Performance Prediction Modeling of GPU Applications

Marcos Amarís[1], Alfredo Goldman(advisor)[1], Raphael Y. de Camargo(co-advisor)[2],
[1]Institute of Mathematics and Statistics - University of São Paulo - São Paulo, Brazil. {amaris, gold}@ime.usp.br
[2]Center for Mathematics, Computation and Cognition - Universidade Federal do ABC - Santo André, Brazil. raphael.camargo@ufabc.edu.br

IME-USP

Problem Statement

The prediction of application execution times over GPUs is a great challenge and is essential for efficient JMS. There are different approaches to do this, such as analytical modeling and machine learning techniques.

Major Research Highlights

We have developed a simple BSP-based model [1] to predict execution time of CUDA kernel functions. In this poster, we present two works. The first is a fair comparison of our BSP analytical model [2] with three machine learning techniques [3]. The second is a comparison of these three ML techniques when using a process of feature extraction[1].

Simple BSP-based model for GPU Applications [2]

$$\mathbf{T_k} = \frac{\mathbf{t} \cdot (\mathbf{Comp} + \mathbf{Comm_{GM}} + \mathbf{Comm_{SM}})}{\mathbf{R} \cdot \mathbf{P} \cdot \lambda} \qquad (1)$$

$\mathbf{Comp}$ is the computational cost of one thread, $\mathbf{Comm_{GM}}$ is the communication cost of global memory accesses of one thread (Equation 3), $\mathbf{Comm_{SM}}$ is the communication cost of shared memory accesses of one thread (Equation 2), λ adjusts application performance according to the optimal usage of the GPU resources.

$$\mathbf{Comm_{SM}} = (\mathbf{ld_0} + \mathbf{st_0}) \cdot \mathbf{g_{SM}} \qquad (2)$$

$$\mathbf{Comm_{GM}} = (\mathbf{ld_1} + \mathbf{st_1} - \mathbf{L1} - \mathbf{L2}) \cdot \mathbf{g_{GM}} + \mathbf{L1} \cdot \mathbf{g_{L1}} + \mathbf{L2} \cdot \mathbf{g_{L2}} \qquad (3)$$

Machine Learning Techniques [3]

- Linear Regression
- Support Vector Machines
- Random Forests

The same parameters were used for analytical model and ML techniques in the the first work. For extraction features we have used correlation analysis and hierarchical clustering techniques.

Work 1 - Developed BSP-based Vs. ML

In this work regular applications were used. T_k is the approximated execution time of a kernel function with t threads.

