Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Performance of a trunk reservation policy in distributed data centres

Fabrice Guillemin [a],*, Guilherme Thompson [b]

[a] CNC/NCA, Orange Labs2, Avenue Pierre Marzin, 22300 Lannion, France
[b] INRIA Paris, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France

ARTICLE INFO

Article history:
Received 18 September 2017
Received in revised form 27 June 2018
Accepted 5 July 2018
Available online xxxx

Keywords:
Virtualized network function
Mobile edge computing
Random walk
Boundary value problem

ABSTRACT

In this paper, we analyse a system composed of two data centres with limited capacity in terms of computing (namely virtual cores). When one request for a single core is blocked at the first data centre, this request is forwarded to the second one. To protect the requests originally assigned to the second data centre, a trunk reservation policy is introduced (i.e., a redirected request is accepted only if the number of free cores at the second data centre is greater than a given threshold). After rescaling the system when there are many cores in both data centres and high request arrival rates, we are led to analyse a random walk in the quarter plane, which has the particularity of having nonconstant reflecting conditions on one boundary of the quarter plane. Computing the stationary distribution of the present random walk requires the determination of three unknown functions, one polynomial and two infinite generating functions. We show that the coefficients of the polynomial are solutions to a linear system. After numerically solving this linear system, we are able to compute the two other unknown functions and the blocking probabilities at both data centres. Numerical experiments are eventually performed to estimate the gain achieved by the trunk reservation policy.

1. Introduction

The emergence of new networking paradigms such as Network Function Virtualization (see for instance ETSI White Paper [1]) and Mobile (or Multi-access) Edge Computing (MEC) [2] incites network operators to deploy distributed data centres at the edge of their networks. This trend is naturally in line with the Fog computing framework [3–5]. In some sense, Fog computing is an evolution of Cloud computing, which has proved very efficient for the past decade to host on-line applications. The objective of Fog computing and distributed data centres is to support on-line services with more stringent latency requirements (CPU offloading, gaming, etc.) and to perform computations at the edge of the network instead of increasing the load of the core network by transmitting data up to centralised platforms.

Because of the potentially large number of data centres to be deployed at the edge of the network, these data centres may have much smaller capacity than big centralised data centres generally used in Cloud computing. In this paper, we focus on data centres, which are indeed much smaller than those of centralised cloud platforms but are nevertheless equipped with sufficient capacity to host Virtualized Network Functions (VNFs) and MEC applications. To be more specific, with new architectures based on Next Generation Points of Presence [6] or on Main Central Offices (MCOs) combined with Core Central Offices (CCOs) [7], edge data centres may comprise a few thousands of vCPUs. Such numbers are indeed small when compared

* Corresponding author.
E-mail addresses: Fabrice.Guillemin@orange.com (F. Guillemin), Guilherme.Thompson@inria.fr (G. Thompson).

https://doi.org/10.1016/j.peva.2018.07.002
0166-5316/© 2018 Elsevier B.V. All rights reserved.

with millions of vCPUs in big data centres but this is sufficient to investigate scaling limits. In addition, we suppose that the demand for resources in edge data centres is sufficiently high and volatile so that the load of some edge data centres may reach high values. Under these assumptions, it may happen that some user requests are blocked if resources are exhausted at an edge data centre. This is a key difference with cloud computing, where resources are often considered as infinite. In this article, we consider the case, where users request computing units (virtual cores) available in a data centre. If no cores are available, then a user request may be blocked.

To reduce the blocking probability, it may be suitable that data centres at the edge of the network collaborate. This is certainly a key issue, which makes the design of distributed data centres very different from that of centralised big cloud platforms. Along this line of investigations, an offloading scheme has been investigated in Fricker et al. [8], where requests blocked at a data centre are forwarded to another one with a given probability. Motivated by the design of resource allocation schemes for distributed data centres hosting virtualised network functions and mobile edge computing applications, we investigate in this paper the case when a blocked request is systematically forwarded to another data centre. To protect those requests, which originally arrive at this data centre, a redirected request is accepted only if there is a sufficient large number of idle cores. In the framework of telephone networks, such a policy is known as *trunk reservation* [9].

In the following, we consider the case of two data centres, where the trunk reservation policy is applied in one data centre only; the analysis of the case when the policy is applied in both data centres is a straightforward extension of the case considered in this paper but involves much more intricate computations. We further simplify the system by reasonably assuming that both data centres have a sufficiently large number of virtual cores, as discussed above, even if the capacity is not infinite. From a theoretical point of view, this leads us to rescale the system and to consider limiting processes, see Kelly [10] for instance. The goal of the present analysis is to estimate the gain achieved by the trunk reservation policy under overload conditions.

Considering the number of free cores in both data centres, we are led after rescaling to analyse a random walk in the positive quarter plane. This kind of process has been extensively studied in the technical literature (see for instance the book by Fayolle et al. [11] and that by Cohen and Boxma [12]). For the random walk appearing in this article, even if the kernel is similar to that analysed in Fayolle and Iasnogorodski [13] (and more recently in Fricker et al. [8]), the key difference is that the reflecting conditions on the boundaries of the quarter plane are not constant. More precisely, the reflecting coefficients in the negative vertical direction along the y-axis take two different values depending upon a given threshold (namely, the trunk reservation threshold).

This simple difference between the random walk considered in this paper and classical ones makes the analysis much more challenging. Contrary to the usual case, which consists of determining two unknown functions, we have in the present case to decompose one unknown function into two pieces (one polynomial and one generating function with an infinite number of terms) and thus to determine three unknown functions. We show that the coefficients of the unknown polynomial can be computed by solving a linear system. Once this polynomial is determined, the two other functions can be determined. This eventually allows us to compute the blocking probabilities at the two data centres and to estimate the efficiency of the trunk reservation policy for data centres hosting VNFs and MEC applications.

The contribution of the present paper is precisely to formulate for the trunk reservation system under consideration the boundary value problem exhibiting three unknown functions (two infinite generating functions and a polynomial) and to establish functional relations between these functions. This eventually allows us to determine the linear system satisfied by the coefficients of the unknown polynomial. It is then possible to numerically solve this linear system and to estimate the blocking probabilities at the two data centres.

It is worth noting that queueing systems solvable through boundary value problems exhibiting more than two unknown generating functions have already been studied in the technical literature. This is notably the case for the shorter queue polling model analysed in [14] and the longest or shortest queue models considered in [15,16]. Moreover, boundary value problems arising in telecommunication systems have for instance been studied in [17–20]. With regard to [8], the model in the present paper implements an admission control policy based on trunk reservation. This was not the case in [8], where any deflected request can enter the data centre.

This paper is organised as follows: In Section 2, we introduce the model and the notation, and we show convergence results for the rescaled system. In Section 3, we analyse the limiting random walk, in particular its kernel. The associated boundary value problems are formulated and solved in Section 4; the proof of some technical results as well as some elements of the theory of Riemann–Hilbert problems are deferred to the Appendix. To evaluate the gain achieved by the trunk reservation policy, some numerical results are discussed in Section 5. Concluding remarks are presented in Section 6.

2. Model description and rescaled system

2.1. Model description

We consider in this paper two data centres in series. In the architecture of distributed data centres in the framework of 5G networks, the first data centre is intended to host MEC applications and is located at the very edge of the network; this corresponds to the MCO [7]. The second data centre (namely a CCO) is located higher in the network, say, at a Point of Presence, and hosts VNFs. We assume that the computing capacity (in the form of virtual cores) is the limiting resource and there are sufficient memory and disk capacities. The above tasks occupy some computing resources, namely a single virtual