2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Stability of non-cooperative load balancing with time-varying latency

Alessandro Giuseppi[1], Danilo Menegatti[1] and Antonio Pietrabissa[1]

Abstract— **The problem of non-cooperative load balancing arises in multi-agent scenarios where users/services compete for some limited resources. This study, leveraging on results from set stability and switched systems control theory, analyses the convergence properties of a class of load-balancing strategies towards a set of approximated non-cooperative equilibria in a scenario in which the performance of the resource providers is described by a time-varying latency function.**

Index Terms— **switched systems; Wardrop equilibrium; non-cooperative load balancing.**

I. Introduction

Dynamic load balancing is among the most impactful processes for optimizing service provision in numerous application domains [1]–[3]. In load balancing, a stream of infinitely-many users, each one demanding some resources (in terms of jobs/load per unit of time), have to be distributed over some providers. In the non-cooperative setting, such users compete for the usage of the providers' resources, implying that the balancing strategy of each agent is autonomous and is taken unilaterally with the sole objective of minimizing some measure of *latency* experienced (e.g., queue length, delay, price) [4].

Wardrop equilibria have been broadly studied for both routing and load balancing problems ([5]–[8] and references threrin), with most works focusing on time-invariant setting, with the exception of some studies focusing on time-varying loads [9] and network topologies [10]. The present paper employs the load-balancing problem formulation originally used in [11] and in the previous work [10] with the addition of time-varying latency functions. Up to the authors' knowledge on non-cooperative routing and load balancing – including the cited papers and our previous study [9], addressing the case of a time-varying demand – the main innovation of the present work is related to the analysis of the convergence conditions under time-varying latency functions, which may reflect, e.g., changes in the providers' operating modes for energy-saving, security or malfunction occurrences. This result is obtained by means of set stability theory and Lyapunov arguments applied to the switched systems domain. Specifically, the variations of the latency functions will determine *switching events*, leading to the identification of a minimum *dwell-time* (i.e., a minimum time interval between two switches) that assures the convergence of the system to an approximate Wardrop equilibrium in finite-time.

II. Preliminaries

A. Load balancing and Wardrop equilibria

Given a finite set of providers $\mathcal{P}$, let $\mathcal{I}$ be the set of $|I|$ commodities, each one characterized by its job demand or load rate $\lambda^i > 0$, $i \in \mathcal{I}$ representing the amount of load per unit of time to be elaborated by the set of providers $\mathcal{P}^i \subseteq \mathcal{P}$ over which the i-th commodity may distribute its load, and let $\lambda := \sum_{i\in\mathcal{I}} \lambda^i$ be the total load demand, assumed to consist of infinitely many decision-making agents, each responsible for its allocation strategy [10].

Let x_p^i and $x_p = \sum_{i\in\mathcal{I}} x_p^i$ be the fraction of the load of commodity i and of the overall load of all commodities allocated on provider $p \in \mathcal{P}$. The *load vector* is then defined as $\mathbf{x} = [x_p^i]_{p\in\mathcal{P}^i, i\in\mathcal{I}}$ and the state space can be expressed as:

Definition 2.1: Given a load demand vector $\boldsymbol{\lambda} = [\lambda^i]_{i\in\mathcal{I}}$, the feasible state space is the set of the feasible load vectors:

$$\mathcal{X} := \Big\{ \mathbf{x} = [x_p^i]_{p\in\mathcal{P}^i, i\in\mathcal{I}} \Big| x_p^i \geq 0, \forall p \in \mathcal{P}^i, \sum_{p\in\mathcal{P}^i} x_p^i = \lambda^i, \forall i \in \mathcal{I} \Big\}. \qquad (1)$$

Load balancing problems depend on the definition of a performance index on which a balancing criterion may be evaluated which, in the Wardrop framework, takes the name of *latency functions*. Such functions are used to evaluate and capture the performance of the given state of the system and are typically related to some application-specific quantity (e.g., the average response time to process a set amount of load).

As customary in the literature [9], [12], we assume that latency functions are non-negative and strictly increasing for all $p \in \mathcal{P}$, as they evaluate the degrading performance of the providers as a function of its loads x_p. In addition, we assume no explicit knowledge on the their structure, as their nature depends heavily on the application and the considered performance. We then consider a broad class of latency functions limited only the following assumption.

Assumption 2.2: The latency functions $l_p : [0, \lambda] \to \mathbb{R}_{\geq 0}$, for all $p \in \mathcal{P}$, are Lipschitz continuous and strictly increasing over the interval $[0, \lambda]$.

A *stable* load vector corresponds to a network state, known as a Wardrop equilibrium, in which no commodity may unilaterally (i.e., without cooperation) improve its load allocation.