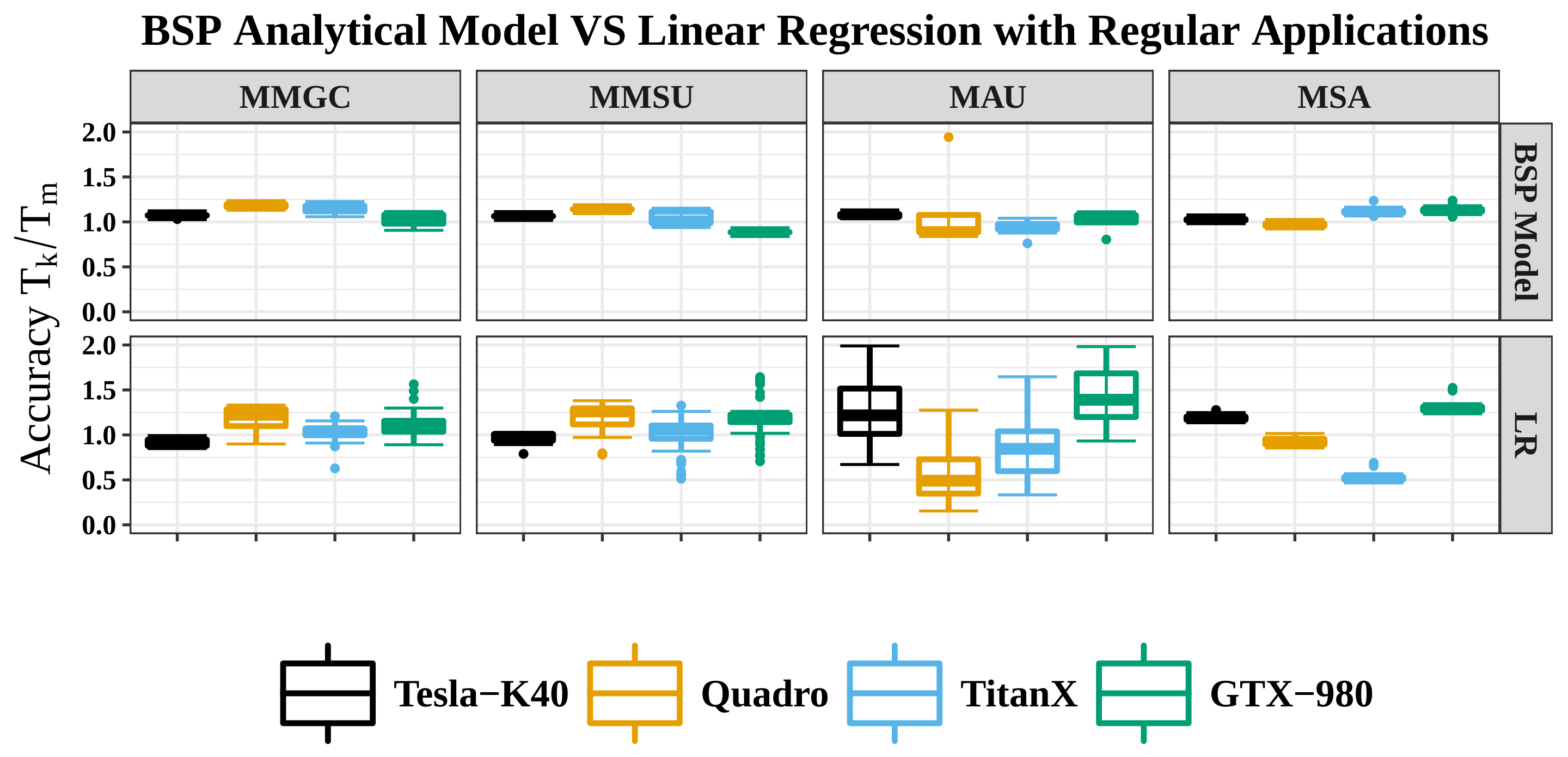

Figure 1: BSP Model Vs. Linear Regression. 4 Kernels over 4 GPUs

Work 2 - ML Vs. ML with Feature Extraction

In this work Irregular applications of Rodinia Benchmark Suite were used.

