



HAL
open science

Modeling Dynamic Resource Allocation in the Edge

Juan Antonio Cordero, Wei Lou

► **To cite this version:**

Juan Antonio Cordero, Wei Lou. Modeling Dynamic Resource Allocation in the Edge. The 8th IFIP/IEEE International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN), Nov 2019, Paris, France. 10.23919/PEMWN47208.2019.8986967. hal-03172297

HAL Id: hal-03172297

<https://polytechnique.hal.science/hal-03172297>

Submitted on 17 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modeling Dynamic Resource Allocation in the Edge

Juan Antonio Cordero
École Polytechnique, France, EU
juan-antonio.cordero-fuertes@polytechnique.edu

Wei Lou
The Hong Kong Polytechnic University, Hong Kong SAR, PRC
csweilou@comp.polyu.hk.edu

Abstract—Edge computing technologies and integrated architectures have been deployed to accommodate Internet traffic growth. These architectures include facilities (cloudlets, micro DCs) to cache and serve contents close to consumers. Resulting systems adapt to observed request/consumption patterns by allowing cloudlet coordination for content caching/dissemination. This paper presents a novel analytical model of transient dynamics of the cloudlets set. The model is used to study system convergence, stability and delivered content locality. Results from this model are validated via simulations.

Index Terms—ICN; dynamic; content; model; wired; wireless

I. INTRODUCTION

Two major trends can be noted in the Internet evolution. First, cloud-based services have significantly expanded. Second, networking and computing capabilities of wireless and mobile devices have substantially improved. This allows new applications to be deployed, by leveraging more powerful and pervasive devices in the Internet (mostly wireless) edge.

Internet traffic and communication patterns are also evolving, from the traditional host-centric, end-to-end paradigm, to the information-centric networking (ICN) paradigm. This is being acknowledged and enabled by industry, which leads towards new technological developments, e.g. IPv6-Centric Networking (6CN) or Cisco Hybrid ICN.

The resulting traffic growth, and its increasing variability, put under pressure current networking infrastructures, as noticed by network operators [1], both in the Internet core, in the wireless last mile, or in the access network backhauls [2]. This has encouraged the emergence of edge or fog computing architectures [3], [4], and various integrated infrastructures. In these architectures, contents, and resources for data processing or computing, are placed in the Internet “edge”, *i.e.*, closer to the end devices that request and consume them. Figure 1 displays an example of this integrated architecture for mobile Internet access. Modelling resource allocation and distribution in the edge computing infrastructure, in particular transient behavior in highly-dynamic scenarios, is thus useful to understand and optimize its performance, and allow cloud-based services and content consumption to scale on the Internet.

A. Related Work

Edge computing architectures [4], and similar integrated wired/wireless architectures, such as fog computing or cloudlet-based architectures [5], have been proposed to accommodate the emergence of cloud-based services, (multimedia)

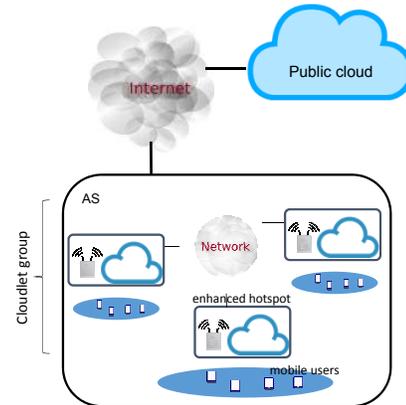


Fig. 1. Internet mobile access architecture diagram.

content consumption and IoT – and the corresponding explosion of Internet edge traffic.

The basic structure of edge computing infrastructures has three tiers [2]: end device, local cloud (*cloudlet*) or micro-datacenter, and public cloud. Based on this hierarchy, different interactions have been explored [2], [6]–[9].

In-network caching and resource offloading are key primitives of integrated architectures. Caching strategy, content popularity estimation, and modelling, have been largely explored in ICN literature, as they have a deep impact in ICN systems performance. Different caching strategies have been proposed [10]–[13], and their performance and main trade-offs (between caching size in the network, bandwidth consumption, delivery delay, route optimality, and heuristic complexity) evaluated, mostly through simulations. Analytical modeling of caching algorithms’ performance has been addressed in some contributions [14], mainly focusing on the steady behavior of the examined ICNs. To the best of our knowledge, however, no substantial efforts have been devoted to the modeling of caching transient dynamics in the edge of information-centric integrated architectures.

B. Contribution

The contribution of this paper is three-fold. First, a model for dynamic content placement within edge networks is presented. Second, system stability, convergence and traffic locality in edge networks are studied analytically, based on this model. Third, these theoretical results are validated by way of discrete-event simulations of the edge network.

II. SYSTEM DESCRIPTION

End devices (user equipments or “things”) are connected to the Internet by way of gateways (*enhanced hotspots* in Figure 1, e.g. wireless access points deployed by Internet access providers) equipped with devices with storage and/or computing capabilities (cloudlets, micro-datacenters) – denoted *cloudlets* hereafter. Gateways (with cloudlets) are connected to each other through a local network or an Autonomous System of the provider – i.e. communication between gateways does not entail Internet traffic. The set of interconnected gateways with cloudlets is the *edge network*.

