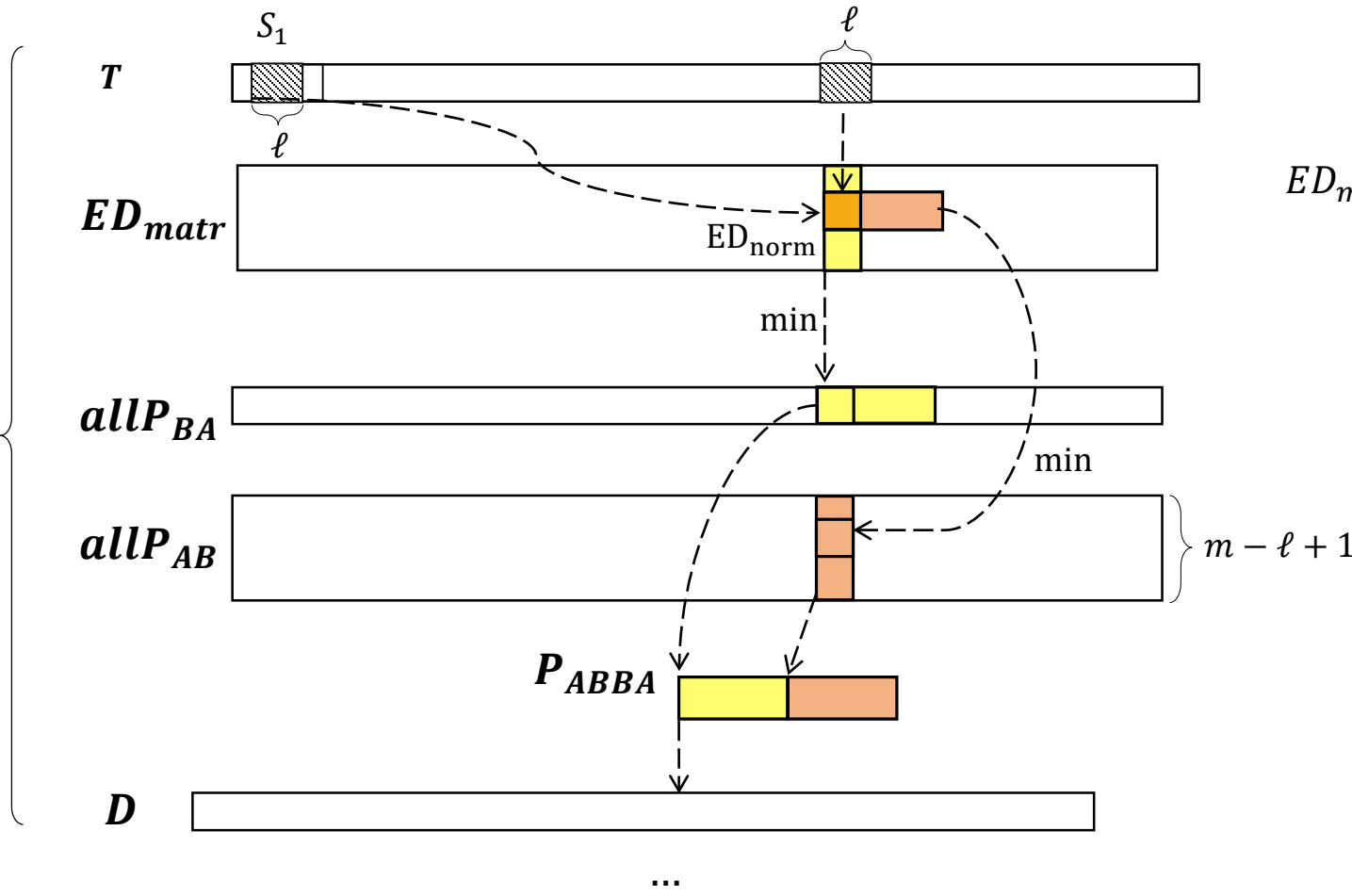
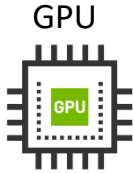