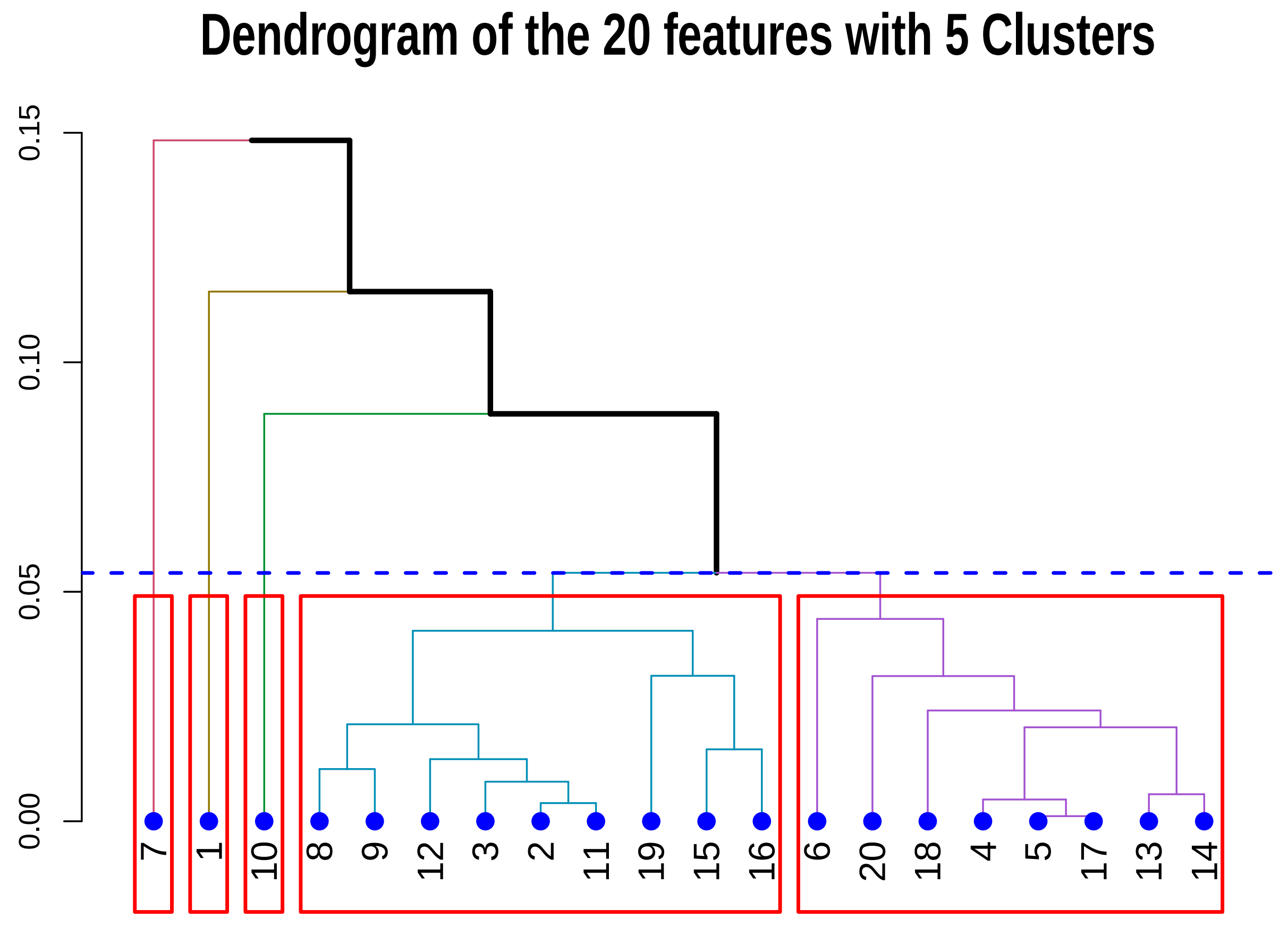

Figure 2: Dendrogram with 5 clusters to select 1 features from each cluster. 20 Features had a high correlation with the execution times

We can see that 2 parameters provided the smaller error for 5 application features. Consequently, we decided to use 2 GPU parameters in the experiments.

