Performance Evaluation and Applications 2023-2024

Project B

Performance of a Big-Little Architecture

Lamperti Federico 10680961 - QR code ID 58135110

Architecture

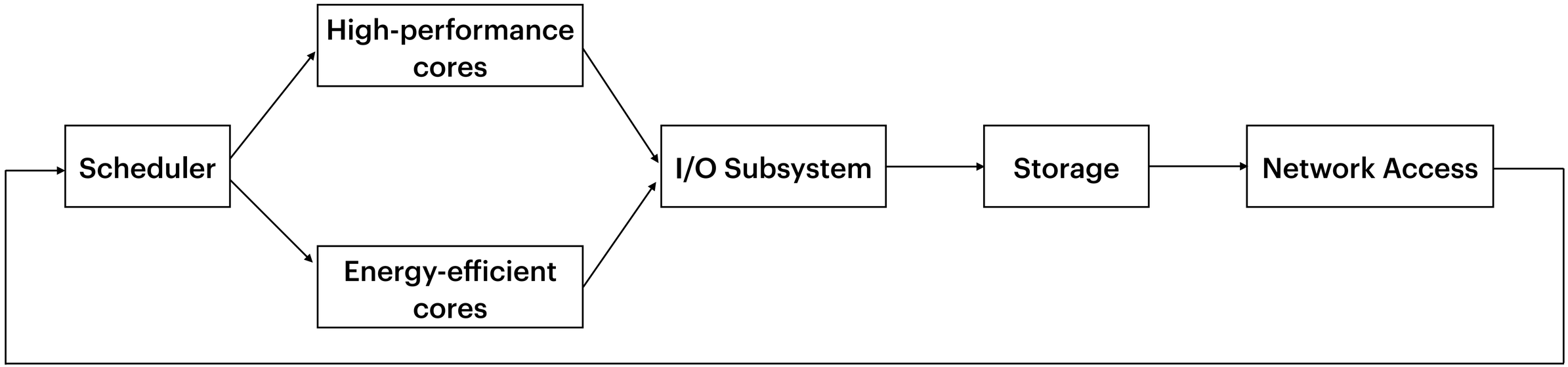
- 2 types of tasks which are 10 Heavy Computation and 32 Low Computation
- CPU composed of 4 High-efficiency cores and 8 Energy-efficient cores and a scheduler
- I/O subsystem, Storage component and Network Access

Architecture

- The scheduler will mainly schedule heavy computation tasks on the high-performance cores, while low computation tasks on the energy-efficient cores.
- Each department works independently of the others
- The two sets of cores are considered as two independent sharing stations
- I/O subsystem, Storage component and Network Access are considered working in processor sharing

Goal of the Project

Find the best routing probability distribution for the scheduler and determine the throughput in each scenario analyzed



Fitting

Using a Matlab script and having at disposal the traces that describe the computation times for the two types of tasks in each of the two types of cores, I was able to define the characteristics of the various service times, exploiting the method of moments and the maximum likelihood method and fitting the traces according to several distributions, such as Exponential, Erlang, Weibull, Pareto, Hyper-Exponential and Hypo-Exponential.

Then looking at the coefficient of variation and the curves in the graphs I was able to select the best distribution with the appropriate parameters' values.

Fitting - HE

Moments:

First Moment = 96,560531

Second Moment = 29425.248819

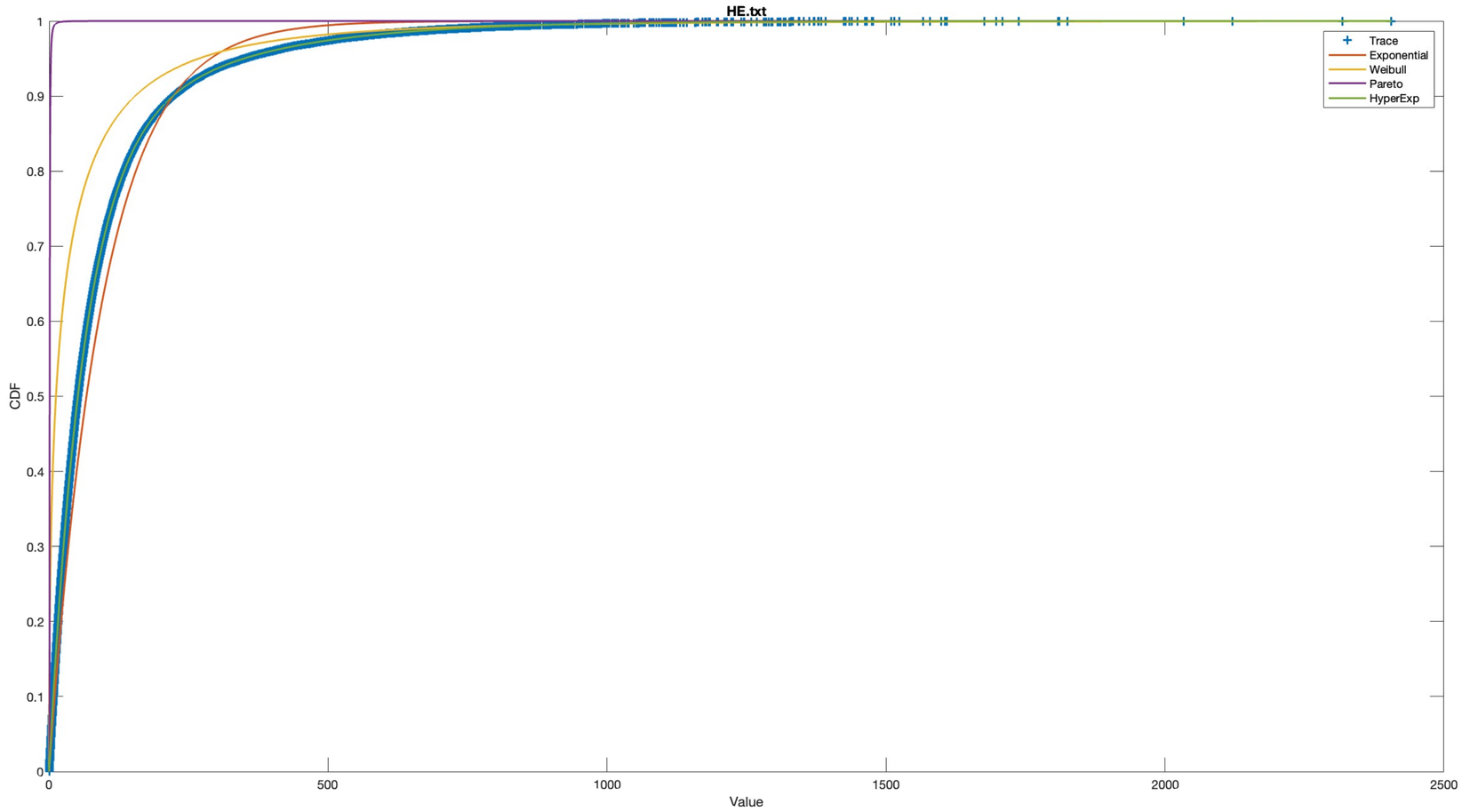
Third Moment = 18315896.156154

CV > 1

Distribution:

Hyper-Exponential

- λ1 = 0,016712
- λ2 = 0,004106
- P = 0,800100



Fitting - HH

Moments:

First Moment = 24,125513

Second Moment = 1820,784049

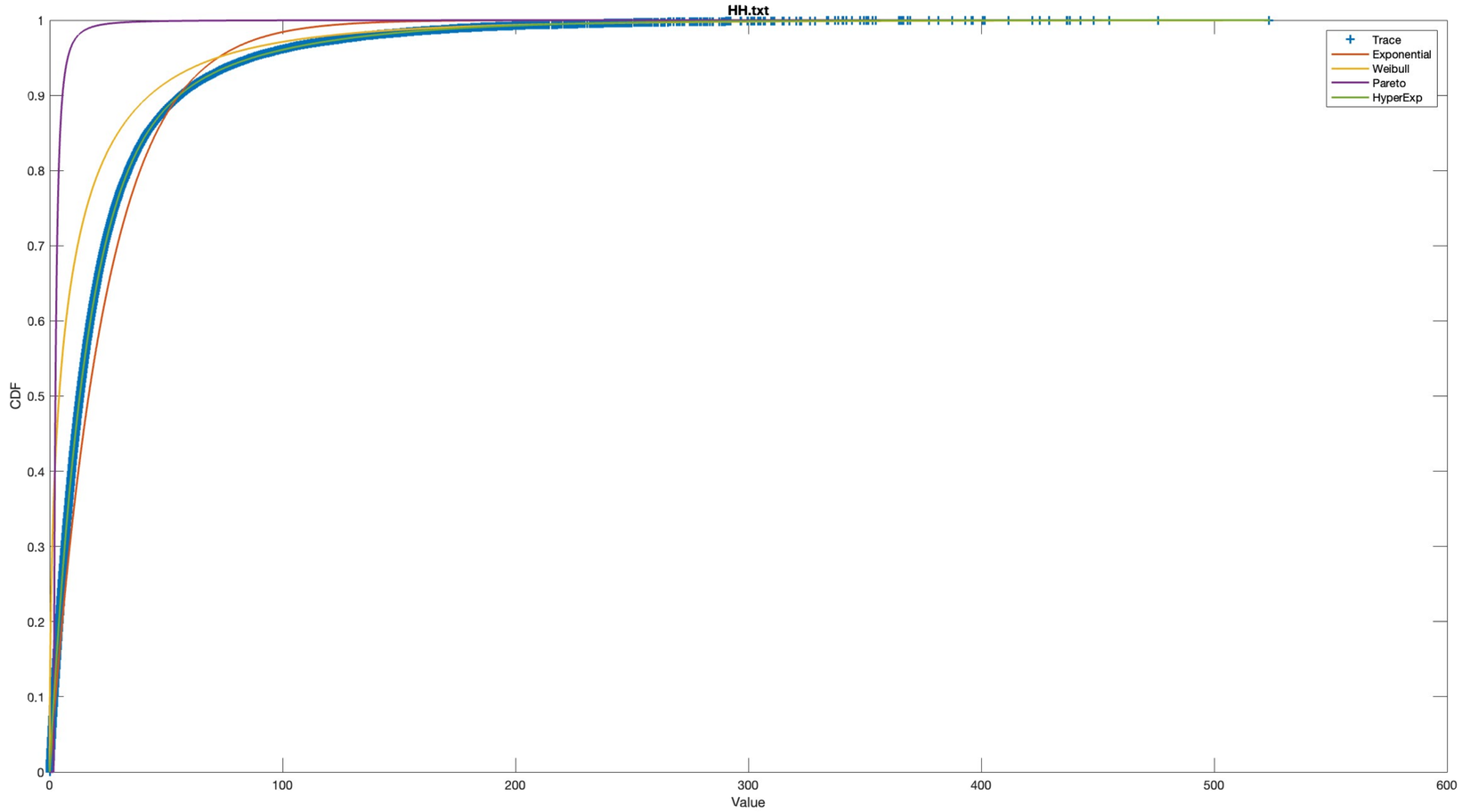
Third Moment = 278560,276419

CV > 1

Distribution:

Hyper-Exponential

- λ1 = 0,066928
- λ2 = 0,016758
- P = 0,794677



Fitting - LE

Moments:

First Moment = 4,804131

Second Moment = 24,991874

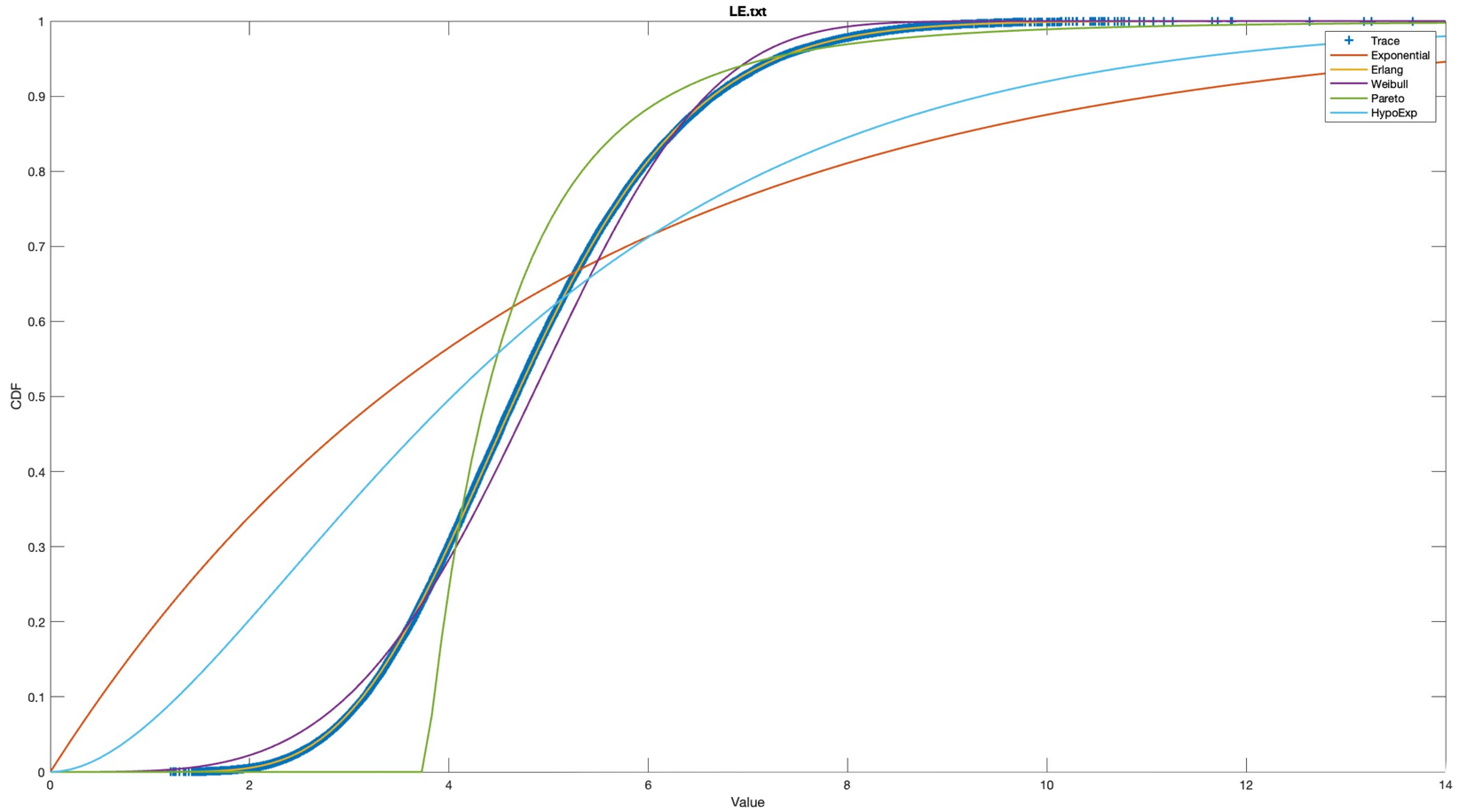
Third Moment = 139,942299

CV < 1

Distribution:

Erlang

- K = 12
- λ = 2,497850



Fitting - LH

Moments:

First Moment = 1,502497

Second Moment = 2,445363

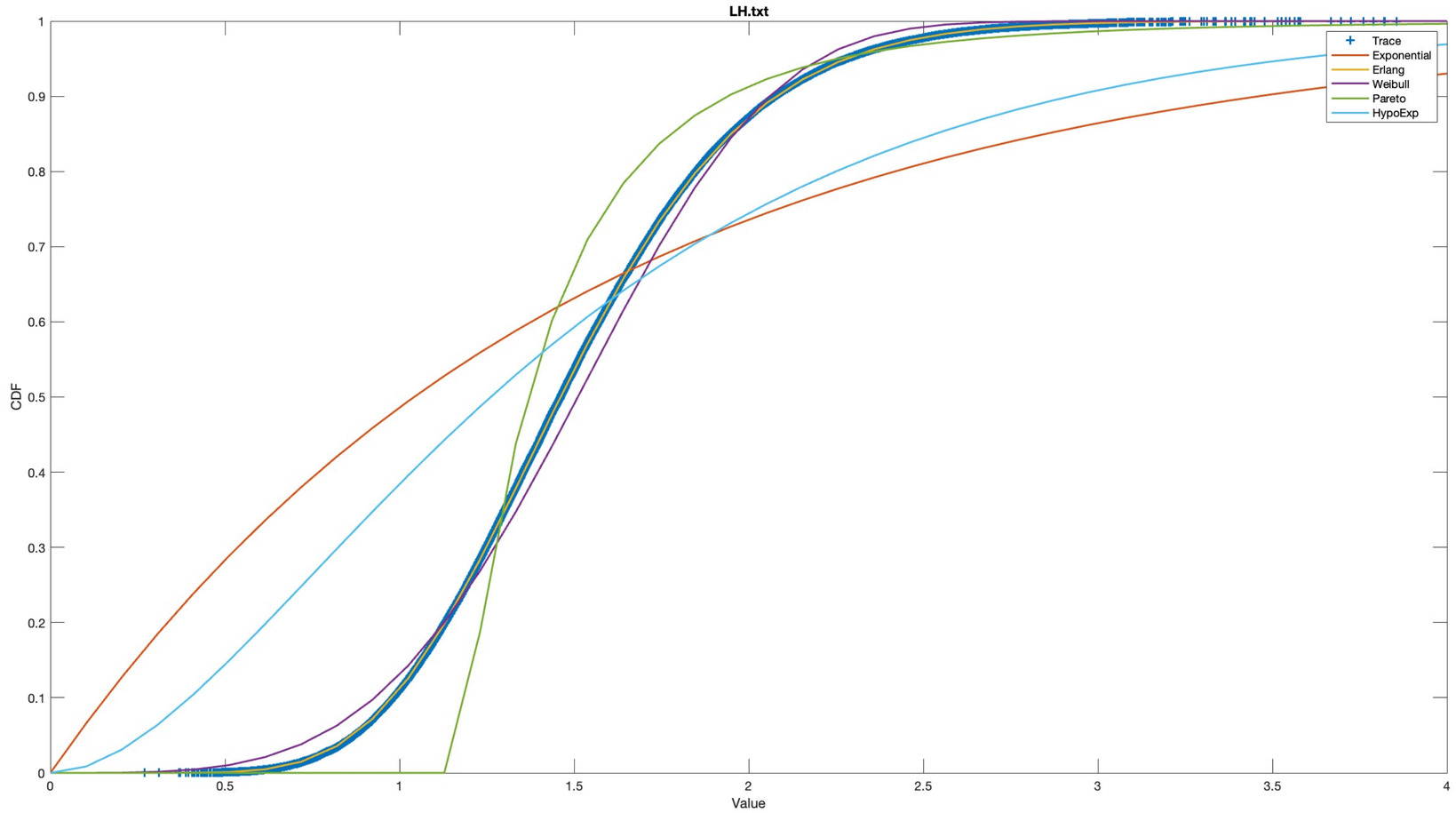
Third Moment = 4,286979

CV < 1

Distribution:

Erlang

- K = 12
- λ = 7,986706

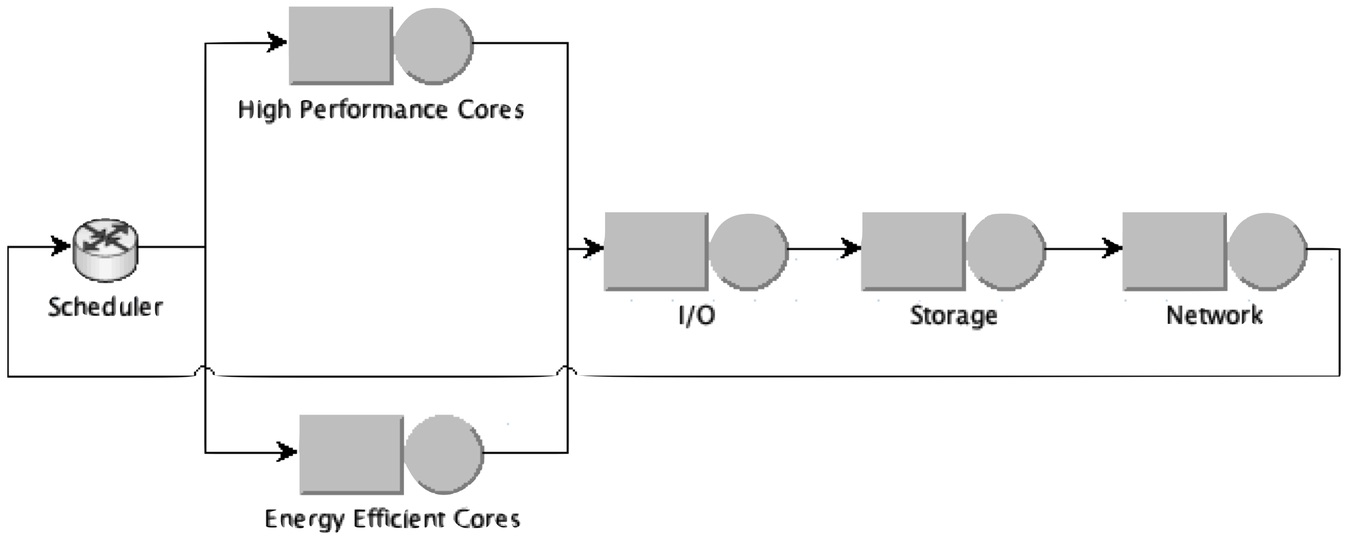


Model Definition

Using JSIM Graph, a tool from JMT, I created a model that can fit as much as possible with the requirements.
I chose to implement a closed model defining the service times distributions according to the parameters found during the fitting phase and to the parameters that were provided.
Every component is modeled as a queue station with infinite capacity, FCFS policy, and non-preemptive scheduling, except for the scheduler which is a router that forwards jobs with probability routing.
The Scheduler is also the reference station for the system.
The high-performance cores component has 4 servers to simulate the number of cores present while the energy-efficient one has 8.