# Автоматизированная разметка на GPU кластере



\*Narendra Karmarkar and Richard M. Karp, "The differencing method of set partitioning", Tech report UCB/CSD 82/113, Computer science division, University of California, Berkeley, 1982

# Работа оптимизатора на одном узле



$$ED_{matr}(S, T, \ell) = \begin{pmatrix} d_{1,1} & \dots & d_{1,n-\ell+1} \\ \dots & \ddots & \dots \\ d_{m-\ell+1,1} & \dots & d_{m-\ell+1,n-\ell+1} \end{pmatrix},$$

$$d_{i,j} = ED^2(S_{i,\ell}, T_{j,\ell})$$

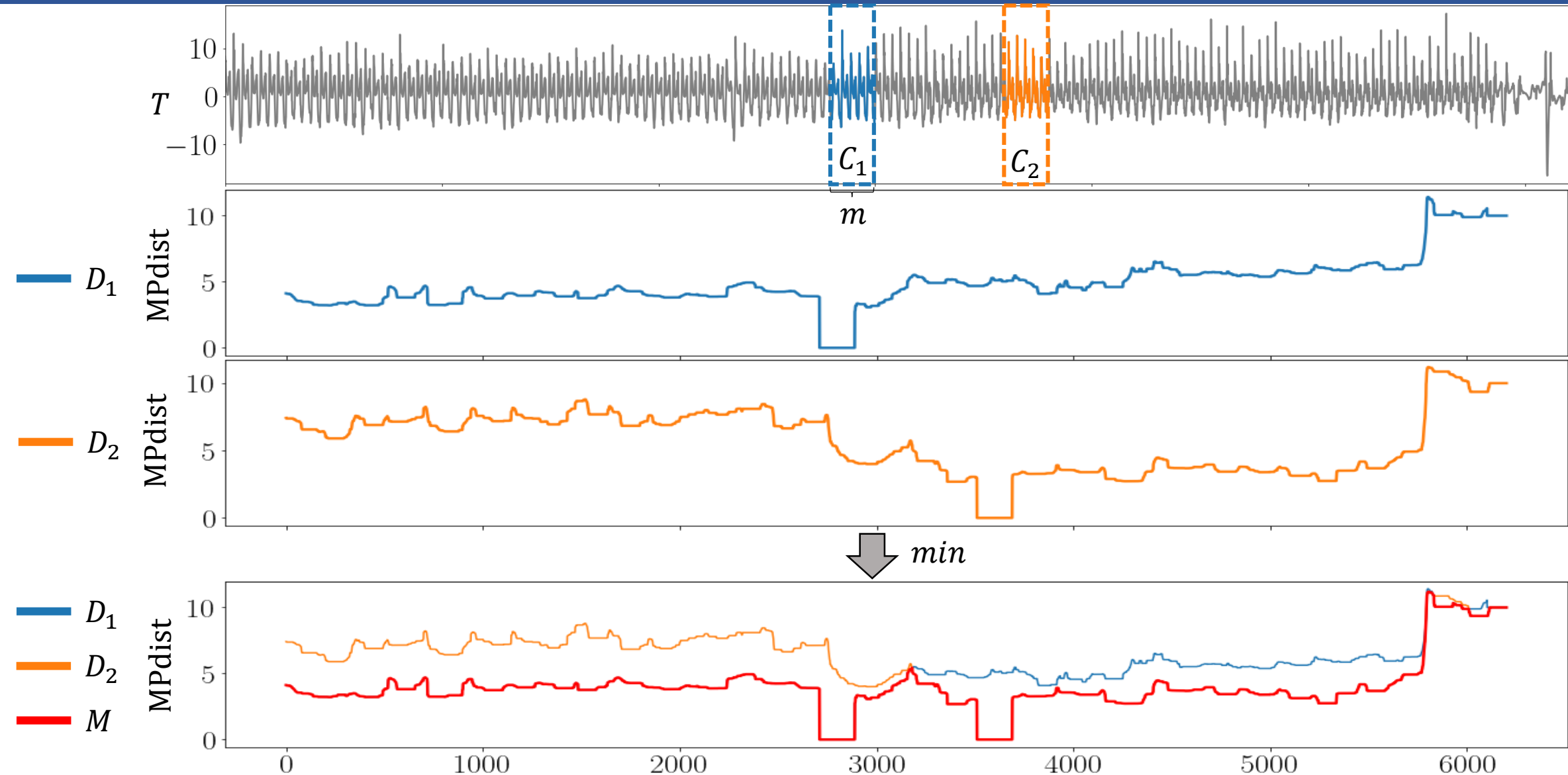
$$allP_{BA}(j) = \min_{1 \leq i \leq m-\ell+1} ED_{matr}(i, j)$$