Definition 2.3: (from [12]) Under a given load rate $\boldsymbol{\lambda}$, a Wardrop equilibrium is defined as a state in which the load vector $\mathbf{x} \in \mathcal{X}$ is such that $l_p(x_p) \leq l_m(x_m)$ $\forall p \in \mathcal{P}^i$ for which $x_p^i > 0 \forall m \in \mathcal{P}^i$ and $\forall i \in \mathcal{I}$.

[1] University of Rome La Sapienza, Via Ariosto 25, 00185, Rome, Italy. email: [giuseppi,menegatti,pietrabissa]@diag.uniroma1.it

This work has been partially supported by the University of Rome La Sapienza in the scope of the project "Intelligent Control Systems: Analysis, Design and Applications", n. RP1221816BE0C9AB.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

In practice, following [13], Wardrop equilibria are characterized by the fact that the latencies experienced by the agents (i.e., the latencies of the providers $p \in \mathcal{P}$ such that $x_p > 0$) have the same value $\forall i \in \mathcal{I}$. Thus, it is possible to define the set of Wardrop equilibria as

$$\mathcal{W}^{\mathbf{l}} := \Big\{ \mathbf{x} \in \mathcal{X} \big| l_p(x_p) - l_m(x_m) \leq 0, \forall p, m \in \mathcal{P}^i \\ \text{s.t. } x_p^i > 0, \forall i \in \mathcal{I} \Big\}, \tag{2}$$

where the apex $\mathbf{l}$ indicates that the set depends on the latency functions and $\mathbf{l} := [l_p]_{p \in \mathcal{P}}$ is a vector of latency functions satisfying Assumption 2.2.

We employ the Beckmann potential [14], as conventional in the Wardrop literature:

$$\Phi^{\mathbf{l}}(\mathbf{x}) = \sum_{p \in \mathcal{P}} \int_0^{x_p} l_p(\xi) d\xi. \tag{3}$$

Property 2.4: Under Assumption 2.2, the Beckmann potential (3) is continuous and the following properties hold:

1) there exists a unique feasible load vector, denoted by $\mathbf{w}^{\mathbf{l}} \in \mathcal{X}$, that minimizes $\Phi^{\mathbf{l}}(\mathbf{x})$, with $\mathbf{x} \in \mathcal{X}$;
2) correspondingly, there exist a unique, positive minimum of $\Phi^{\mathbf{l}}(\mathbf{x})$, denoted by $\Phi^{\mathbf{l}}_{min} := \Phi^{\mathbf{l}}(\mathbf{w}^{\mathbf{l}}) > 0$.
3) $\mathbf{w}^{\mathbf{l}}$ is at the Wardrop equilibrium and, therefore, the equilibrium set collapses into $\mathcal{W}^{\mathbf{l}} = \{\mathbf{w}^{\mathbf{l}}\}$.

In this work, we will consider a load balancing problem with latency functions that vary over time. These changes induce a change of the set of equilibria 2, which will reflect on the control strategy and overall system dynamics. Therefore, in the following subsection we will introduce some notions on switched systems useful to model the time-varying problem.

B. Switched systems

The set of providers with time-varying latency functions and the load balancing problem will be modelled as a switched nonlinear system [15]. Following the approach of [15], [16], we consider our switching system to be defined by a time-dependent switching rule and a set of different flows of a continuous-time dynamical system, indexed with r:

$$\dot{\mathbf{x}}(t) = f^{(r)}(\mathbf{x}(t)), \tag{4}$$

where $f^{(r)}$ takes the name of *flow* of the r-th system.

Considering a family of distinct systems of the form (4) whose indexes are contained in the set $\mathcal{R} \subseteq \mathbb{N}$, the discrete-time *switch* that makes the system flow change from the one associated to the current index $r \in \mathcal{R}$ to the $(r+1)$-th one takes the name of *switching event*. Switching events are orchestrated by a *switching signal*, which in our setting will be driven by the uncontrollable variations of the providers' latency functions.

Let $\mathcal{X}$ be a bounded invariant set for the flow

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t)\big), \mathbf{x}(t) \in \mathcal{X}, \forall t \geq 0, \tag{5}$$

and let $\mathcal{A}$ be a closed subset of $\mathcal{X}$. Denoting with $d(\mathbf{u}, \mathbf{v})$ the euclidean distance between the two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, the distance between a point $\mathbf{x} \in \mathcal{X} \setminus \mathcal{A}$ and the set $\mathcal{A}$ is defined as as $d(\mathbf{x}, \mathcal{A}) = \min_{\mathbf{y} \in \mathcal{A}} d(\mathbf{x}, \mathbf{y})$.

