

Parsimonious Edge Computing to Reduce Microservice Resource Usage

Mathieu Simon*, Alessandro Spallina[†], Loïc Dubocquet*, Andrea Araldo*

*Télécom SudParis - Institut Polytechnique de Paris, first-name.last-name@telecom-sudparis.eu

[†]Dept. of Electrical, Electronics and Information Engineering, University of Catania

Abstract—Cloud Computing (CC) is the most prevalent paradigm under which services are provided over the Internet. The most relevant feature for its success is its capability to promptly scale service based on user demand. When scaling, the main objective is to maximize as much as possible service performance. Moreover, resources in the Cloud are usually so abundant, that they can be assumed infinite from the service point of view: an application provider can have as many servers it wills, as long it pays for it.

This model has some limitations. First, energy efficiency is not among the first criteria for scaling decisions, which has raised concerns about the environmental effects of today’s “wild” computations in the Cloud. Moreover, it is not viable for Edge Computing (EC), a paradigm in which computational resources are distributed up to the very edge of the network, i.e., co-located with base stations or access points. In edge nodes, resources are limited, which imposes different *parsimonious* scaling strategies to be adopted.

In this work, we design a scaling strategy aimed to instantiate, *parsimoniously*, a number of microservices sufficient to guarantee a certain Quality of Service (QoS) target. We implement such a strategy in a Kubernetes/Docker environment. The strategy is based on a simple Proportional-Integrative-Derivative (PID) controller. In this paper we describe the system design and a preliminary performance evaluation.

I. INTRODUCTION

Despite the many improvements in resource optimization for cloud computing, these remote systems fall victim to unexpected variations in service demands. In order to always guarantee service requirements, overprovisioning of computation resources can be adopted in the Cloud, with consequent economic and environmental costs. Due to the limited resources in the Edge, overprovisioning is not an option for Edge Computing (EC), where instead allocation strategies should be *parsimonious*, i.e., use only the resources that are needed by a certain service and not more than those.

Microservice architectures have been proposed to easily maintain and scale a service with respect to the demand and have also been considered for EC [1]. While infrastructure owners/administrators usually aim to guarantee a good QoS we believe that, additionally, microservice architectures are an opportunity to achieve *parsimony* and finely tune the resource consumption of an application by dynamically allocating the right amount of microservices. In our vision, parsimony is particularly important in edge nodes, where computational resources are limited (not infinite, as assumed in the cloud). In such cases, if multiple applications co-exists in the same edge node, it is crucial that each of them consumes the right amount

of resources, not more, in order to leave room to the other applications. At the same time, energy consumption can be dramatically reduced by reducing the number P of processing microservices running [2], [3]. However, if the number P of microservices is reduced too much, service requirements could not be met anymore. Therefore, a cautious control law must be adopted in order to find the “right” P . The calculation of this law must be light enough not to impact resource consumption at the edge. The process of scaling up and down the number P of microservices is called *horizontal scaling*.

Our contribution in this paper is to demonstrate the feasibility of *Parsimonious Edge Computing* (PEC). To this aim, we present a proof of concept of a microservice architecture managed to limit the consumed resources. We implement it in a Docker / Kubernetes-orchestrated environment, where clients generate a load of requests and multiple microservices can serve such requests. Requests are placed in a queue. A control law, based on Proportional-Integrative-Derivative (PID), performs horizontal scaling with the goal of maintaining such a queue on a certain target level. Differently from adopting PID in the Cloud, where we can afford complex monitoring systems [4], we propose a lightweight architecture.

Preliminary results obtained in a small scale testbed show that the system is able to dynamically orchestrate creation/removal of microservices in order to center the target. Our architecture and experiments are reproducible and the code is released as open source [5].