End devices request and retrieve Internet contents (or resources) through the edge network. Contents belong to a finite catalog and are always available from one or more Internet providers (clouds, datacenters), located outside the edge network [12] – i.e., from the *public cloud*. Cloudlets may store locally (cache) requested contents in order to speed up service, and to reduce traffic. For simplicity, it is first assumed that caches can store a *single* content.

Contents in the edge network are placed in function of content popularity in the communities served by gateways. A content is moved from one cloudlet to another if its popularity in the latter cloudlet is higher than in the former. A content overrides another content in the same cloudlet if the popularity of the former is higher than the latter’s. This is a simplification of the Last Frequently Used (LFU) caching policy.

A. System States, Inputs and Notation

N is the number of available contents, M is the number of cloudlets/caches in the system. $H(\cdot)$ denotes the Heaviside function: $H(x) = 1$ if $x \geq 0$, 0 otherwise.

The *state* of the system is described by the state of its cloudlets. For N contents and M caches (typically, $N \gg M$), the set of all possible states, S , is contained in $\{c_1, c_2, \dots, c_N, 0\}^M$, where c_i denotes the i -th content and 0 denotes that no content is allocated.

The following modeling assumptions hold:

- Requests for a content i arrive at cache j , as a Poisson homogeneous process with rate $r_{i,j}$ (in requests/sec).
- All contents have the same size. Content homogeneity is a common assumption in CCN literature [13], [15].
- At most, one content replica is allowed in the system.

This assumption follows the approach of restricting the number of replicas in the system [10]–[12].

Given these assumptions, valid system states can be classified in (i) *void state*, $(0, 0, \dots, 0)$, in which no content is allocated in the system, (ii) *individual states*, in which a single content is allocated in the system, (iii) *incomplete states*, in which several caches (but not all) are allocated, and (iv) *complete states*, in which all caches are allocated.

The *inputs* of the system are described by a matrix $\mathbf{R} = (r_{i,j})_{N \times M}$, where $r_{i,j}$, the request rate for content i received at cache j , is also denoted *content-cache rate* (i, j) (measured in requests/sec). The sum of all request rates in the system is $\Sigma = \sum_{i,j} r_{i,j}$. The maximum request rate is $m = \max_{i \leq N, j \leq M} r_{i,j}$.

1) *Specific Notation for $M = 2, N = 2$ Case:* In the case of $M = 2$ contents, $N = 2$ caches, and given a content i , the other content (not i) is the *complementary content* of i , and denoted \bar{i} . Symmetrically, given a cache j , the other cache (not j), is the *complementary cache* of j , \bar{j} .

For a content-cache $x = (i, j)$, $\bar{x} = (\bar{i}, \bar{j})$ denotes the complementary content of x (in the same cache); $x_c = (i, \bar{j})$, the content of x in the complementary cache; and $\bar{x}_c = (\bar{i}, j)$, the complementary content of x in the complementary cache.

B. Evaluation Criteria

1) *System Transitions:* There is a *system transition* when the state of the system changes upon a request arrival. Two types of transitions can be observed:

- *Internal transition*, when the set of contents stored in the system after the transition is contained in the set of contents stored in the system before the transition.
- *External transition*, when a new content (previously not present in the system) is installed.

2) *Convergence Time:* A system has ϵ -converged at time t_0 when the probability that a system transition occurs at $t \geq t_0$ is lower or equal to ϵ . The ϵ -convergence time is the minimum time for which the system has ϵ -converged.

Empirically, the K -convergence time is the minimum time at which K consecutive arrivals have been received in the system without causing a transition (see II-B1). Both definitions are used (the probabilistic one in section III, the empiric one in section VI) to study the convergence evolution with respect to other parameters. In the limit, ϵ - and K -convergences tend ($K \rightarrow \infty, \epsilon \rightarrow 0$) to the same value.

3) *Content Delivery Locality:* When a cache receives a content request, this content can be served:

- *Locally*, from the contacted cache.
- *Group-locally*, from a cache from another cloudlet (not the one receiving the request) within the edge network.
- *Remotely*, from the Internet (public cloud).

III. BASIC CASE: $N = 2, M = 2$

The system with $N = 2$ contents, $M = 2$ caches is a simplification of real systems that allows a detailed modelling (in this section) and analysis (sec. IV).

For $N = 2, M = 2$, α and β denote the two contents that can be requested. The set of states is then $S = \{(0, 0), (\alpha, 0), (0, \alpha), (\beta, 0), (0, \beta), (\alpha, \beta), (\beta, \alpha)\} = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6\}$. The content-cache rates in this case are then denominated, for convenience, $r_{1,1} = \alpha_1, r_{1,2} = \alpha_2, r_{2,1} = \beta_1$ and $r_{2,2} = \beta_2$.