Definition 2.5: [17] The function $V : \mathcal{X} \to \mathbb{R}_{\geq 0}$ is *positive definite* with respect to a closed set $\mathcal{A} \subset \mathcal{X}$ if there exists an increasing continuous function $\Psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\Psi(0) = 0$ and $\Psi\big(d(\mathbf{x}, \mathcal{A})\big) \leq V(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{A}$.

Theorem 2.6: [18] Given a closed set $\mathcal{A} \subset \mathcal{X}$, if there exists a continuously differentiable function $V : \mathcal{X} \to \mathbb{R}_{\geq 0}$ positive definite with respect to $\mathcal{A}$, such that its derivative along the trajectories of the system (5), $\dot{V}(\mathbf{x})$, is so that $-\dot{V}(\mathbf{x})$ is also positive definite with respect to $\mathcal{A}$, then $V(\mathbf{x})$ is a Lyapunov function on $\mathcal{X} \setminus \mathcal{A}$ and $\mathcal{A}$ is a Globally Asymptotically Stable Set (GASS) for the system.

As a consequence of the theorem, it follows that if $\exists \gamma > 0$ such that $\dot{V}(\mathbf{x}) < -\gamma$ $\forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{A}$, the state of the system enters $\mathcal{A}$.

III. Main Result

A. Modelling and dynamics

The *switching signal* that characterizes the system dynamics is defined as the piece-wise constant function

$$s : \mathcal{T} \to \mathcal{R}, \tag{6}$$

with $\mathcal{T} = [0, \infty)$. Signal (6) then maps all time instants t onto the set of indexes $\mathcal{R}$. Let $\tau_r \in \mathcal{T}$ denote the r-th switching instant, with $\tau_0 = 0$ and $\tau_{r-1} < \tau_r$ for all $r \in \mathcal{R}$. Let $l_p^{(r)}$ denote the provider latency functions, for $t \in [\tau_r, \tau_{r+1})$, and let $H^{(r)} := \tau_{r+1} - \tau_r$ denote the r-th *holding time* of the switching signal.

The considered switched system

$$\dot{\mathbf{x}}(t) = f^{s(t)}(\mathbf{x}(t)), t \in \mathcal{T}, \tag{7}$$

is the result of the switching signal (6) orchestrating the switches over the continuous-time dynamics

$$\dot{\mathbf{x}}(t) = f^{(r)}(\mathbf{x}(t)), t \in \mathcal{T}, \mathbf{x}(0) \in \mathcal{X}, \tag{8}$$

that are parametrized by r, with $f^{(r)} : \mathcal{X} \to \mathcal{X}$.

The instant just before the r-th switch, which occurs at the end of the holding time H_{r-1}, is denoted by τ_r^-. In our system, switches occur when there is a change in provider latency, that does not affect the load vector, i.e.:.

$$\mathbf{x}(\tau_r) = \mathbf{x}(\tau_r^-). \tag{9}$$

During every flow r the latency functions satisfy Assumption 2.2. Let $\beta_p^{(r)}$ be the local Lipschitz constant of $l_p^{(r)}$, for all providers $p \in \mathcal{P}$. We make an additional assumption that states that the Lipschitz constants of the latency functions are bounded according to the following statement.

Assumption 3.1: Let $\mathcal{L}$ denote the space of the functions satisfying Assumption 2.2. For every flow $r \in \mathcal{R}$, the vector $\mathbf{l}^{(r)}$ of the providers' latency functions is such that $\mathbf{l}^{(r)} \in \mathcal{L}$.

The variation of the value of the latency functions at the switches is assumed to be bounded:

Assumption 3.2: For every load vector $\mathbf{x} \in \mathcal{X}$ and provider $p \in \mathcal{P}$, the latency variation at the switch $r \in \mathcal{R}$ is bounded as $|l_p^{(r)}(x_p) - l_p^{(r-1)}(x_p)| \leq \bar{\alpha}$, with $\bar{\alpha} > 0$.

We remark that the latencies variations of assumption 3.2 make so each flow (8) is characterized by a distinct equilibrium $\mathbf{w}^{(r)}$ that depends on its specific latency functions.