$$allP_{AB}(i, j) = \min_{j \leq c \leq j+m-\ell+1} ED_{matr}(i, c)$$

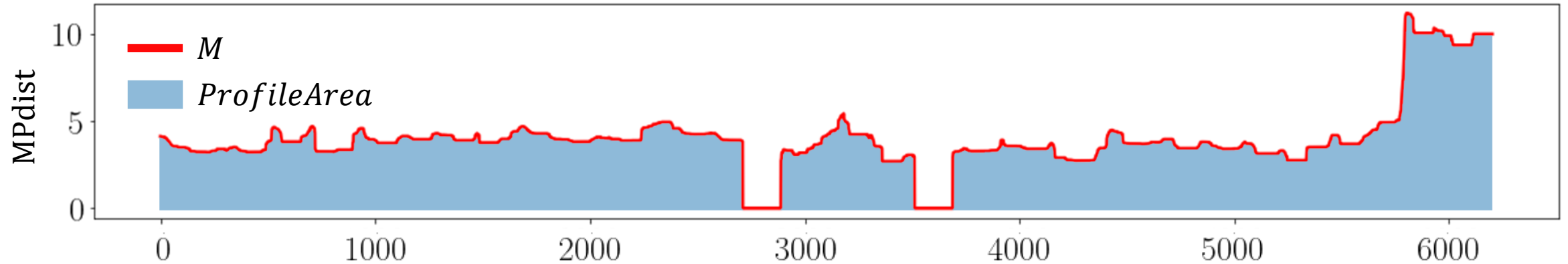
$$P_{ABBA} = allP_{AB}(i, m - \ell) \cdot allP_{BA}(i)$$