A. State Occupancy

The probability that the system is in state s_i at time t is $s_i(t) \in [0, 1]$. Dynamics of this system *after* the first arrival (i.e. excluding state $s_0 = (0, 0)$), assumed at $t = 0$, are studied through the evolution of $s_i(t)$. Initial conditions for this system are: $s_1(0) = \alpha_1/\Sigma, s_2(0) = \alpha_2/\Sigma, s_3(0) = \beta_1/\Sigma$, and $s_4(0) = \beta_2/\Sigma$, with $\Sigma = \alpha_1 + \alpha_2 + \beta_1 + \beta_2$. The system of ordinary differential equations (ODEs) showing state

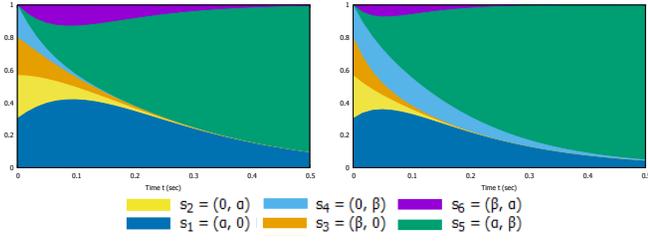


Fig. 2. Probabilities of state occupancy. $\alpha_1 = 9, \alpha_2 = 8, \beta_1 = 7, \beta_2 = 6$ req/sec (left (a)), and $\alpha_1 = 9, \alpha_2 = 7, \beta_1 = 6, \beta_2 = 8$ req/sec (right (b)).

occupancy evolution is described in Theorem 1. (Proofs are skipped due to space limitations.)

Theorem 1. *State occupancies for the cloudlet group ($M = 2, N = 2$) are described by the system of ODEs (1).*

$$(1) \quad \begin{cases} \frac{d}{dt} s_1(t) = -s_1(t)(\beta_2 + \beta_1 H(\beta_1 - \alpha_1) + \alpha_2 H(\alpha_2 - \alpha_1)) \\ \quad + s_2(t)\alpha_1 H(\alpha_1 - \alpha_2) + s_3(t)\alpha_1 H(\alpha_1 - \beta_1) \\ \quad + s_6(t)\alpha_1 H(\alpha_1 - \alpha_2) H(\alpha_1 - \beta_1) \\ \frac{d}{dt} s_2(t) = -s_2(t)(\beta_1 + \beta_2 H(\beta_2 - \alpha_2) + \alpha_1 H(\alpha_1 - \alpha_2)) \\ \quad + s_1(t)\alpha_2 H(\alpha_2 - \alpha_1) + s_4(t)\alpha_2 H(\alpha_2 - \beta_2) \\ \quad + s_5(t)\alpha_2 H(\alpha_2 - \alpha_1) H(\alpha_2 - \beta_2) \\ \frac{d}{dt} s_3(t) = -s_3(t)(\alpha_2 + \alpha_1 H(\alpha_1 - \beta_1) + \beta_2 H(\beta_2 - \beta_1)) \\ \quad + s_4(t)\beta_1 H(\beta_1 - \beta_2) + s_1(t)\beta_1 H(\beta_1 - \alpha_1) \\ \quad + s_5(t)\alpha_1 H(\alpha_1 - \alpha_2) H(\alpha_1 - \beta_1) \\ \frac{d}{dt} s_4(t) = -s_4(t)(\alpha_1 + \alpha_2 H(\alpha_2 - \beta_2) + \beta_1 H(\beta_1 - \beta_2)) \\ \quad + s_3(t)\beta_2 H(\beta_2 - \beta_1) + s_1(t)\beta_2 H(\beta_2 - \alpha_2) \\ \quad + s_6(t)\alpha_2 H(\alpha_2 - \alpha_1) H(\alpha_2 - \beta_2) \\ \frac{d}{dt} s_5(t) = \beta_2 s_1(t) + \alpha_1 s_4(t) \\ \quad - \alpha_2 s_5(t) H(\alpha_2 - \alpha_1) H(\alpha_2 - \beta_2) \\ \quad - \beta_1 s_5(t) H(\beta_1 - \alpha_1) H(\beta_1 - \beta_2) \\ \frac{d}{dt} s_6(t) = \beta_1 s_2(t) + \alpha_2 s_3(t) \\ \quad - \alpha_1 s_6(t) H(\alpha_1 - \alpha_2) H(\alpha_1 - \beta_1) \\ \quad - \beta_2 s_6(t) H(\beta_2 - \alpha_2) H(\beta_2 - \beta_1) \end{cases}$$

Corollary. *Given any constant input matrix \mathbf{R} , the system (1) is a linear system of ODEs with constant coefficients. The Initial Value Problem (IVP) of (1) with its initial conditions has therefore a unique solution.*

ODEs for individual states $\{s_1, \dots, s_4\}$ share the same structure, as well as ODEs for the complete states $\{s_5, s_6\}$. The system is symmetric both in terms of contents and caches. Solutions for this system depend on \mathbf{R} and are described in Proposition 1. Figure 2 shows the probability of state occupancy with respect to time, for two examples of rate configurations, representative of the two possible cases described in Proposition 1.

