11th International Conference on Sensitivity Analysis of Model Output, April 23–25 2025, Grenoble, France

Sensitivity Analysis for Sizing Standalone Green Datacenters

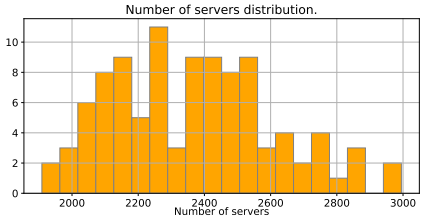
Manal BENAISSA, Noura DRIDI, Jean-Marc NICOD
SUPMICROTECH, CNRS, institut FEMTO-ST, F-25000 Besançon, France
[manal.benaissa,noura.dridi,jm.nicod]@femto-st.fr

Georges Da-Costa
IRIT, Université de Toulouse, CNRS, F-31062 Toulouse, France
georges.da-costa@irit.fr

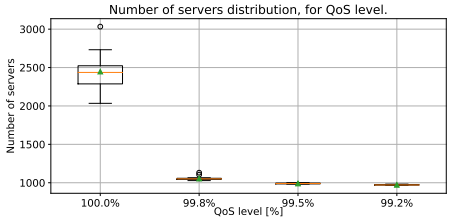
Since the widespread use of social media and IT services (cloud, video streaming, etc.), the ecological impact of IT infrastructures such as datacenters has become a significant concern. This is due to the large amount of electricity these infrastructures consume, and the reliance on fossil fuels to meet this demand. Datacenters powered by renewable energy have been proposed as a solution. However, due to the intermittency of renewable energy sources, these platforms remain connected to the conventional power grid. The DATAZERO2 project [1] has developed an approach with a fully autonomous electrical infrastructure including wind turbines, solar panels, short- and long-term energy storage devices (batteries and hydrogen systems respectively). The work presented here focuses on the sizing stage of such an infrastructure. Although this stage encompasses both the data center and the electrical infrastructure needed to power it, the focus here is on IT sizing. The challenge is to determine the optimal size of the infrastructure: large enough to meet customer demand without service interruption, but small enough to avoid overprovisioning, which would lead to unnecessary costs, pollution from excess equipment and wasted energy on idle servers. This raises the question of balancing system robustness against frugality. The sizing process must account for the potential variability in client usage (such as the increase in demand observed during lockdown due to remote working and streaming) while simultaneously minimizing the infrastructure to avoid purchasing unnecessary hardware.

The IT (datacenter) sizing takes as inputs a workload (requests submitted by clients) and the characteristics of the servers used. The workload is structured as a time series over one year, where requests (jobs) are presented hourly across 8760 hours. Each job J_i is defined by two values: its submission time s_i (the date the datacenter receives the request), and the job size w_i (the amount of work servers must do to complete the task). The sizing process then returns the minimum number of servers $mOPT$ required to satisfy this demand. IT sizing can be monitored by various metrics as the cost of the infrastructure (in US dollar), its carbon footprint (in kgeqCO2) resulting from the construction of the datacenter. The objective is to meet the expected QoS while minimizing both the ecological and economic impact of the infrastructure. This is where the first question of sensitivity analysis comes in: how does the workload, and therefore the number of servers, contribute to variations in quality of service? Besides, to ensure robustness, the datacenter must be able to meet the expected QoS even when the workload fluctuates over the years. To deal with, we propose the following framework: considering m servers, the sizing is challenged across a set $setW$ of workloads (tests) and for each test, the resulting QoS obtained. A guarantee is then established, for m servers, the datacenter commits to meeting an expected QoS of $Q\%$, with a success rate of $R\%$. This success rate represents the ratio of tests that meet the expected QoS, relative to the total number of tests. The robustness of the datacenter is evaluated on a set of 100 workloads, providing the associated guarantee. These workloads are based on a generator using Google traces [2], with distributions determined the number of jobs submitted per hour (which follows a Pareto distribution with an index of $\alpha = 5$) and the size of the jobs in Million Instructions [MI] (which follows a log-normal distribution with parameters $\sigma = 1.42$ and $\mu = 6.25$). The second question in the sensitivity analysis is how the number of servers is governed by the variation in the number of tasks and their size. Figure 1a shows the distribution of the number of servers obtained by considering $|setW| = 100$ workloads tested. Further analysis suggests that the number of servers $mOPT$ (the minimum number of servers required for a QoS of 100%) follows a log-normal distribution. This result is confirmed using the normality test of D'Agostino with a p-value around 0.2. Figure 1b shows the distribution of the number of servers for a given expected QoS. Notably, a 0.2% reduction in the number of servers results in a significant decrease in the

value of m, which subsequently converges to a constant value. However, further sensitivity analysis indices will be explored to provide more accurate insights about the QoS on the number of servers m. Based on these two observations, the objective is to generalize the guarantee presented above



(a) Distribution of the number m of servers for $|setW| = 100$ workloads.



(b) Distribution of the number m of servers for variable QoS.

for any given set $setW$, such that the guarantee no longer depends on the chosen set. This study is conducted in two phases: (1) Evaluating the impact of the workload on the sizing when the expected quality of service is at its maximum (100%). The goal is to precisely determine which factor, whether the number of submitted jobs or their size, has the greatest impact on the resulting sizing; (2) Evaluating the impact of the fixed number of servers on the final QoS. The aim is to clarify the impact of the workload on the resulting QoS for a given fixed sizing. To do so, a global sensitivity analysis is performed. The focus is on the study of the impact of the input parameter variation on the variability of the output in the whole domain of interest [4].

Among the existing GSA approaches we focus on: Variance based method where the idea is to quantify how the variability in model outputs can be attributed to the variations in input parameters, this can be performed using sensitivity indices such as Sobel indices [5]. Second approach is the screening methods where the variations in the output is measured for different variation of the input parameter. The third approach is based on Shapely coefficients and is more suitable for correlated model inputs [3]. In this work, it is attested that the size of job has greater impact on the servers numbers than the number of jobs. Further sensitivity analysis indices will be explored to accurately measure the impact of each variable.

References:

[1] Manal Benaissa, Georges da Costa, and Jean-Marc Nicod. "Standalone Data-Center Sizing Combating the Over-Provisioning of the IT and Electrical Parts". In: *Workshop on Cloud Computing (WCC 2022) @ SBAC-PADW 2022*. Bordeaux, France, Nov. 2022, pp. 1–8. URL: https://hal.science/hal-03876011.

[2] Georges da Costa, Léo Grange, and Inès de Courchelle. "Modeling and Generating Large-Scale Google-Like Workload". In: *International Workshop on Resilience and/or Energy-Aware Techniques for High-Performance Computing*. Hangzhou, China, Nov. 2016. DOI: 10.1109/IGCC.2016.7892623. URL: https://laas.hal.science/hal-01472021.

[3] B. L. Nelson E. Song and J. Staum. "Shapley effects for global sensitivity analysis: theory and computation". In: *SIAM/ASA Journal on Uncertainty Quantification 4.1* (2016). DOI: 10.1137/15M1048070.

[4] A. Saltelli et al. "Global Sensitivity Analysis". In: *The Primer. John Wiley and Sons, Ltd* (2007). DOI: 10.1002/9780470725184.

[5] Ilya M. Sobol. "Sensitivity Estimates for Nonlinear Mathematical Models". In: *Mathematical Modelling and Computational Experiments* 4 (1993).