$$MPdist(S_i, T_{i,m}) = SortedP_{ABBA}(k), k = [0.1m]$$

# Вычисление кривой $M$



# Критерий выбора $m_{best}$

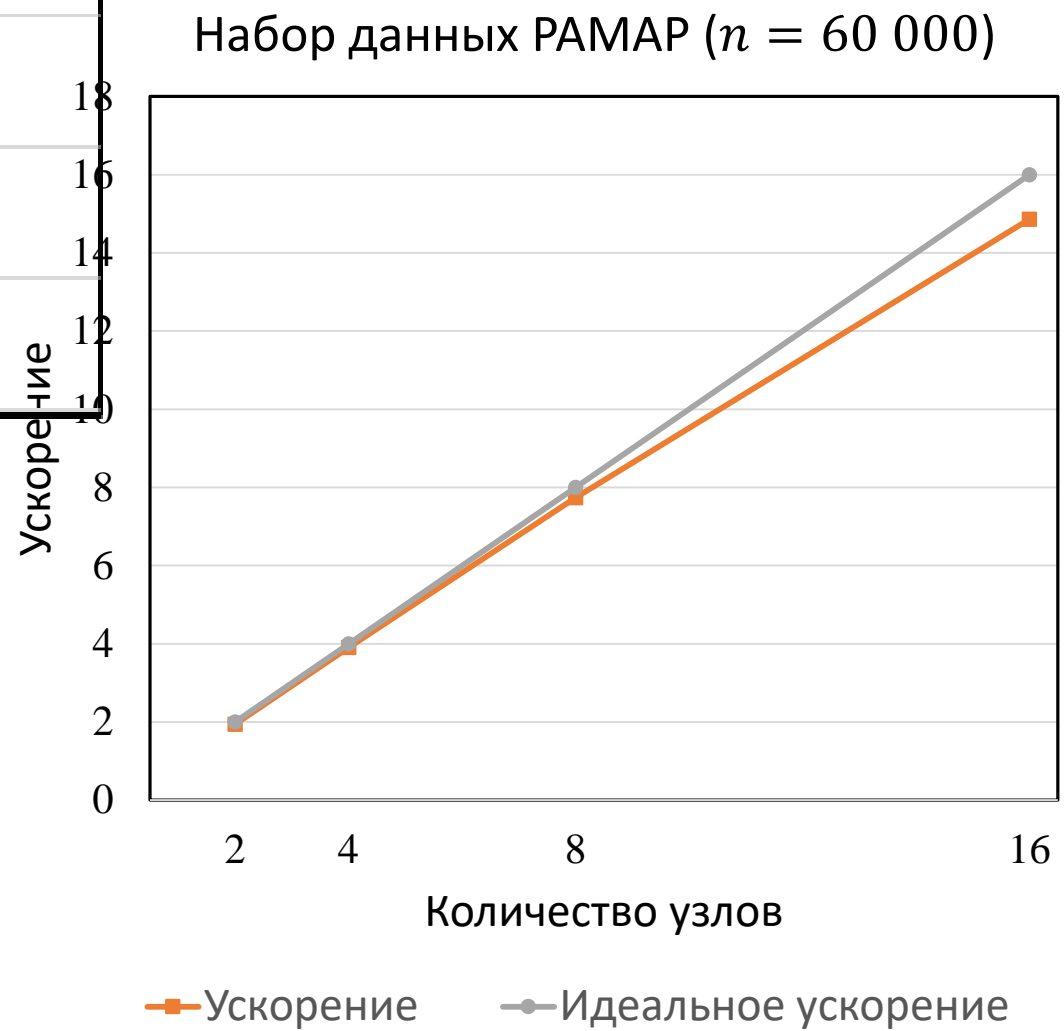
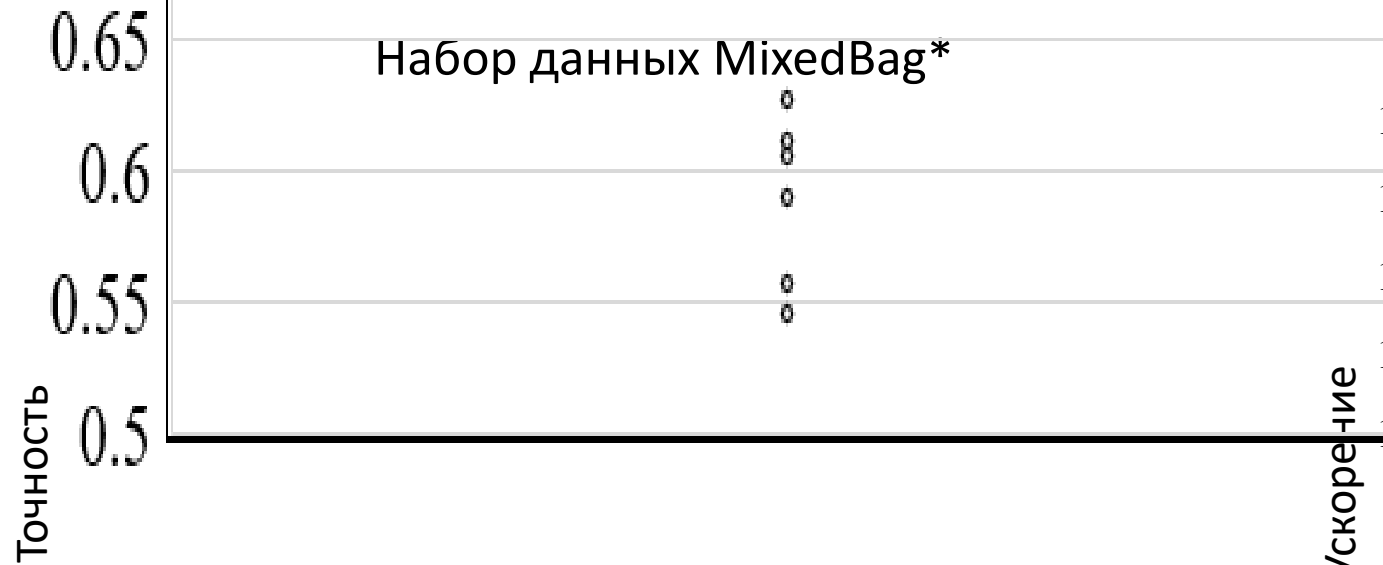