Focusing on the subsystem (8), one has that, during flow r, for all $t \in \mathcal{T}$, $p \in \mathcal{P}^i$ and $i \in \mathcal{I}$,

$$\dot{x}_p^i(t) = \sum_{m \in \mathcal{P}^i} r_{mp}^i(\mathbf{x}(t)) - \sum_{m \in \mathcal{P}^i} r_{pm}^{i,(r)}(\mathbf{x}(t)), \tag{10}$$

where $r_{pm}^{i,(r)}(\mathbf{x}(t))$ is the migration rate, directed from provider p towards provider m. It is then possible to define the total migration rate, for a given state $\mathbf{x}$, from provider p to m as

$$r_{pm}(\mathbf{x}) = \sum_{i \in \mathcal{I}} r_{pm}^{i,(r)}(\mathbf{x}). \tag{11}$$

Equation (11) defines the state-dependant rate according to which the load exchange between providers p and m occurs. A typical solution to characterize this rate is to assume its structure as follows [6], [7], [10]:

$$r_{pm}^{i,(r)}(\mathbf{x}) = \sigma_{pm}^i(\mathbf{x})\mu_{pm}^{i,(r)}(\mathbf{x})x_p^i, \forall p, m \in \mathcal{P}^i, \forall i \in \mathcal{I} \tag{12}$$

in which the variable $\sigma_{pm}^i(\mathbf{x})$ captures the concept of *sampling probability*, that is the probability according to which a fraction of the load of provider p is re-allocated onto provider m, and the variable $\mu_{pm}^{i,(r)}(\mathbf{x})$ represents the *migration policy*, that is the law that defines if a migration among the two providers actually occurs and its magnitude. In order to fully define a selfish routing policy defined by (12), it is hence needed to characterize both $\sigma_{pm}^i(\mathbf{x})$ and $\mu_{pm}^{i,(r)}(\mathbf{x})$.

In general, $\sigma_{pm}^i(\mathbf{x})$ can be described as the distribution

$$\sum_{p,m \in \mathcal{P}^i} \sigma_{pm}^i(\mathbf{x}) = 1, \text{with } \sigma_{pm}^i(\mathbf{x}) > \sigma, \forall p, m \in \mathcal{P}^i, \forall i \in \mathcal{I}, \tag{13}$$

while $\mu_{pm}^{i,(r)}(\mathbf{x})$ assumes the form

$$\mu_{pm}^{i,(r)}(\mathbf{x}) = \\ = \begin{cases} \geq \mu & \text{if } l_p^{(r)}(\mathbf{x}) - l_m^{(r)}(\mathbf{x}) > 0 \\ 0 & \text{otherwise} \end{cases}, \forall p, m \in \mathcal{P}^i, \forall i \in \mathcal{I}, \tag{14}$$

in which $\sigma > 0$ and $\mu > 0$ are fixed values that can be used to characterize the migration policy behaviour.

Several possible choices exist for both $\sigma_{pm}^i(\mathbf{x})$ and $\mu_{pm}^{i,(r)}(\mathbf{x})$, among which we mention the *uniform sampling probability*, according to which the target provider is selected according to a uniform distribution, i.e.,

$$\sigma_{pm}^i(\mathbf{x}) = \sigma^i = 1/|\mathcal{P}^i|, \tag{15}$$

and the so-called *proportional sampling probability*, according to which the probability of selecting a given target provider for the migration is proportional to its load, that is

$$\sigma_{pm}^i(\mathbf{x}) = x_m^i/\lambda^i. \tag{16}$$

Regarding $\mu_{pm}^{i,(r)}(\mathbf{x})$, a first possible choice is to consider a structure of the form:

$$\mu_{pm}^{i,(r)}(\mathbf{x}) = \begin{cases} 1 & \text{if } l_p^{(r)}(\mathbf{x}) - l_m^{(r)}(\mathbf{x}) \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

which takes the name of *better-response migration policy*, or one may consider one of its variants, such as the *linear migration policy*

$$\mu_{pm}^{i,(r)}(\mathbf{x}) = \begin{cases} \dfrac{l_p^{(r)}(\mathbf{x}) - l_m^{(r)}(\mathbf{x})}{\bar{l}} & \text{if } l_p^{(r)}(\mathbf{x}) - l_m^{(r)}(\mathbf{x}) \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{18}$$

which considers an upper-bound on the latency functions, $\bar{l}$.

For the purposes of this study, we do not restrict our analysis to a particular choice of $\sigma_{pm}^i(\mathbf{x})$ and $\mu_{pm}^{i,(r)}(\mathbf{x})$, so we assume that the migration rate (12) is defined in terms of a sampling probability that obeys (13) and an arbitrary migration policy of the form (14).

B. Subsystem convergence between switches

In this subsection we focus on the finite-time convergence of (8)-(14) to a particular set of approximate equilibria $\mathcal{W}_{\delta,\varepsilon}^{(r)}$.