Proposition 1. *Due to system symmetry, and given a matrix \mathbf{R} of request rates, two basic cases can be distinguished, any other case being reducible to one of them by a combination of content and cache exchanges:*

- *The two most popular content-caches can be simultaneously allocated (because they correspond to different contents and different caches). Without loss of generality,*

assume the most popular content-caches are α_1 and β_2 . Then, solutions are as follows ($\Sigma = \alpha_1 + \alpha_2 + \beta_1 + \beta_2$):

$$\begin{cases} s_1(t) = \frac{e^{-\beta_2 t}}{\Sigma} (-(\alpha_2 + \beta_1)e^{-\alpha_1 t} + \alpha_1 + \alpha_2 + \beta_1) \\ s_2(t) = \frac{\alpha_2}{\Sigma} e^{-(\alpha_1 + \beta_1 + \beta_2)t} \\ s_3(t) = \frac{\beta_1}{\Sigma} e^{-(\alpha_1 + \alpha_2 + \beta_2)t} \\ s_4(t) = \frac{e^{-\alpha_1 t}}{\Sigma} (-(\alpha_2 + \beta_1)e^{-\alpha_1 t} + \beta_1 + \beta_2 + \alpha_2) \\ s_5(t) = \left((\alpha_2 + \beta_1)e^{-(\alpha_1 + \beta_2)t} - (\alpha_2 + \beta_1 + \beta_2)e^{-\alpha_1 t} \right. \\ \quad \left. - (\alpha_2 + \alpha_2 + \beta_1)e^{-\beta_2 t} \right) / \Sigma \\ s_6(t) = \frac{e^{-(\alpha_1 + \beta_2)t}}{\Sigma} (-\alpha_2 e^{-\beta_1 t} - \beta_1 e^{-\alpha_2 t} + \alpha_2 + \beta_1) \end{cases}$$

- *The two most popular content-caches cannot be simultaneously allocated. Without loss of generality, assume these are α_1 and α_2 . Then, solutions are as follows:*

$$\begin{cases} s_1(t) = \frac{\alpha_1 e^{-\beta_2 t}}{\Sigma} \left[1 + \int_0^t dx e^{(\beta_2 - \alpha_1)x} \left(\beta_2 e^{-(\beta_2 + \alpha_2)x} \right. \right. \\ \quad \left. \left. + (\alpha_2 + \beta_2)e^{-\beta_2 x} + \beta_2 e^{-(\beta_1 + \alpha_2)x} \right. \right. \\ \quad \left. \left. + (\beta_1 + \beta_2)e^{-\alpha_2 x} \right) \right] \\ s_2(t) = \frac{e^{-(\alpha_1 + \beta_1)t}}{\Sigma} (\alpha_2 + \beta_2(1 - e^{-\alpha_2 t})) \\ s_3(t) = \frac{e^{-(\alpha_1 + \alpha_2)t}}{\Sigma} (\beta_1 + \beta_2(1 - e^{-\beta_1 t})) \\ s_4(t) = \frac{\beta_2}{\Sigma} e^{-(\alpha_1 + \alpha_2 + \beta_1)t} \\ s_5(t) = \frac{\beta_2 \alpha_1}{\Sigma} \left[\int_0^t dx \left(e^{-(\alpha_1 + \alpha_2 + \beta_1)x} + e^{-\beta_2 x} \times \right. \right. \\ \quad \left. \left. \left(1 + \int_0^x dy e^{(\beta_2 - \alpha_1)y} \left(-\beta_2 e^{(\alpha_2 - \beta_1)y} \right. \right. \right. \right. \\ \quad \left. \left. \left. + (\alpha_2 + \beta_2)e^{-\beta_1 y} - (\alpha_2 + \beta_2)e^{-\beta_1 y} \right. \right. \right. \right. \\ \quad \left. \left. \left. + \alpha_2 + \beta_1 + \beta_2 \right) \right) \right] \\ s_6(t) = \frac{e^{-\alpha_1 t}}{\Sigma} \left(e^{-\beta_1 t} (\alpha_2 + \beta_2) + \beta_2 e^{-(\alpha_2 + \beta_1)t} \right. \\ \quad \left. - (\beta_1 + \beta_2)e^{-\alpha_2 t} + \alpha_2 + \beta_1 + \beta_2 \right) \end{cases}$$

B. Discussion

For m denoting the most popular (*i.e.*, with highest rate) content-cache in the system:

- Unsurprisingly, the system converges to the state $\{m, \bar{m}_c\}$ (see section II-A1), that is, a state in which the most popular content-cache is allocated together with its complementary (both in content and in cache).
- Before convergence, two different transient evolutions can be distinguished. If the two most-requested content-caches are compatible (*i.e.*, they can be simultaneously allocated), individual states $\{m, 0\}$ ($s_1 = (\alpha, 0)$ in Figure 2) and $\{m_c, 0\}$ ($s_2 = (0, \alpha)$ in Figure 2) are more probable than (suboptimal) $\{m_c, \bar{m}\}$ ($s_6 = (\beta, \alpha)$ in Figure 2) – this is the case displayed in Figure 2(b). Otherwise, $\{m_c, \bar{m}\}$ is more probable than any individual state (except $\{m, 0\}$), and the system shows a slower convergence – this is the case in Figure 2(a).