II. MODEL

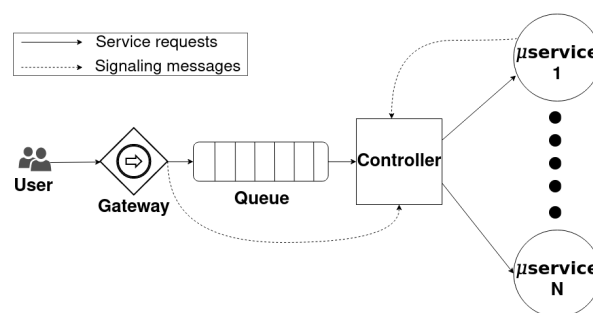


Fig. 1. Design of the model

Fig. 1 depicts the implemented system. Users generate a sequence of service requests, e.g., antivirus calculations, or

livecast video transcoding offloaded from end devices to an edge node [6]–[8]. Requests are placed at the edge in a queue. There are P microservices, denoted with μ services hereafter, able to process such requests. While in our experiments they are Docker containers, they could also be ephemeral execution environments in the context of serverless computing. Immediately after a μ service becomes free, it dequeues a request and serves it, which takes a certain processing time. In other words, our system is a multiserver queue.

A *Controller* dynamically adjusts the number P of μ services. The controller monitors at several instants t the queue size $W(t)$, i.e., how many requests are still waiting in the queue to be served. Based on this observation, it adjusts the number P of μ services. This adjustment is calculated with a simple PID control law. In particular, the Proportional–Integral–Derivative controller (PID) issues a *correction* $P_{out}(t)$ (see (2)), i.e., how many μ services should be added (if $P_{out} > 0$) or removed (if $P_{out} < 0$). This correction $P_{out}(t)$ is added to the number $P(t)$ of currently deployed μ services and gives the number of μ services wanted (see (1)), i.e., the ones that are necessary to drive the queue length $W(t)$ close to a pre-fixed target T (which is equivalent to keep the error (3) as low as possible). The control law is summarized by the following equations:

$$\mu\text{services wanted } P_w(t) = P(t) + P_{out}(t) \quad (1)$$

$$\text{Control variable } P_{out}(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{d}{dt} e(t) \quad (2)$$

$$\text{Error value } e(t) = W(t) - T \quad (3)$$

By imposing the number $W(t)$ to be close to the target, we can limit the response time since we prevent the queue to explode. Consider a low latency application, with real time requirements, i.e. requests have deadlines of few milliseconds and failing to serve them within the deadline is equivalent to losing the requests. In this case, we can set the target queue size T to a very small number. Our control law will aim to achieve such a target by deploying *just the right amount* of μ services, in order not to over-use computational resources.

III. IMPLEMENTATION

We create a Docker/Kubernetes cluster that implements the concept of Parsimonious Edge Computing. In what follows, we denote with *italic* font the *service requests* generated by users (simply *request* hereafter) and with underlined font the signaling messages needed to manage the system.

a) Service request handling.: We emulate **users**, by generating a sequence of *service requests* directed to the **Gateway**. Although we do not deal with encryption in this preliminary work, in the future extensions the Gateway would act as TLS termination proxy. This latter (i) places the *requests* into a Kafka **queue** [9] of one topic (therefore, the gateway acts as a “Kafka producer”) and (ii) sends a notification to the Controller, via another HTTP GET, with an ID of the



Fig. 2. Number of requests in queue and μ services (moving average, 10 samples per window).

request. The gateway increments these IDs at every request so that consecutive requests have consecutive IDs. The Controller is a Docker container, which implements a web server able to handle HTTP requests. The controller is the only Kafka “consumer”, i.e., it dequeues *requests* from the queue via Kafka Python API [10]. The Controller computes the queue length $W(t)$ as the difference between the ID of the last arrived *request* in the queue and the ID of the lastly dequeued *request*. The μ service that processes the *service requests* is a microservice implemented as a Docker container. We implement a pull-based mechanism: (i) whenever the μ service becomes free (it has finished to process the previous *request*), it asks for a new *requests* by sending an HTTP GET to the Controller, (ii) the Controller dequeues a *request* from the queue and (iii) sends it to the μ service via an HTTP Reply.

b) Dynamic microservice deployment.: Each μ service has an ID assigned by the Controller, which keeps a list of active μ services.