Definition 3.3: For a given $0 \leq \delta < 1$ and a given $\varepsilon > 0$, the (δ, ε)-Wardrop equilibrium set under the r-th latency vector is defined as

$$\mathcal{W}_{\delta,\varepsilon}^{(r)} := \Big\{ \mathbf{x} \in \mathcal{X} \Big| l_p^{(r)}(x_p) - l_m^{(r)}(x_m) \leq \varepsilon, \forall p, m \in \mathcal{P}^i \\ \text{s.t. } x_p^i \geq \delta \lambda^i, \forall i \in \mathcal{I} \Big\} \supset \{\mathbf{w}^{(r)}\}. \tag{19}$$

In the above definition, the constant ε can be seen as the maximum tolerated distance among latency values, whereas δ represents the minimum load portion x_p^i of the commodity i that is required to consider the provider p as loaded by the commodity. From definition 3.3 it follows that at a (δ, ε)-Wardrop equilibrium, the latencies of all the δ-loaded providers of a commodity i (i.e., $p \in \mathcal{P}^i$ *s.t.* $x_p^i > \delta$) are equalized up to a tolerance of ε.

Note that, since the constraints that appear in (19) are continuous, $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ is a compact and closed subset of $\mathcal{X}$.

Definition 3.4: Let $V^{(r)} : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be the continuously differentiable function

$$V^{(r)}(\mathbf{x}) = \Phi^{\mathbf{l}^{(r)}}(\mathbf{x}) - \Phi_{\min}^{\mathbf{l}^{(r)}}. \tag{20}$$

Given a constant $c > 0$, the level set of the function $V^{(r)}$ is the contour $\partial\mathcal{V}^{(r)}(c)$ of the sublevel set

$$\mathcal{V}^{(r)}(c) := \left\{\mathbf{x} \in \mathcal{X} \middle| V^{(r)}(\mathbf{x}) \leq c\right\}. \tag{21}$$

Definition 3.5: Let $c_1^{(r)}$ and $c_2^{(r)}$ be the constants such that $\mathcal{V}_1^{(r)} := \mathcal{V}^{(r)}(c_1^{(r)})$ is the maximum sublevel set included in $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ and $\mathcal{V}_2^{(r)} := \mathcal{V}^{(r)}(c_2^{(r)})$ is the minimum sublevel set containing $\mathcal{W}_{\delta,\varepsilon}^{(r)}$, respectively.

We note that Assumption 2.2 implies that the two sets $\mathcal{V}_1^{(r)}$, $\mathcal{V}_2^{(r)}$ are unique. Furthermore, ε is respectively the minimum and maximum latency mismatch that may occur on the contours $\delta\mathcal{V}_1^{(r)}$ and $\delta\mathcal{V}_2^{(r)}$. We also mention that the specific values assumed by $c_1^{(r)}$ and $c_2^{(r)}$ depend on the latency functions, but we do not assume any knowledge on their structure, as we assume them to be unknown and only measured.

From the discussions of the previous work [9], we have that the following Lemmas and Theorem 3.8 hold:

Lemma 3.6: The set $\mathcal{X}$ is a positive invariant set for the nonlinear system (8)-(14).

Lemma 3.7: Under Assumption 2.2, for any $\delta > 0$ and $\varepsilon > 0$, the functions

$$\Psi_j\big(d(\mathbf{x}, \mathcal{V}_j^{(r)})\big) = \gamma_j^{(r)} d(\mathbf{x}, \mathcal{V}_j^{(r)}), i = 1, 2, \tag{22}$$

with

$$\begin{gathered}\gamma_1^{(r)} = \frac{c_1^{(r)}}{d_1^{(r)}}, \quad \gamma_2^{(r)} = \frac{\sigma\mu\delta\lambda_{\min}\varepsilon}{d_1^{(r)}}, \\ d_1^{(r)} = \max_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{W}_{\delta,\varepsilon}^{(r)}} d(\mathbf{x}, \mathcal{V}_1^{(r)})\end{gathered} \tag{23}$$

are such that $V^{(r)}$ and $-\dot{V}^{(r)}$ are positive definite with respect to $\mathcal{X}_{\delta,\varepsilon}^{(r)}$ and therefore $V^{(r)}$ is a Lyapunov function on $\mathcal{X}\setminus\mathcal{W}_{\delta,\varepsilon}^{(r)}$, implying, from Theorem 2.6, that the set $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ is GASS for the nonlinear system (8)-(14).

Theorem 3.8: Under Assumption 2.2, for any $\delta > 0$ and $\varepsilon > 0$, trajectories of the nonlinear system (8)-(14) enter the set $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ in finite time, with minimum convergence velocity

$$\dot{V}^{(r)}(\mathbf{x}) \leq -\sigma\mu\delta\lambda_{\min}\varepsilon, \forall\mathbf{x} \in \mathcal{X} \setminus \mathcal{W}_{\delta,\varepsilon}^{(r)}, \tag{24}$$

with $\lambda_{\min} := \min_{i\in\mathcal{I}} \lambda^i$, and the equilibrium load vector $\mathbf{w}^{(r)}$, where the latencies of all the loaded providers are equalized for every commodity, is the unique asymptotically stable equilibrium state for the nonlinear system (8)-(14).