$$M_i = \min_{1 \leq j \leq K} \{d_i \mid d_i \in D_j\},$$

$$ProfileArea(m) = \sum_{i=1}^{n-m+1} M_i$$

$$m_{best} = \arg \min_{m_{min} \leq m \leq m_{max}} ProfileArea(m)$$

# Эксперименты: точность и производительность

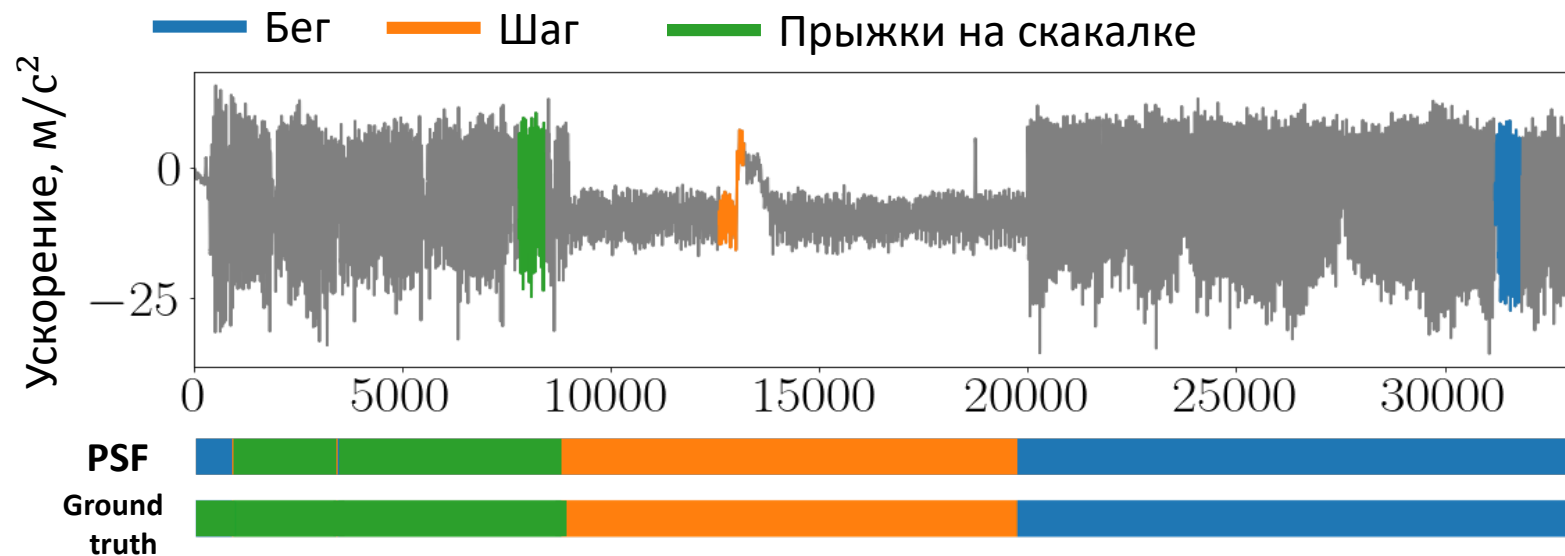


$$\text{Точность} = \frac{\text{True Positive}}{n}$$

\*Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020). DOI: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)