IV. ANALYSIS OF $N = 2, M = 2$ SYSTEM

The $N = 2, M = 2$ system is a 4-dimensional system, described by matrix \mathbf{R} (dimensions 2×2). The evolution of the evaluation criteria presented in section II-B depends on these 4 dimensions, and on time. In practice, content-cache rates are not independent. This section explores different dependencies between content-cache rates (projections) to show the behavior of the analyzed parameters:

- **Rate factor**, $F \in [1, \infty]$. Given a fixed base configuration \mathbf{R} , the parameter analysis with respect to F allows to study the impact of increasing the total amount of requests ($F \times \mathbf{R}$) in the system behavior, while preserving the spectrum (set of eigenvalues) of the input matrix.
- **Content discovery factor**, $x = \sqrt{\frac{\alpha_1}{\beta_1}} = \sqrt{\frac{\alpha_2}{\beta_2}}$. In x -based analysis, relationship between contents in both caches is the same. Higher values of x correspond to more unbalanced distribution of content requests in the system.
- **Cache discovery factor**, $y = \sqrt{\frac{\alpha_1}{\alpha_2}} = \sqrt{\frac{\beta_1}{\beta_2}}$. In y -based analysis, relationship between caches is content-independent. Higher values of y correspond to less even distribution of requests at caches in the system.

Analysis of the evolution of evaluation criteria on the F axis and the xy plan are denoted as F - and xy -analysis, respectively.

A. System Transitions

From the state occupancy probabilities, the average number of internal transitions, $N_I(t)$, and external transitions, $N_E(t)$, correspond to the solution to the following Initial Value Problems (IVPs):

$$\begin{cases} \frac{d}{dt} N_E(t) = s_1(t) (\beta_2 + \beta_1 H(\beta_1 - \alpha_1)) \\ \quad + s_2(t) (\beta_1 + \beta_2 H(\beta_2 - \alpha_2)) \\ \quad + s_3(t) (\alpha_2 + \alpha_1 H(\alpha_1 - \beta_1)) \\ \quad + s_4(t) (\alpha_1 + \alpha_2 H(\alpha_2 - \beta_2)) \\ N_E(0) = 1 \end{cases} \quad (2)$$

$$\begin{cases} \frac{d}{dt} N_I(t) = s_1(t) \alpha_2 H(\alpha_2 - \alpha_1) + s_2(t) \alpha_1 H(\alpha_1 - \alpha_2) \\ \quad + s_3(t) \beta_2 H(\beta_2 - \beta_1) + s_4(t) \beta_1 H(\alpha_1 - \beta_2) \\ \quad + s_5(t) [\beta_1 H(\beta_1 - \alpha_1) H(\beta_1 - \beta_2) \\ \quad + \alpha_2 H(\alpha_2 - \beta_2) H(\alpha_2 - \alpha_1)] \\ \quad + s_6(t) [\beta_2 H(\beta_2 - \alpha_2) H(\beta_2 - \beta_1) \\ \quad + \alpha_1 H(\alpha_1 - \beta_1) H(\alpha_1 - \alpha_2)] \\ N_I(0) = 0 \end{cases} \quad (3)$$

Figures 3 and 4 show the xy -analysis for the number of external and internal transitions, that is, the evolution of $N_E(t)$ and $N_I(t)$ (for a fixed time $t = 1$) depending on the content and cache diversity factors (x and y). Figure 3 shows that the maximum number of external transitions occurs when content rates are equal (content and cache diversity factors $x, y = 1$), and decreases as cache or content rates diverge. When rates are similar, system states are likely to oscillate and content replacements are frequent within the system, therefore contents need to be retrieved more often from the Internet.

Internal transitions behave differently. When caches are comparable ($y \rightarrow 1$), increases in content diversity entail more internal transitions. On the contrary, for caches serving at very different rates, the number of internal transitions decreases exponentially as content rates diverge. Content allocation is more stable (*i.e.* the number of internal transitions is

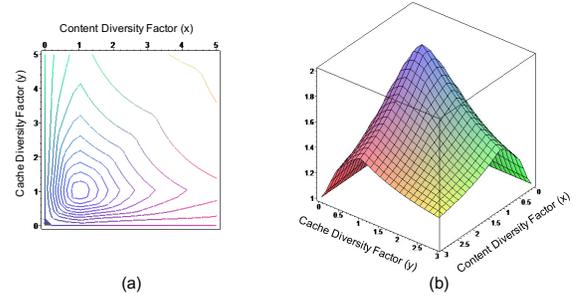


Fig. 3. Average number of external transitions, $N_E(t)$, for different configurations ($\alpha_1 = xy, \alpha_2 = y/x, \beta_1 = x/y, \beta_2 = 1/(xy)$, $x, y \in (0, 3]$), at a fixed time $t = 1$. Contour lines (left, (a)), and 3-D view (right, (b)).

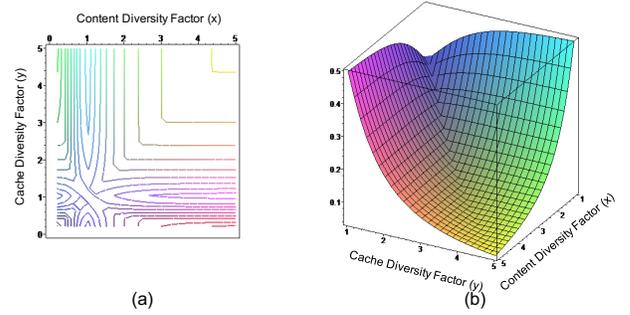


Fig. 4. Average number of internal transitions, $N_I(t)$, for different configurations ($\alpha_1 = xy, \alpha_2 = y/x, \beta_1 = x/y, \beta_2 = 1/(xy)$, $x, y \in [1, 5]$), at a fixed time $t = 1$. Contour lines (left, (a)), and 3-D view (right, (b)).

minimum) when cache diversity and content diversity are the same ($x = y$).