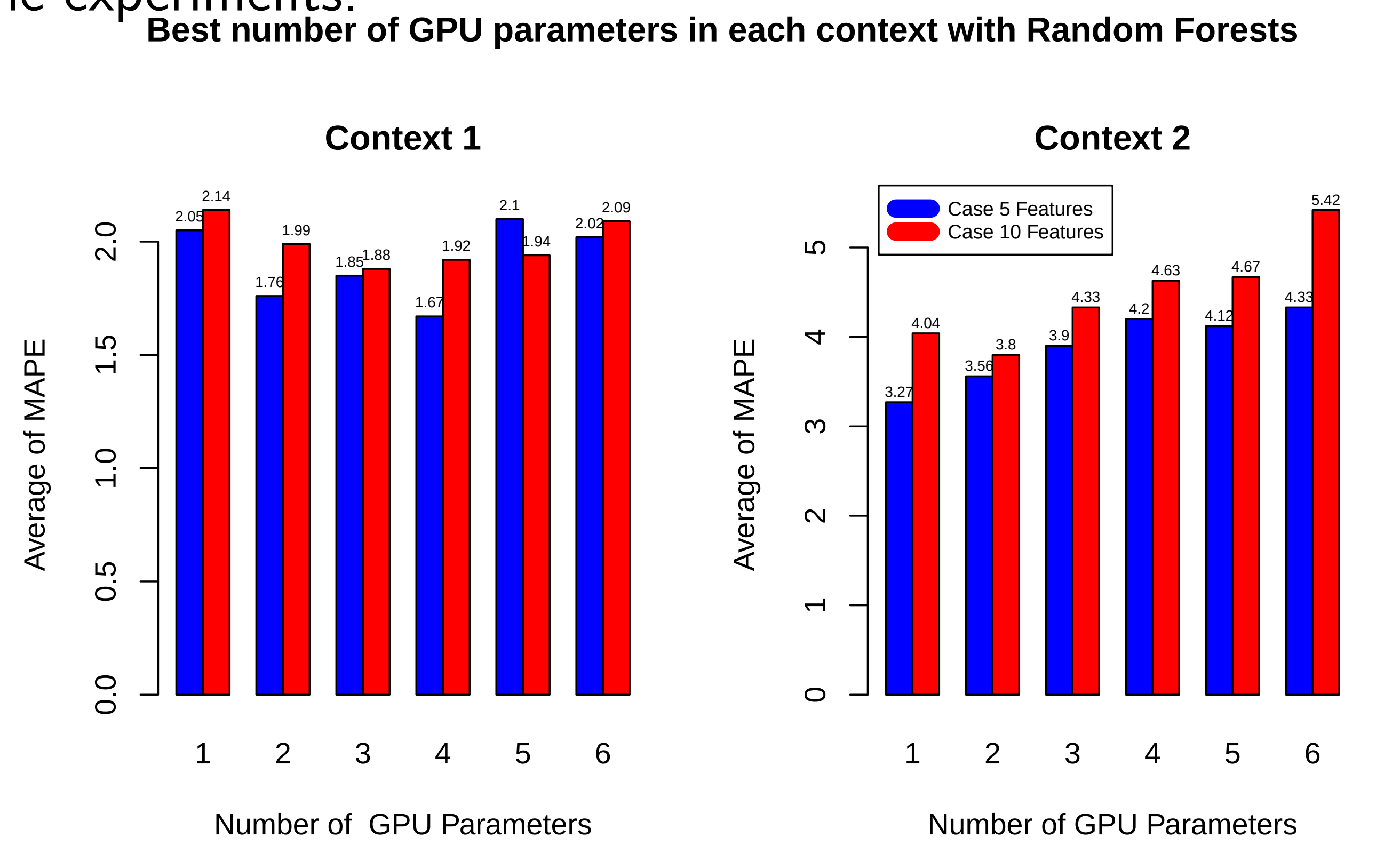

Figure 3: Best number of GPU parameters in each one of the contexts

2 Contexts were tested with each ML

1. GPUs: we evaluated if the ML algorithm could predict the execution time over a previously unseen GPU
2. Kernels: in this context we predict the execution time of a previously unseen CUDA kernel.