Model Definition

- Closed Model (N always the same)
- Multi-Class
- Queue stations with ∞ capacity
- FCFS policy
- Non-preemptive scheduling



Analysis Method

In my system analysis, I experimented with varying percentages of scheduling distributions. Specifically, I tested scenarios where 5% to 20% of tasks were allocated to the opposite set of cores for each of the two classes. I increased the percentage of 5% for each test.

After running these tests, I evaluated the results based on various metrics. This evaluation process helped me identify the most optimal configuration.

Every Measure is evaluated with a confidence interval of 99% and a max relative error equal to 3%

	1	2	3	4
HeavyToHP	0,95	0,9	0,85	0,8
HeavyToEE	0,05	0,1	0,15	0,2
	A	**B**	**C**	**D**
LowToHP	0,05	0,1	0,15	0,2
LowToEE	0,95	0,9	0,85	0,8

Percentage distributions Legend

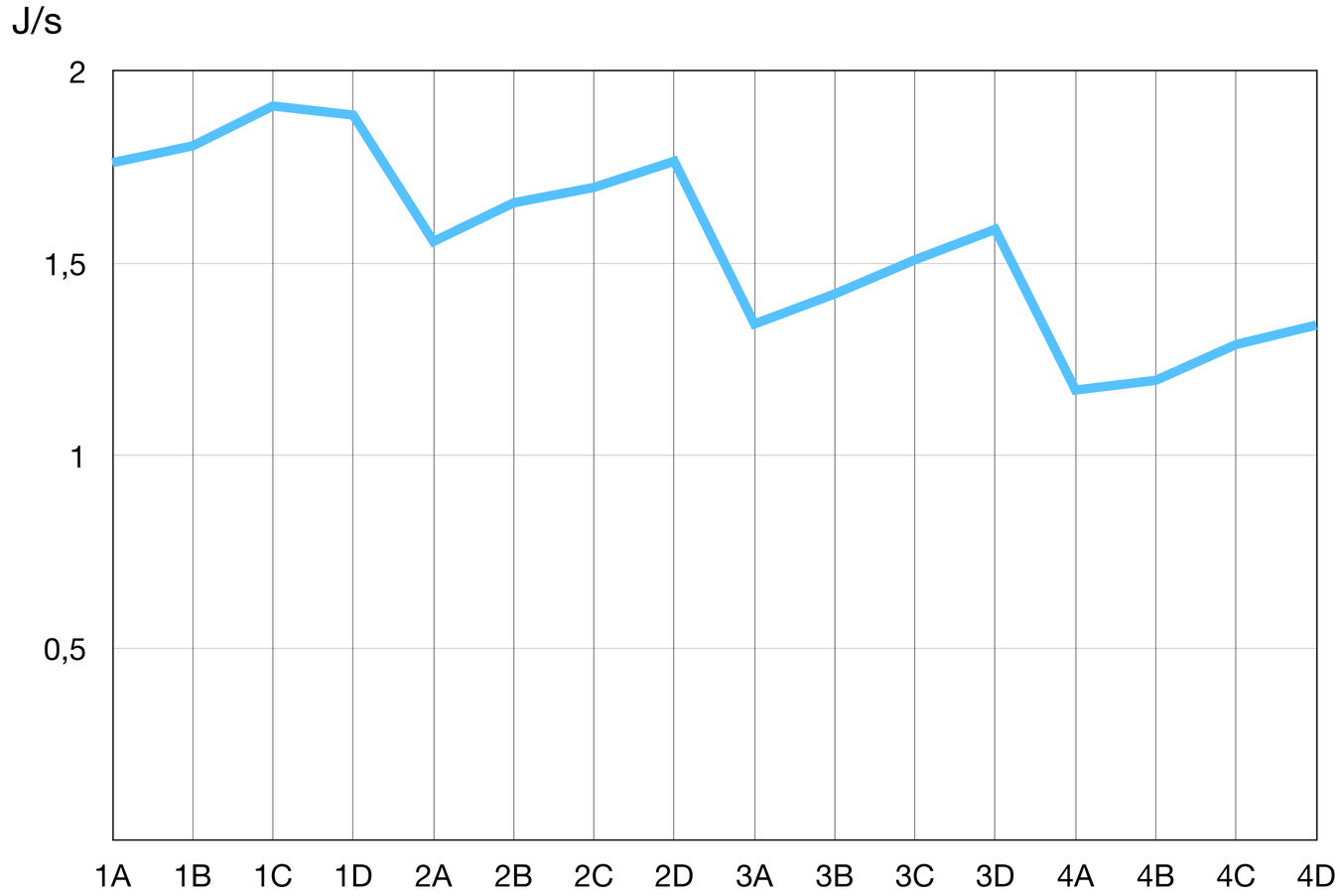
Performance Metrics Analyzed

- **Throughput**: total number of jobs that can be served per time unit
- **Response Time**: average time a job spends in the system
- **System Power**: optimal operational point of a system, corresponds to the maximum throughput with the minimum response time
- **Utilization**: percentage of time a station is used
- **Queue Time**: average time a job spends waiting in a station queue