Figure 5(a) displays the F -evolution of transitions. The number of transitions increases with factor F , but is bounded by a constant limit, determined by the relation between rates.

B. System Convergence

In an $M = 2, N = 2$ system, transitions are distributed over time according to the following function $f_{tr}(t)$:

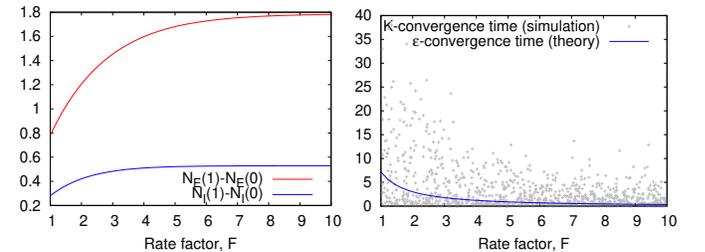


Fig. 5. Impact of the rate factor F in the average number of external and internal transitions, $N_E(t)$ and $N_I(t)$, at time $t = 1$ (left, (a)) and the ϵ -convergence time, for $\epsilon = 0.01$, in sec (right, (b)); configuration $\alpha_1 = 0.9F, \alpha_2 = 0.8F, \beta_1 = 0.7F, \beta_2 = 0.6F$, for $F \in [1, 10]$.

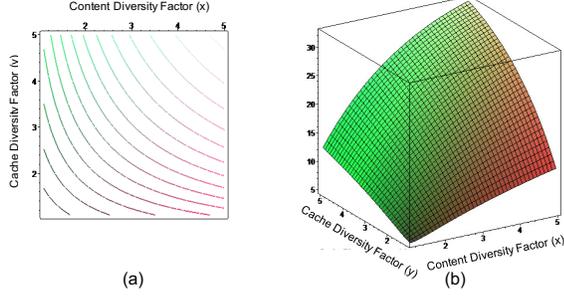


Fig. 6. ϵ -convergence time, for $\epsilon = 0.01$ (in sec), in function of content and cache diversity factors, x and y ($\alpha_1 = xy$, $\alpha_2 = y/x$, $\beta_1 = x/y$, $\beta_2 = 1/(xy)$). Contour lines (left, (a)), and 3-D view (right, (b)).

$$\begin{aligned}
f_{tr}(t) = & [s_1(t)(\beta_2 + \beta_1 H(\beta_1 - \alpha_1) + \alpha_2 H(\alpha_2 - \alpha_1)) \\
& + s_2(t)(\beta_1 + \beta_2 H(\beta_2 - \alpha_2) + \alpha_1 H(\alpha_1 - \alpha_2)) \\
& + s_3(t)(\alpha_2 + \alpha_1 H(\alpha_1 - \beta_1) + \beta_2 H(\beta_2 - \beta_1)) \\
& + s_4(t)(\alpha_1 + \alpha_2 H(\alpha_2 - \beta_2) + \beta_1 H(\beta_1 - \beta_2)) \\
& + s_5(t)\beta_1 H(\beta_1 - \alpha_1) H(\beta_1 - \beta_2) \\
& + s_5(t)\alpha_2 H(\alpha_2 - \beta_2) H(\alpha_2 - \alpha_1) \\
& + s_6(t)\beta_2 H(\beta_2 - \alpha_2) H(\beta_2 - \beta_1) \\
& + s_6(t)\alpha_1 H(\alpha_1 - \beta_1) H(\alpha_1 - \alpha_2)] / \Sigma
\end{aligned} \quad (4)$$

Eq. (4) allows to compute convergence time, which decreases exponentially with the rate factor F (Figure 5(b)).

Figure 6 shows the xy -evolution. Convergence time is minimum when content and cache diversity factors are 1, *i.e.* when all content-caches are requested at the same rate, and grows as content rates or cache requests become more different (higher diversity factors). A balanced configuration thus leads to a faster convergence of the edge network, as expected. This is consistent with the evolution of the F -analysis of the average number of system transitions, depicted in Figure 5(a).

C. Content Delivery Locality

From the state occupancy probabilities, the average number of local content deliveries, $N_L(t)$, edge computing & access network deliveries, $N_G(t)$, and remote content deliveries, $N_R(t)$, correspond to the solution to the following IVPs:

$$\begin{cases} \frac{d}{dt} N_L(t) &= \alpha_1 s_1(t) + \alpha_2 s_2(t) + \beta_1 s_3(t) + \beta_2 s_4(t) \\ &+ s_5(t)(\alpha_1 + \beta_2) + s_6(t)(\alpha_2 + \beta_1) \\ N_L(0) &= 0 \end{cases} \quad (5)$$

$$\begin{cases} \frac{d}{dt} N_G(t) &= \alpha_2 s_1(t) + \alpha_1 s_2(t) + \beta_2 s_3(t) + \beta_1 s_4(t) \\ &+ s_5(t)(\alpha_2 + \beta_1) + s_6(t)(\alpha_1 + \beta_2) \\ N_G(0) &= 0 \end{cases} \quad (6)$$

$$\begin{cases} \frac{d}{dt} N_R(t) &= (\beta_1 + \beta_2)(s_1(t) + s_2(t)) \\ &+ (\alpha_1 + \alpha_2)(s_3(t) + s_4(t)) \\ N_R(0) &= 1 \end{cases} \quad (7)$$

From IVP (7), $\frac{d}{dt} N_R(t) \rightarrow 0$ when $t \rightarrow \infty$, as only s_5 or s_6 can be converging states ($\lim_{t \rightarrow \infty} s_i(t) = 0$ for $i \leq 4$).