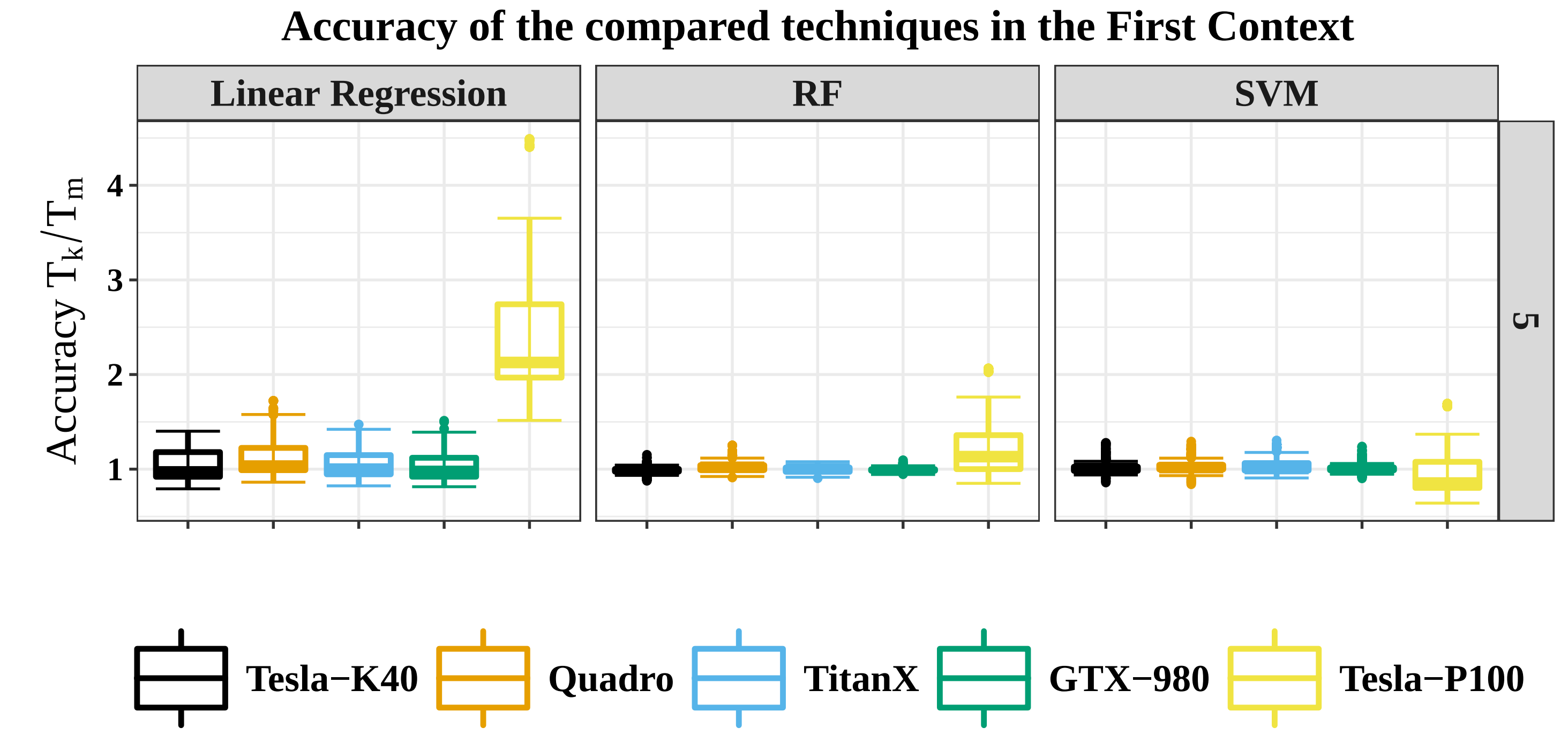

Figure 4: Accuracy of the first context with all the machine learning techniques used over 5 target GPUs