Throughput

- Best 3 values:
  1) 1C = 1.9096
  2) 1D = 1.8860
  3) 1B = 1.8059

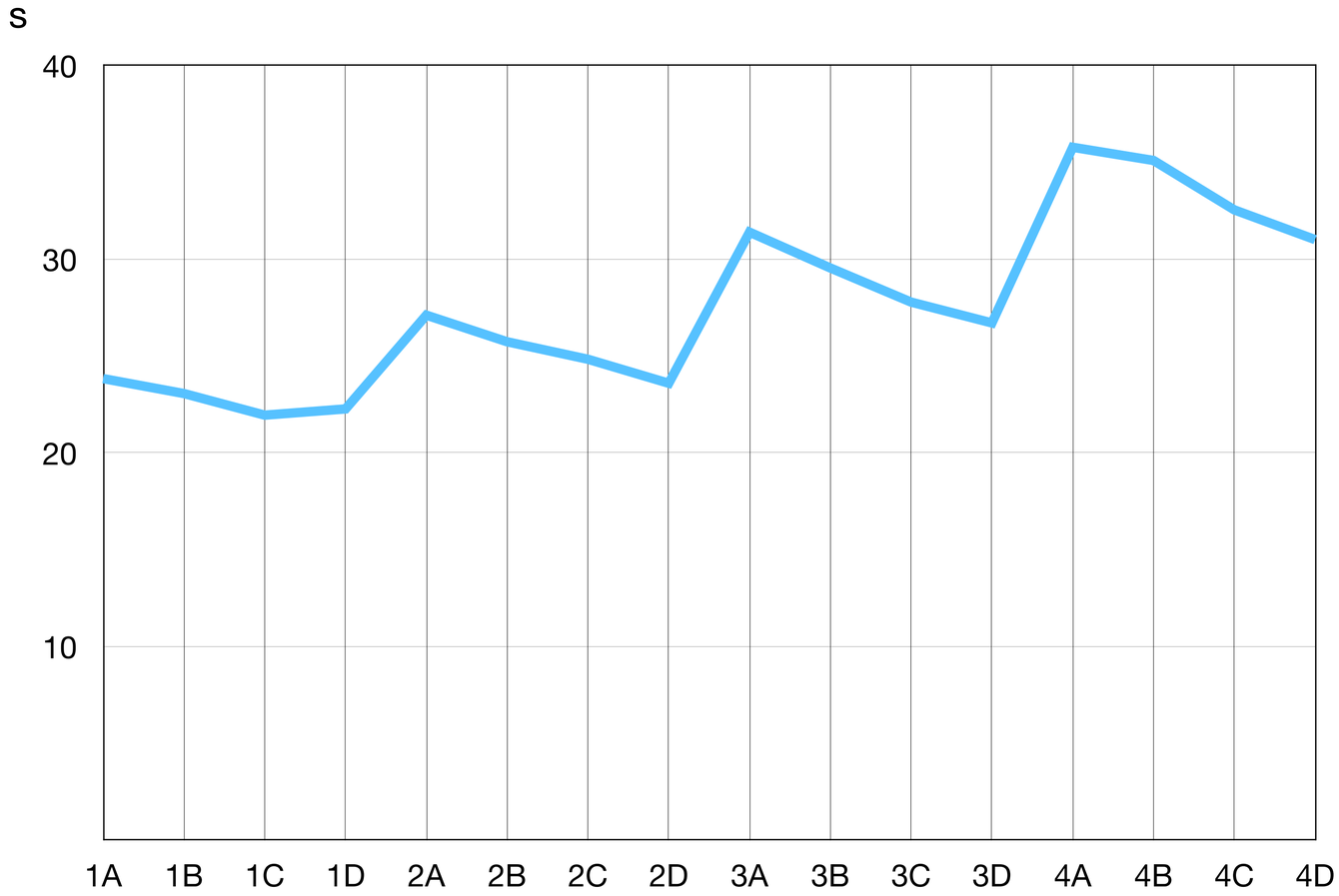
- Worst 3 values:
  1) 4A = 1.1713
  2) 4B = 1.1965
  3) 4C = 1.2896