Whenever the Controller receives a notification by the Gateway of a new *request* arriving in the queue, it checks the number $P(t)$ of active μ services and compare it with the number $P_w(t)$ of μ services wanted, established by PID control law. If $P_w(t) > P(t)$, the Controller creates $P_w(t) - P(t)$ new μ services, assigning new IDs using Kubernetes primitives. Suppose that t_1 and t_2 are the arrival times of two consecutive requests. In the interval $t \in [t_1, t_2]$, by construction (1)–(3), $W(t) \leq W(t_1)$ because requests can only leave the queue. Therefore, $e(t) \leq e(t_1)$, i.e., the PID law would never trigger creation of new μ services between two service requests arrivals.

At any instant t when a μ service s asks the Controller for a new *service request*, the latter first checks if $P_w(t) < P(t)$: in this case, there are in the system more μ services than needed and the controller destroys μ service s via Kubernetes primitives. This mechanism ensures that we never destroy any μ service that is currently processing a *request*.

IV. EVALUATION

We generate requests with exponentially distributed inter-arrival time with mean 1 second. The time for one μ service to process one request is exponentially distributed with mean 5 seconds. We fix the target T of requests waiting in the queue

to 25. Based on observation, we tune the PID parameters to $K_p = -0.9$, $K_i = 0$ and $K_d = -0.2$. Note that the creation of μ services requires some non-negligible startup time [11], and thus the effects of the control values coming from the PID controller are delayed. Moreover, to keep you monitoring environment lightweight, we only get the current queue length $W(t)$ when new requests arrive or a μ service asks for a new request to consume. Therefore, we sample the queue at non-regular time intervals. Moreover, there is an imbalance between creation and destruction of μ services: we destroy them, at most one at a time, when a μ service has finished processing its current request. On the other hand, we can create many μ services in one shot. This results in the visible imbalance between the above-target and the below-target values of $W(t)$ in Fig. 2. Despite such difficulties, our system is able to get the queue length $W(t)$ around the target.

V. OTHER HORIZONTAL SCALING STRATEGIES FOR MICROSERVICE ARCHITECTURES

The subject of our work is horizontal autoscaling (Sec. I) to which we limit our discussion. Vertical autoscaling [12] is out of scope.

The standard horizontal scaling strategy in Kubernetes [13] requires the administrator to decide a threshold on CPU utilization. Once a μ service exceeds such a threshold, new μ services are created. However, there is no clear relation between application QoS and CPU utilization, so that it is difficult, if not impossible, for the administrator to set the right threshold for different classes of applications. As a consequence, if the administrator wants to be sure a certain QoS is satisfied, he needs to set stringent threshold, which may deploy more μ services than actually needed, and which goes against the parsimony that we advocate here.

In Knative autoscaling [14], instead, the administrator sets a “concurrency” value x , and the strategy deploys the amount of μ services so that each one gets x requests. As for the standard Kubernetes horizontal scaling, understanding what is the right value to set is challenging. Similarly, OpenFaaS triggers μ services based on thresholds, defined on incoming requests per second (RPS). It has been found that such concurrency- or RPS- based approaches are inefficient to meet request demand while limiting the number of triggered μ services [15, Sec. 5].

The horizontal scaling of a recent strategy, called Libra [16], scales μ services based on heuristic rules, which consists in thresholds on the requests served per second and request serving time. It will be interesting in our future work to compare our control-theoretical low with Libra’s heuristic.

VI. CONCLUSION

We have proposed a lightweight architecture and control law for *Parsimonious Edge Computing*. We have shown an implementation based on Docker/Kubernetes/Kafka and released it as open source. We have numerically verified that our system manages to meet some pre-defined system-level performance target while limiting the amount of microservices, via dynamic allocation/destruction. In future work, we will

compare our strategy with other horizontal scaling techniques (Sec. 5). We will implement with this model different classes of real services, with different requirements (low latency, high throughput). We will also consider different concurrent services running at the same node and competing for resources. Other control strategies, e.g., Reinforcement Learning, will be considered.