# Эксперименты



Временной ряд РАМАР

$n = 30\,003$

$m_{min} = 100, m_{max} = 1\,500$

$m_{best} = 630$

Точность = 0.96

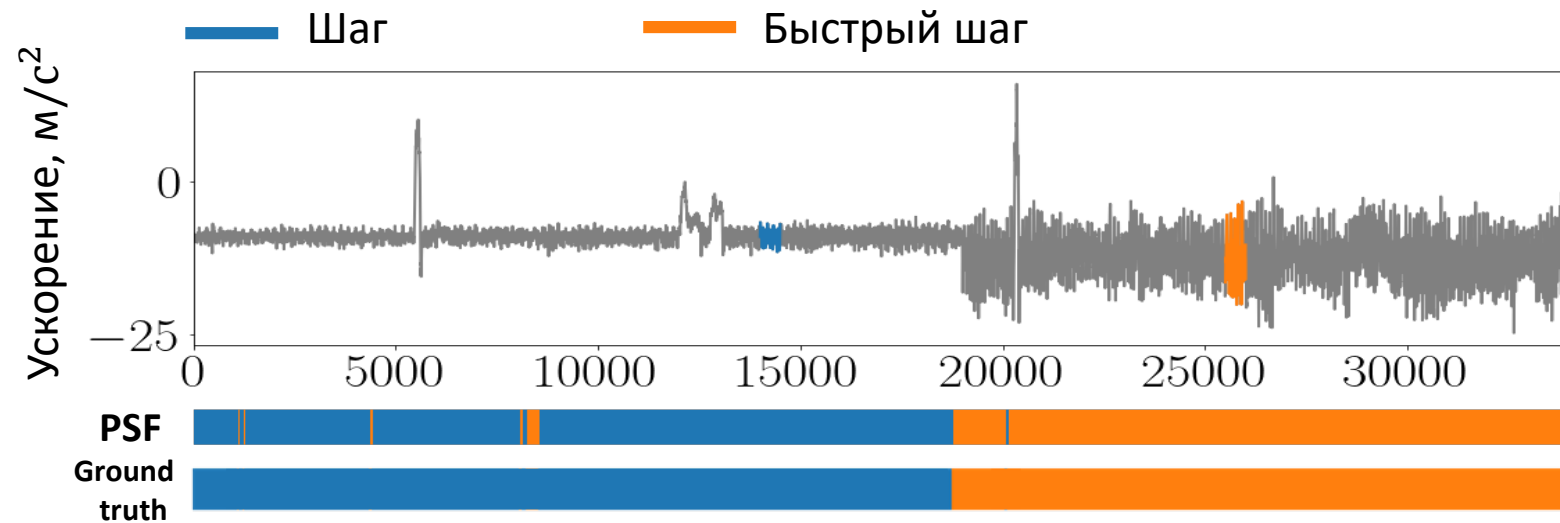
Временной ряд РАМАР

$n = 34\,001$

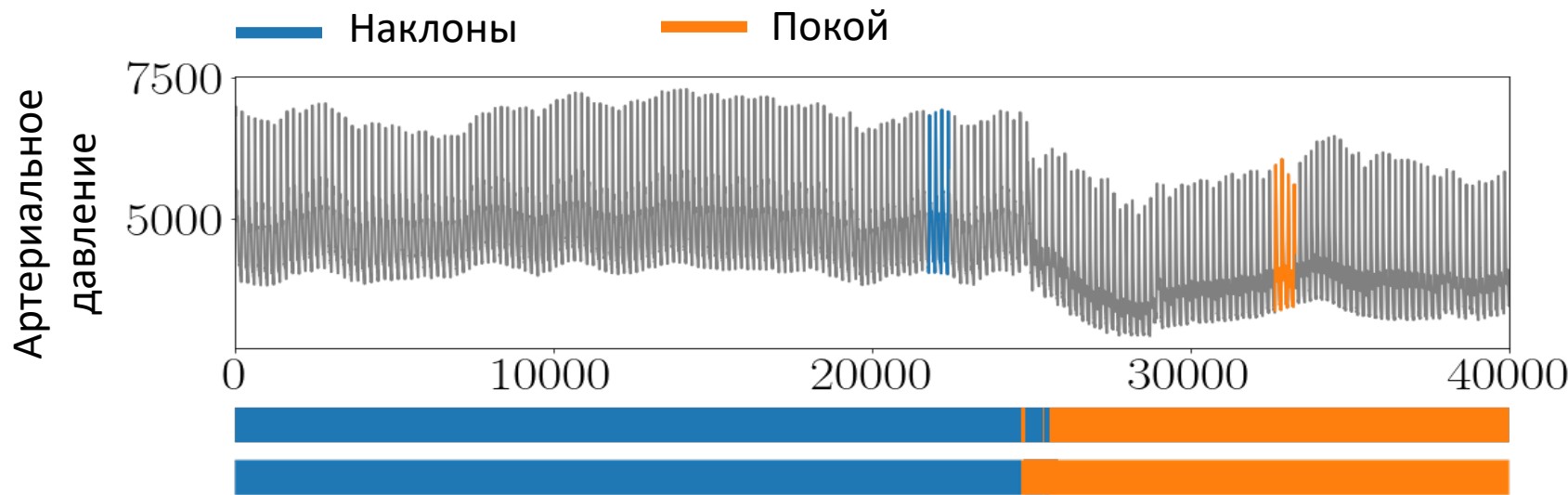
$m_{min} = 100, m_{max} = 1\,500$

$m_{best} = 598$

Точность = 0.97

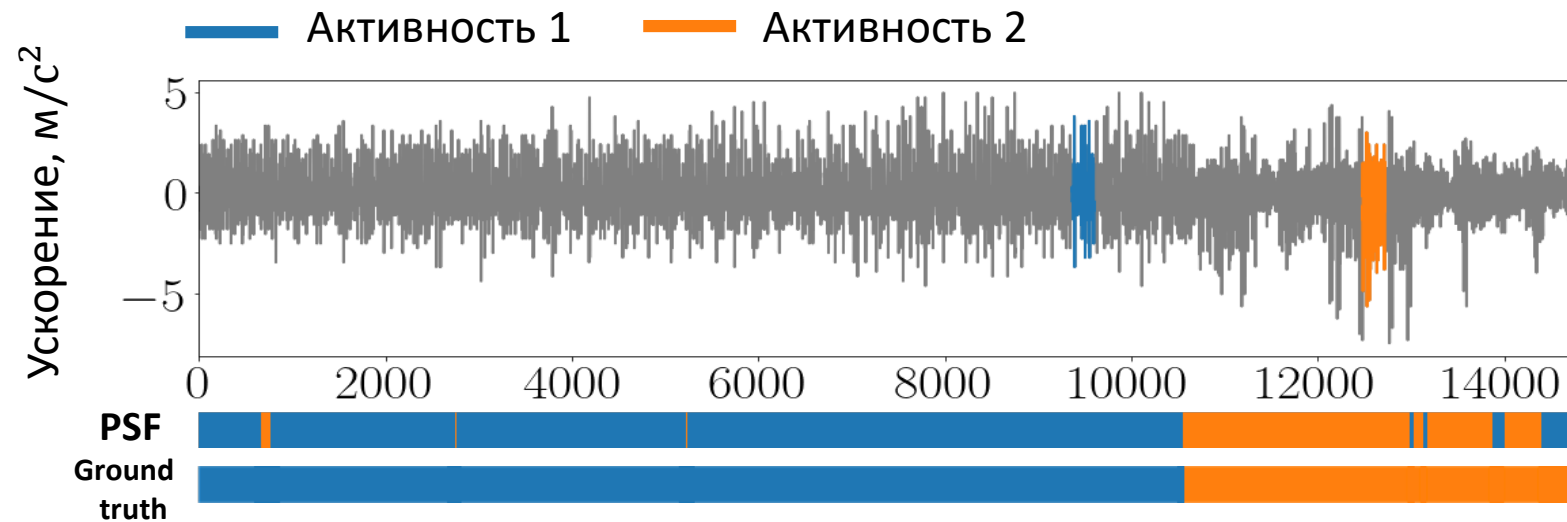


# Эксперименты



Временной ряд TitltABP  
 $n = 40\,000$   
 $m_{min} = 100, m_{max} = 2\,000$   
 $m_{best} = 615$   
Точность = 0.98

Временной ряд RoboticDogActivityY  
 $n = 14\,699$   
 $m_{min} = 50, m_{max} = 1000$   
 $m_{best} = 242$   
Точность = 0.94



# Заключение

- Разработан алгоритм разметки больших временных рядов на кластере с GPU узлами.
- Проведены эксперименты, показывающую высокую точность и масштабируемость разработанного алгоритма.

Спасибо за внимание! Вопросы?

Андрей Гоглачев, [goglachevai@susu.ru](mailto:goglachevai@susu.ru)