C. Convergence of the overall switched system

We now prove that the overall system (7) is stable with respect to the union of the approximate equilibrium sets (19).

Given (9), that describes a switch event for the switched system, we have that Lemma 3.6 implies the property:

Property 3.9: The set $\mathcal{X}$ is positive invariant for the switched nonlinear system (8)-(14).

For the sake of presentation, in Figure 1 we visualize the sets $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ and the Wardrop equilibria for a toy example.

Let us define c_1, c_2 as $c_1 := \min_{r\in\mathcal{R}} \left(c_1^{(r)}\right)$ and $c_2 := \max_{r\in\mathcal{R}} \left(c_2^{(r)}\right)$. Under Assumption 2.2, one has that, $\forall r \in \mathcal{R}$, the two sets $\mathcal{V}^{(r)}(c_1)$ and $\mathcal{V}^{(r)}(c_2)$ are unique, and also that

$$\mathcal{V}^{(r)}(c_1) \subseteq \mathcal{V}^{(r)}(c_1^{(r)}) \subseteq \mathcal{W}_{\delta,\varepsilon}^{(r)} \subseteq \mathcal{V}^{(r)}(c_2^{(r)}) \subseteq \mathcal{V}^{(r)}(c_2). \tag{25}$$

We can now state the following Theorem.

Theorem 3.10: For any choice of $\delta > 0$ and $\varepsilon > 0$, if Assumption 2.2 holds and if the minimum dwell-time $H_{\min}$ verifies

$$H_{\min} > \frac{2\bar{\alpha}\lambda}{\sigma\mu\delta\lambda_{\min}\varepsilon}, \tag{26}$$

the state of the switched system (8)-(14) enters $\mathcal{W}_{\delta,\varepsilon}^{(\bar{r})}$ for a finite value of $\bar{r} \in \mathcal{R}$. Furthermore, after entering $\mathcal{W}_{\delta,\varepsilon}^{(\bar{r})}$, the evolution of the system state remain in the set $\mathcal{V}^{(r)}(c_2+2\bar{\alpha}\lambda)$, for all $t \in [\tau_r, \tau_{r+1}^-)$ and $r > \bar{r}$.

Proof: From property 3.9 it follows that $\mathbf{x}(t) \in \mathcal{X}$ $\forall t \in [\tau_r, \tau_{r+1}^-)$ and $r \in \mathcal{R}$. It is then possible to retrace the proof of Theorem 3.8 to prove that $V^{(r)}(\mathbf{x})$ and $-\dot{V}^{(r)}(\mathbf{x})$ are positive definite with respect to $\mathcal{V}^{(r)}(c_1)$ by setting $\gamma_1 := c_1/\bar{d}_1$ and $\gamma_2 := \min_{r\in\mathcal{R}} \left(b_1^{(r)}/\bar{d}_1\right)$, with $\bar{d}_1 := \max_{r\in\mathcal{R},\mathbf{x}\in\mathcal{X}} d\big(\mathbf{x}, \mathcal{V}^{(r)}(c_1)\big)$. Equation (24) guarantees that, during the holding time $H^{(r)}$, on has

$$\dot{V}^{(r)}(\mathbf{x}) \leq -\sigma\mu\delta\lambda_{\min}\varepsilon < 0, \tag{27}$$

$\forall\mathbf{x} \in \mathcal{X} \setminus \mathcal{W}_{\delta,\varepsilon}^{(r)}$ and for all $r \in \mathcal{R}$.

At the r-th switch, $V^{(r-1)}\big(\mathbf{x}(\tau_r^-)\big) - V^{(r)}\big(\mathbf{x}(\tau_r)\big)$ can be expanded as follows:

$$\begin{aligned}&V^{(r)}\big(\mathbf{x}(\tau_r)\big) - V^{(r-1)}\big(\mathbf{x}(\tau_r^-)\big) = \\ &= \sum_{p\in\mathcal{P}} \Big(\int_0^{x_p(\tau_r)} l_p^{(r)}(\xi)d\xi - \int_0^{w_p^{(r)}} l_p^{(r)}(\xi)d\xi\Big) + \\ &- \sum_{p\in\mathcal{P}} \Big(\int_0^{x_p(\tau_r^-)} l_p^{(r-1)}(\xi)d\xi - \int_0^{w_p^{(r-1)}} l_p^{(r-1)}(\xi)d\xi\Big) = \\ &= \sum_{p\in\mathcal{P}} \int_0^{x_p(\tau_r)} \Big(l_p^{(r)}(\xi) - l_p^{(r-1)}(\xi)\Big)d\xi + \\ &- \sum_{p\in\mathcal{P}} \int_0^{w_p^{(r)}} l_p^{(r)}(\xi)d\xi + \sum_{p\in\mathcal{P}} \int_0^{w_p^{(r-1)}} l_p^{(r-1)}(\xi)d\xi,\end{aligned} \tag{28}$$