VII. ACKNOWLEDGMENT

This work has been carried out in the context of the Chaire “Les Réseaux du Future pour les Services de Demain” at the Très Haut Débit (THD) Platform of Télécom SudParis.

REFERENCES

- [1] L. Ma, S. Yi, and Q. Li, “Efficient service handoff across edge servers via docker container migration,” in *Proceedings of the Second ACM/IEEE Symposium on Edge Computing*, 2017, pp. 1–13.
- [2] E. A. Santos, C. McLean, C. Solinas, and A. Hindle, “How does docker affect energy consumption? Evaluating workloads in and out of Docker containers,” *Journal of Systems and Software*, vol. 146, no. July, pp. 14–25, 2018.
- [3] P. E. N, F. J. P. Mulerickal, B. Paul, and Y. Sastri, “Evaluation of docker containers based on hardware utilization,” in *2015 International Conference on Control Communication Computing India (ICCC)*, Nov 2015, pp. 697–700.
- [4] M. Cerqueira De Abranches and P. Solis, “An algorithm based on response time and traffic demands to scale containers on a Cloud Computing system,” in *IEEE International Symposium on Network Computing and Applications, NCA*, 2016.
- [5] M. Simon, “Parsimonious Edge Computing: GitHub Repository,” <https://github.com/hemesipp/MDP-container-scaling-prevision>, 2021.
- [6] Y. Kim, H. W. Lee, and S. Chong, “Mobile Computation Offloading for Application Throughput Fairness and Energy Efficiency,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 3–19, 2019.
- [7] T. Q. Dinh, Q. D. La, T. Q. Quek, and H. Shin, “Learning for computation offloading in mobile edge computing,” *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6353–6367, 2018.
- [8] Q. He, C. Zhang, X. Ma, and J. Liu, “Fog-Based Transcoding for Crowdsourced Video Livecast,” *IEEE Communications Magazine*, vol. 55, no. 4, pp. 28–33, 2017.
- [9] R. Shree, T. Choudhury, S. C. Gupta, and P. Kumar, “Kafka: The modern platform for data management and analysis in big data domain,” in *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*. IEEE, 2017, pp. 1–5.
- [10] “Kafka Python client,” <https://github.com/dpkp/kafka-python>, 2020.
- [11] C. C. Spoiala, A. Calinciuc, C. O. Turcu, and C. Filote, “Performance comparison of a webrtc server on docker versus virtual machine,” in *2016 International Conference on Development and Application Systems (DAS)*. IEEE, 2016, pp. 295–298.
- [12] G. Rattihalli, M. Govindaraju, H. Lu, and D. Tiwari, “Exploring potential for non-disruptive vertical auto scaling and resource estimation in kubernetes,” in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*. IEEE, 2019, pp. 33–40.
- [13] T.-T. Nguyen, Y.-J. Yeom, T. Kim, D.-H. Park, and S. Kim, “Horizontal pod autoscaling in kubernetes for elastic container orchestration,” *Sensors*, vol. 20, no. 16, p. 4621, 2020.
- [14] “Configuring Knative Serving autoscaling,” https://docs.openshift.com/container-platform/4.5/serverless/knative_serving/configuring-knative-serving-autoscaling.html.
- [15] J. Li, S. G. Kulkarni, K. Ramakrishnan, and D. Li, “Understanding open source serverless platforms: Design considerations and performance,” in *Proceedings of the 5th International Workshop on Serverless Computing*, 2019, pp. 37–42.
- [16] D. Balla, C. Simon, and M. Maliosz, “Adaptive scaling of Kubernetes pods,” in *EEE/IFIP Network Operations and Management Symposium - Management in the Age of Softwarization and Artificial Intelligence*, no. 1, 2020, pp. 8–12.