Response Time

- Best 3 values:
  1) 1C = 21.9553
  2) 1D = 22.2763
  3) 1B = 23.0710

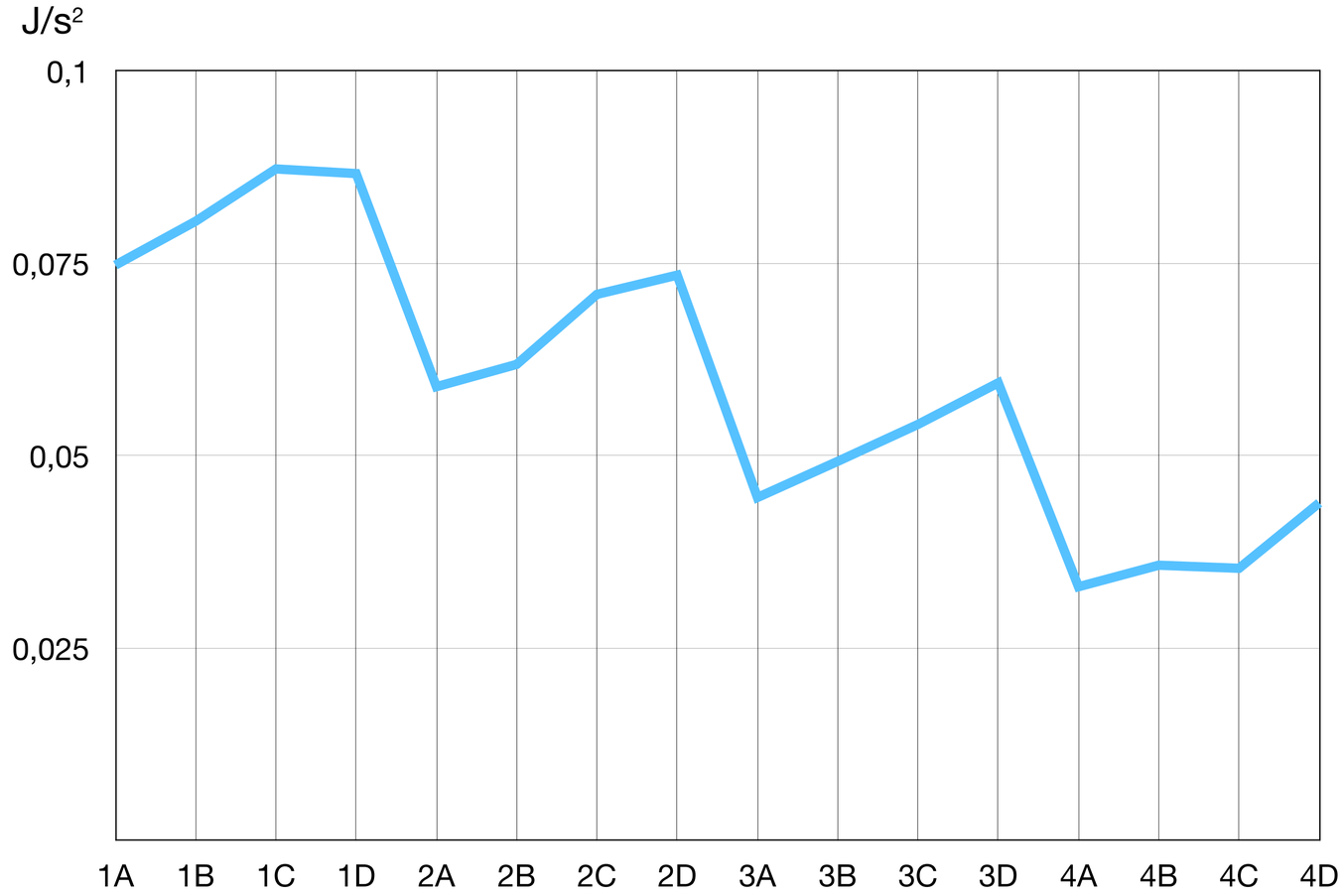
- Worst 3 values:
  1) 4A = 35.7875
  2) 4B = 35.1067
  3) 4C = 32.5554