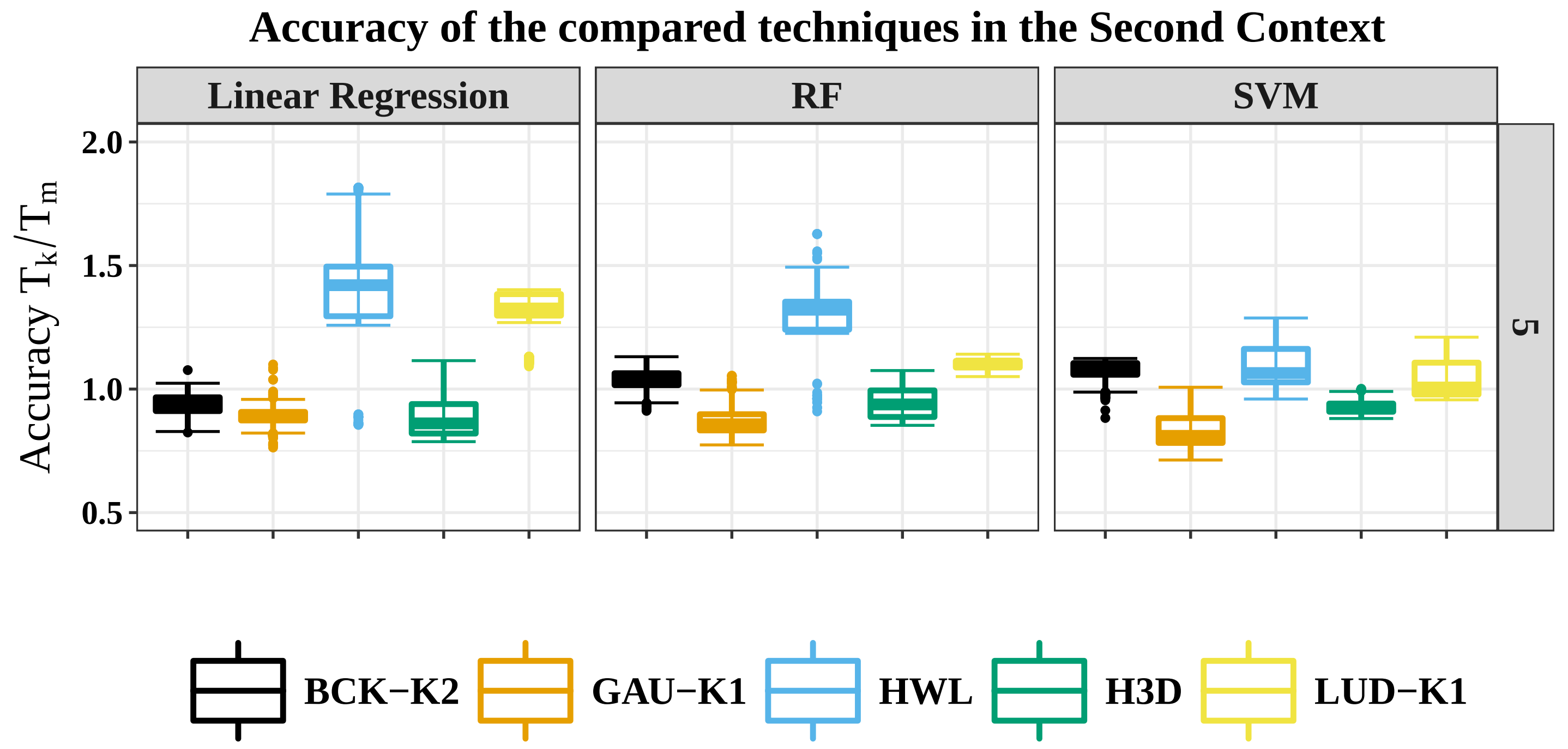

Figure 5: Accuracy of the second context with all the machine learning techniques used over 5 target irregular kernels

Conclusions

Performance of regular GPU applications can be easily predicted with simple analytical models, few parameters about computation or computation of the application are enough for these applications. Machine learning techniques worked better for Irregular Applications. Linear models are enough if the features are linearized. Predictions with high accuracy are possible with an optimum statistical process of feature extraction.

References

[1] L. G. Valiant, "A bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111.

[2] M. Amaris, D. Cordeiro, A. Goldman, and R. Y. Camargo, "A simple bsp-based model to predict execution time in gpu applications," in *(HiPC), 2015*, December 2015, pp. 285–294. Bangalore. India.

[3] M. Amaris, M. Dyad, R. Y. Camargo, A. Goldman, and D. Trystram "A Comparison of GPU Execution Time Prediction using Machine Learning and Analytical Modeling," in *(NCA), 2016 IEEE*, October 2016, pp. 326-333. Boston. USA.

Doctoral Showcase - International Conference for Super Computing (SC 17), 12 - 17 November 2017. Denver, CO - USA