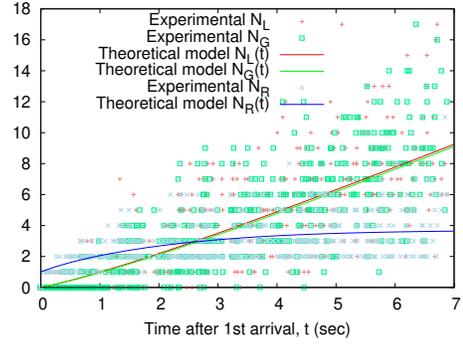


Fig. 7. Time evolution of $N_L(t)$, $N_G(t)$, and $N_R(t)$, for rates $\{\alpha_1 = 0.9, \alpha_2 = 0.8, \beta_1 = 0.7, \beta_2 = 0.6\}$, experimental and theoretical values.

	c_1	c_2	c_3	c_4
CR1	8.92	6.74	4.54	11.24
CR2	14.69	3.79	7.57	8.96
CR3	1.00	2.75	8.54	14.02

TABLE I

CONTENT-CACHE RATES OF A FULL SYSTEM EXAMPLE.

V. TOWARDS A MODEL GENERALIZATION

The model in section III allows to study the evolution of a system with $N = 2$ contents and $M = 2$ caches, by way of states $\{s_1, \dots, s_6\}$. For $N > 2$ and/or $M > 2$, the number of states in the system grows super-exponentially ($\sim O\left(\left(1 + \frac{M}{N-M}\right)^N (N-M)^M\right)$, with Stirling approximation), thus becoming mathematically intractable.

State and evolution of individual content-caches (i, j) in systems with $N \geq 2$, $M \geq 2$ (denoted *full system*), can be however analyzed by approximating the full system as a *restricted* 2×2 system. In this *restricted system*, two classes of contents (one including content i , the other including any other content), and two classes of caches (one including cache j , the other including any other cache), are considered. This approximation allows to reuse the basic model (section III), with only 3 additional states (9 in total), and thus derive *bounds* on the behavior of the studied content-cache.

Figure 8 shows the evolution of the number of transitions in content-cache (1, 2), as observed experimentally (see section VI), and the (lowest) bound obtained from the restricted system approximation, for a $(N = 4, M = 3)$ system with rates detailed in Table I (in bold, the examined content-cache).

VI. SIMULATIONS

An event-driven simulator has been developed in Python to test experimentally the behavior of the described family of systems, and compare to the analytical performance.

A. System with $N = 2$ Caches, $M = 2$ Contents

Figure 9 shows the number of external and internal transitions in 500 simulations of the $(N = 2, M = 2)$ system, for different running times. Figure 7 shows the number of contents delivered locally, group-locally, and remotely. Experimental results are consistent with solutions of IVPs in Eqs. (2)-(3).

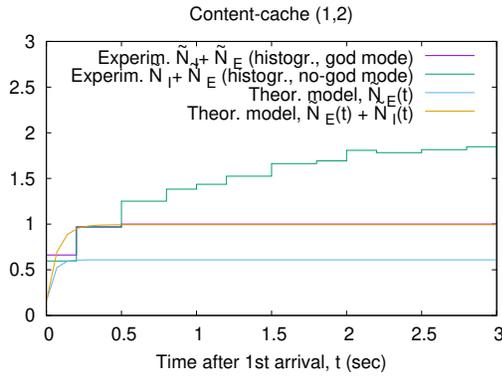


Fig. 8. Number of external and internal transitions for content-cache (1, 2): values from the generalized model, and experimental results on the full system.

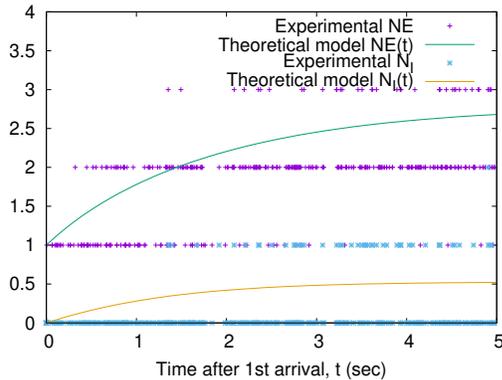


Fig. 9. Number of external and internal transitions, theoretical and experimental values. Basic case ($N = 2, M = 2$) with rates $\{\alpha_1 = 0.9, \alpha_2 = 0.8, \beta_1 = 0.7, \beta_2 = 0.6\}$, $N_E(t)$ and $N_I(t)$.