where $w_p^{(r)}$ represents the load of provider p at the Wardrop equilibrium for flow r and where we considered that $x_p(\tau_r) = x_p(\tau_r^-)$ (equation (9)). Assuming that all the commodity latencies increase proportionally to $\bar{\alpha}$, it is trivial to show that the first term of equation (28) is bounded by

$$\begin{aligned}&\sum_{p\in\mathcal{P}} \int_0^{x_p(\tau_r)} \Big(l_p^{(r)}(\xi) - l_p^{(r-1)}(\xi)\Big)d\xi \leq \\ &\leq \sum_{p\in\mathcal{P}} \int_0^{x_p(\tau_r)} \bar{\alpha}d\xi \leq \bar{\alpha}\lambda.\end{aligned} \tag{29}$$

while for the other terms, the second and the third one, a similar upper-bound is built as:

$$\sum_{p\in\mathcal{P}^1}\int_0^{w_p(\tau_r)}\Big(l_p^{(r-1)}(\xi)-l_p^{(r)}(\xi)\Big)d\xi+ \\ +\sum_{p\in\mathcal{P}^2}\int_0^{w_p(\tau_{r-1})}\Big(l_p^{(r-1)}(\xi)-l_p^{(r)}(\xi)\Big)d\xi\leq\bar{\alpha}\lambda, \tag{30}$$

where $\mathcal{P}^1$ and $\mathcal{P}^2$ consider the providers such that $w_p(\tau_r)\geq w_p(\tau_{r-1})$ and $w_p(\tau_r)<w_p(\tau_{r-1})$, respectively.
As a result:

$$V^{(r)}\big(\mathbf{x}(\tau_r)\big)-V^{(r-1)}\big(\mathbf{x}(\tau_r^-)\big)\leq 2\bar{\alpha}\lambda. \tag{31}$$

In order to proceed, we need to make sure that:

$$V^{(r)}\big(\mathbf{x}(\tau_r)\big)-V^{(r-1)}\big(\mathbf{x}(\tau_r^-)\big)+ \\ +\int_{\tau_r}^{\tau_r+H^{(r)}}\dot{V}^{(r)}(\mathbf{x}(t))dt<0, \tag{32}$$

which, looking at (27) and (31), is true if

$$2\bar{\alpha}\lambda-\sigma\mu\delta\lambda_{\min}\varepsilon H_{\min}<0. \tag{33}$$

Moreover, note (26) holds, the two equations (27) and (33) imply that, $\forall r\in\mathcal{R}$, the values of the Lyapunov functions at times τ_{r+1}^- (i.e., sampled at the end of the holding times) decrease, with the only exception being cases in which the system state reaches $\mathcal{W}_{\delta,\varepsilon}^{(r)}$ within the holding time.
In fact, if $\mathbf{x}(\tau_{r+1}^-)\in\mathcal{X}\setminus\mathcal{W}_{\delta,\varepsilon}^{(r)}$ just before a switch, one has

$$V^{(r)}\big(\mathbf{x}(\tau_{r+1}^-)\big)-V^{(r-1)}\big(\mathbf{x}(\tau_r^-)\big)<-h,\forall r\in\mathcal{R}, \tag{34}$$

in which $h:=H_{\min}-\frac{2\bar{\alpha}\lambda}{\sigma\mu\delta\lambda_{\min}\varepsilon}>0$, meaning that (34) defines a decreasing sequence.

On the contrary, $\forall r\in\mathcal{R}$ such that $\mathbf{x}(\tau_{r+1}^-)\in\mathcal{X}\setminus\mathcal{W}_{\delta,\varepsilon}^{(r)}$ one has:

1) $\mathcal{V}^{(r)}(c_1)\subseteq\mathcal{W}_{\delta,\varepsilon}^{(r)}$ and therefore $d\big(\mathbf{x}(\tau_{r+1}^-),\mathcal{W}_{\delta,\varepsilon}^{(r)}\big)\leq d\big(\mathbf{x}(\tau_{r+1}^-),\mathcal{V}^{(r)}(c_1)\big)$;
2) γ_1 is such that $d\big(\mathbf{x}(\tau_{r+1}^-),\mathcal{V}^{(r)}(c_1)\big)\leq\frac{1}{\gamma_1}V^{(r)}\big(\mathbf{x}(\tau_{r+1}^-)\big)$;
3) $d\big(\mathbf{x},\mathcal{W}_{\delta,\varepsilon}^{(r)}\big)=0$ when $\mathbf{x}\in\partial\mathcal{W}_{\delta,\varepsilon}^{(r)}$.