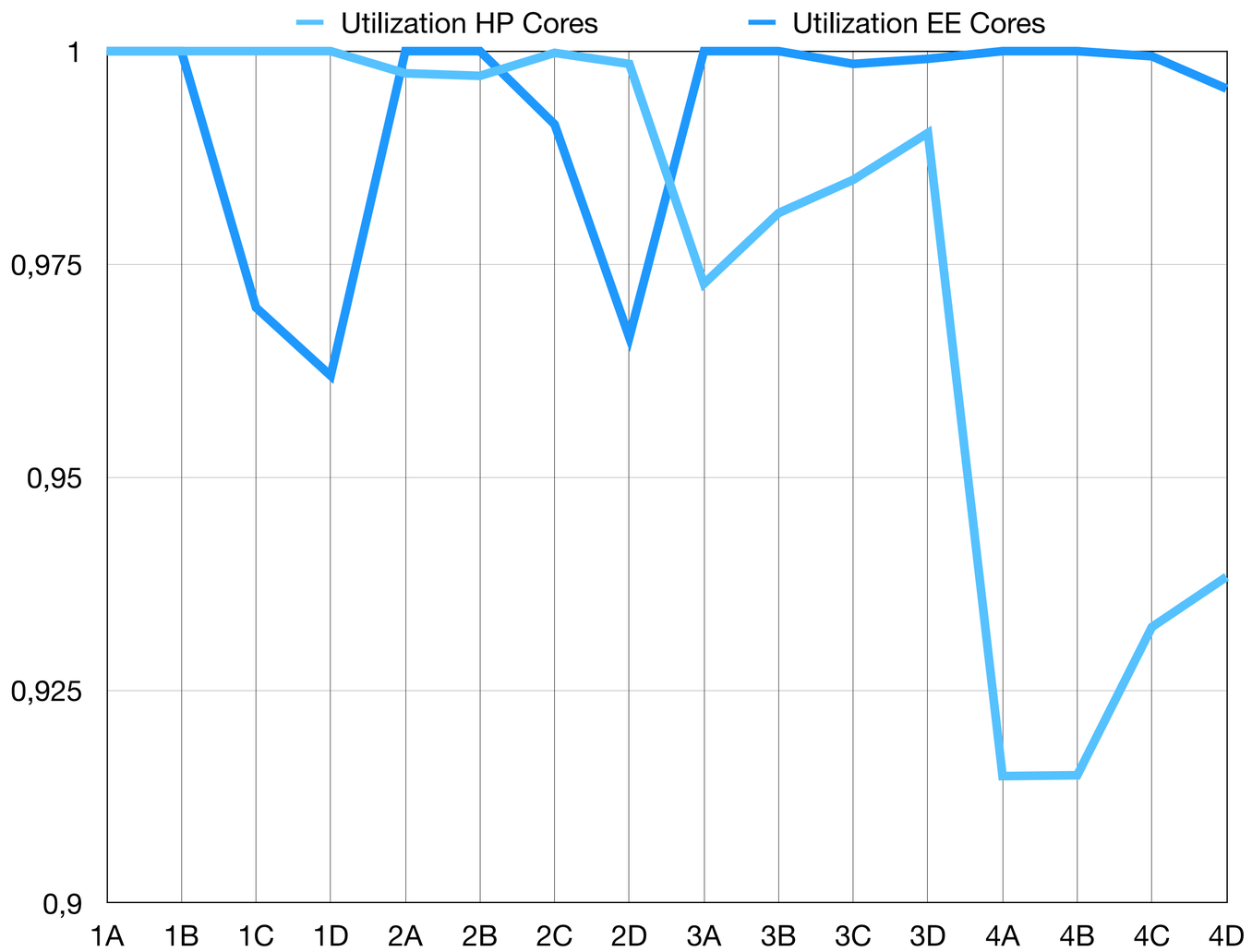
System Power

- Best 3 values:
  1) 1C = 0.0873
  2) 1D = 0.0867
  3) 1B = 0.0805

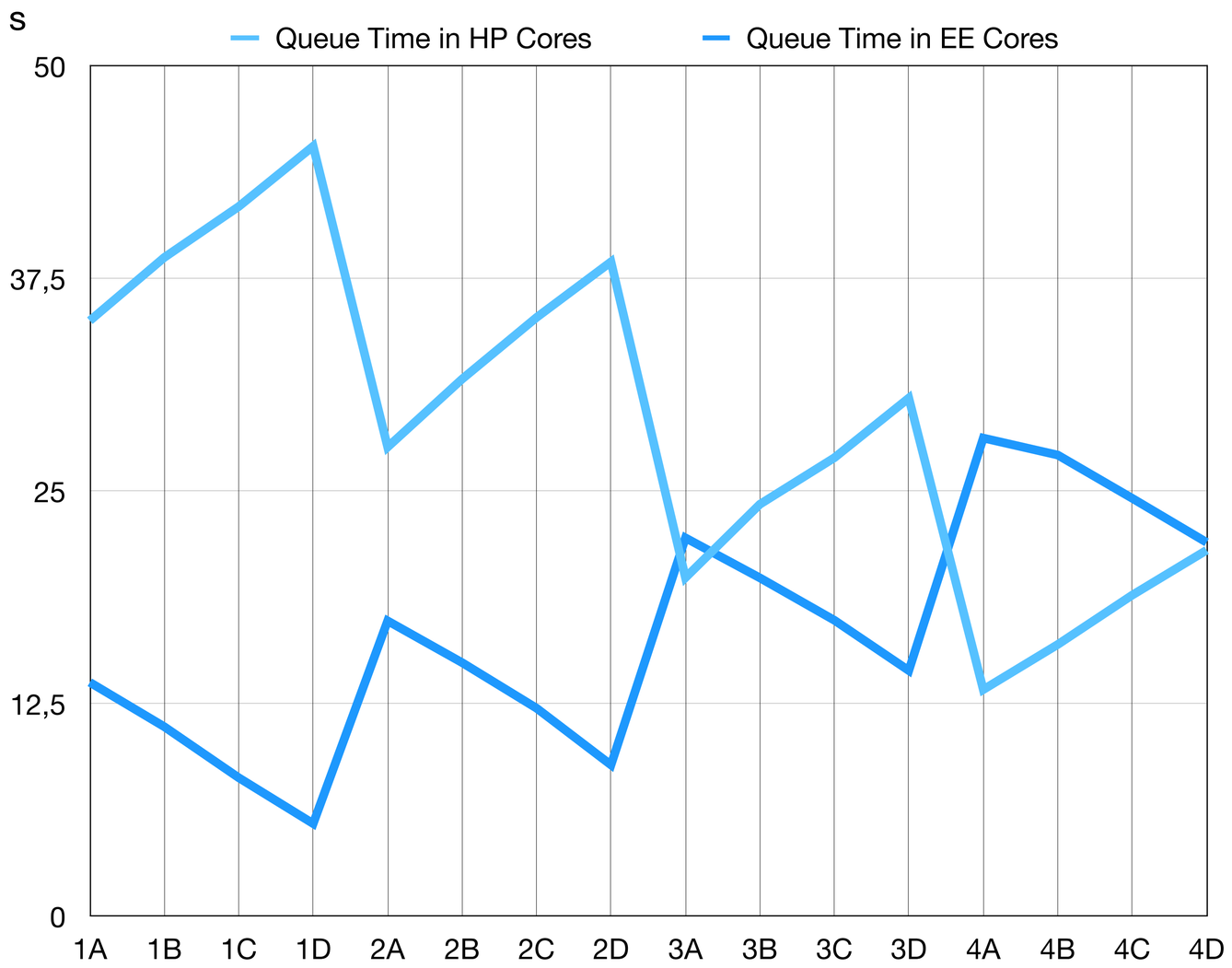
- Worst 3 values:
  1) 4A = 0.0330
  2) 4C = 0.0354
  3) 4B = 0.0358



Utilization



Queue Time


s
Queue Time in HP Cores
Queue Time in EE Cores
50
37,5
25
12,5
0
1A
1B
1C
1D
2A
2B
2C
2D
3A
3B
3C
3D
4A
4B
4C
4D

Conclusions

To provide the best possible solution to the problem, I decided to focus mainly on Throughput and Response Time. However, the other indexes were helpful to have a better understanding of the scenario.
Configuration 1C yielded the best results in terms of Throughput and Response Time, closely followed by Configuration 1D.
The higher Queue Times in configuration 1C compared to 3A indicate the significant impact of core speed differences.
One limitation of this analysis is the lack of information about the system's operating environment. For instance, Configuration 1C exhibits high component Utilization and a growth of the number of jobs in the system could become a real challenge.
Additionally, energy consumption could be a concern as the two types of Cores are almost always used.
In conclusion, while Configuration 1C shows promising results, potential issues such as high utilization and energy consumption need to be considered.

Additional Notes

The folder Additional Material contains:

- Fitting result in JPG format for each of the traces
- *FullAnalysis.xlxs*, an excel file containing all the data obtained as result from the analysis on the model and also confidence intervals for the best configuration for each of the metrics chosen