B. Convergence Time

Figure 5(b) shows the evolution of the K -convergence time, for $K = 50$, in 500 simulations of a 2×2 system with configuration $\{\alpha_1 = 0.9F, \alpha_2 = 0.8F, \beta_1 = 0.7F, \beta_2 = 0.6F\}$, for $F \in [1, 10]$. The F -analysis of K -convergence time (experimental) and ϵ -convergence time (theoretical) agree on the exponential decay with respect to F .

C. Generalized Model Evaluation: $N > 2, M > 2$

Figure 8 shows transitions observed at simulations, for a system with (*god mode*) and without (*no-god mode*) accurate estimations of content-cache rates. Histograms display the average number of transitions for each 0.25-sec step. The approximation from the generalized model provides a lower-bound for the full system actual (experimental) behavior. This is a close bound for god-like experiments, and a conservative bound otherwise, as systems with no prior rate knowledge experience substantially more transitions than god-like systems.

VII. CONCLUSION

This paper has proposed a modelling methodology and a probabilistic finite-states model for describing the dynamics of content dissemination within an edge computing & access network. Provided analysis, detailed for LRU, can be extended

to other caching heuristics (*e.g.*, LRU or Random Replacement principles). Time evolution of the system state has been probabilistically described, and closed expressions have been derived for locality of content delivery, system stability and convergence of content distribution in a simple case ($N = 2$ contents, $M = 2$ cloudlets). The study of this basic case confirms fundamental trade-offs in these systems. In particular, transitions are shown to be minimum in balanced systems (with low content diversity and similar cache traffic rates), and with bounds depending on relative values of traffic rate.

A model generalization for systems with arbitrary caches and contents ($M \geq 2, N \geq 2$), is sketched, that allows to determine analytical performance bounds for large systems.

Further work includes formalization and in-depth analysis of the conditions in which the model generalization can approximate large systems' behavior. Simulations indicate that performance is affected by inaccuracies in content rates estimation – a study of the impact of non-ideal estimations in the model behavior is needed. The model should be extended to capture additional system parameters (*e.g.*, cache size).

ACKNOWLEDGEMENTS

This work is supported by the Hong Kong RGC (GRF PolyU-521312), Hong Kong Polytechnic University (4-BCB6, G-YBJU, G-YBXY, M-N020, G-UAH6), and National Natural Science Foundation of China (No. 61272463). It was also supported in part by the Cisco-Polytechnique Chaire “Internet of Everything” (<http://www.Internet-of-Everything.fr>).

REFERENCES

- [1] F. Rebecchi, M. D. D. Amorim, V. Conan, A. Passarella, R. Bruno, M. Conti, Data offloading techniques in cellular networks: A survey, *IEEE Communications Surveys and Tutorials* 17 (2014) 580–603.
- [2] J. Zhang, T. Xiong, W. Lou, Community clinic: Economizing mobile cloud service cost via cloudlet group, in: *Proc. MASS'14*, 2014.
- [3] M. Satyanarayanan, The emergence of edge computing, *IEEE Computer* 50 (2017) 30–39.
- [4] W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: Vision and challenges, *IEEE Internet of Things Journal* 3 (2016) 637–646.
- [5] M. Satyanarayanan, Z. Chen, K. Ha, W. Hu, W. Richter, P. Pillai, Cloudlets: at the leading edge of cloud-mobile convergence, in: *Proc. 6th MobiCASE*, 2014.
- [6] Q. Xia, W. Liang, W. Xu, Throughput maximization for online request admission in mobile cloudlets, in: *Proc. LCN'13*, 2013.
- [7] D. T. Hoang, D. Niyato, P. Wang, Optimal admission control policy for mobile cloud computing hotspot with cloudlet, in: *Proc. WCNC*, 2012.
- [8] Y. Li, W. Wang, Can mobile cloudlets support mobile applications?, in: *Proc. INFOCOM'14*, 2014.
- [9] J. A. Cordero, W. Lou, Take your time, get it closer: content dissemination within mobile pedestrian crowds, *Wireless Networks* 25 (2019) 3385–3403.
- [10] C. Barakat, A. Kalla, D. Saucez, T. Turletti, Minimizing bandwidth on peering links with defection in named data networking, in: *Proc. ICCIT'13*, 2013.
- [11] W. K. Chai, D. He, I. Psaras, G. Pavlou, Cache “less for more” in information-centric networks, in: *Proc. IFIP Networking 2012*, 2012.
- [12] S. Guo, H. Xie, G. Shi, Collaborative forwarding and caching in content centric networks, in: *Proc. IFIP Networking 2012*, 2012.
- [13] V. Sourlas, L. Gkatzikis, P. Flegkas, L. Tassioulas, Distributed cache management in information-centric networks, *IEEE Trans. on Networks and Service Management* 10 (3) (2013) 286–299.
- [14] G. Carofiglio, M. Gallo, L. Muscariello, D. Perino, Modeling data transfer in content-centric networking, in: *Proc. ITC 2011*, 2011.
- [15] Y. Li, H. Xie, Y. Wen, Z. L. Wang, Coordinating in-network caching in content-centric networks: Model and analysis, in: *Proc. ICDCS'13*, 2013.