Thus, a finite value $\bar{r}$ exists such that the system trajectories reach the set $\mathcal{W}_{\delta,\varepsilon}^{(\bar{r})}$, proving the first part of the theorem.

We now focus on the maximum change in the value of latency functions after a switch, that is defined by (31). Assume that $\tilde{r}$ is such that $\mathbf{x}(\tau_{\tilde{r}}^-)\in\mathcal{W}_{\delta,\varepsilon}^{(\tilde{r})}$; Due to the fact that $\mathcal{W}_{\delta,\varepsilon}^{(\tilde{r})}\subseteq\mathcal{V}^{(\tilde{r})}(c_2)$, at the end of the $\tilde{r}$-th holding time one has $V(\mathbf{x}(\tau_{\tilde{r}+1}^-))\leq c_2$. After the switch, equation (31) implies that $V(\mathbf{x}(\tau_{\tilde{r}+1}))\leq c_2+2\bar{\alpha}\lambda$, meaning that, the state $\mathbf{x}(\tau_{\tilde{r}+1})$ is constrained in the set $\mathcal{V}^{(\tilde{r}+1)}(c_2+2\bar{\alpha}\lambda)$.

We have hence proved both the existence of a finite number $\bar{r}$ such that at time $\bar{t}\in[\tau_{\bar{r}},\tau_{\bar{r}+1})$ the state $\mathbf{x}(\bar{t})$ is in the set $\mathcal{W}_{(\bar{r})}^{\mathcal{L}}$, and the fact that the system trajectory $\mathbf{x}(t)$ never leaves $\mathcal{V}^{(r)}(c_2+2\bar{\alpha}\lambda)$ for all $r>\bar{r}$, concluding the proof. ■

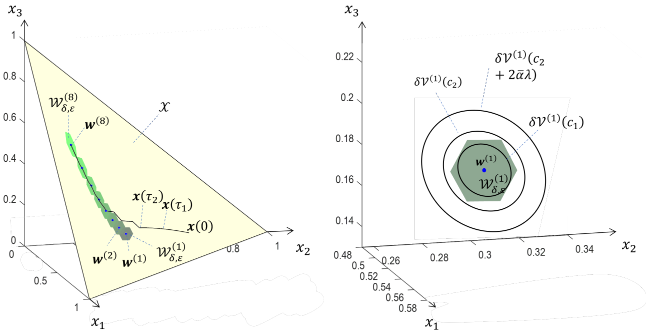

Fig. 1. (left) Single-commodity problem with $|\mathcal{P}|=3$ providers with $l_1(\xi)=e^{0.1\xi}-1$, $l_2(\xi)=e^{0.2(\xi)}-1$, $l_3(\xi)=e^{0.3(\xi)}-1$ and a (scalar) load rate $\lambda=1$. The figure details a trajectory $\mathbf{x}(t)$, the load vectors $\mathbf{w}^{(r)}$ and the sets $\mathcal{W}_{\delta,\varepsilon}^{(r)}$, for a total of eight switches, with $\delta=\varepsilon=0.005$ and $\alpha(\tau_r)=\bar{\alpha}=0.01$ for $r=1,...,8$; (right) Visualization of $\partial\mathcal{V}^{(1)}(c_1^{(1)})$, $\partial\mathcal{V}^{(1)}(c_2^{(1)})$ and $\partial\mathcal{V}^{(1)}(c_2+\bar{\beta}\bar{\alpha}^2\lambda^2)$ in the same setting of the (left) figure, for $r=1$. The maximum latency mismatches on the last two sets are 0.0076 and 0.0201.

Fig. 1 (right) depicts an example of level sets $\partial\mathcal{V}_1^{(1)}$, $\partial\mathcal{V}_2^{(1)}$ and $\partial\mathcal{V}^{(1)}(c_2+2\bar{\alpha}\lambda)$. We observe that the set $\partial\mathcal{V}^{(r)}(c_2+2\bar{\alpha}\lambda)$ is a superset of the equilibrium set $\mathcal{W}_{\delta,\varepsilon}^{(r)}$. While, from a theoretical perspective, considering the proof of Theorem 3.8, the set $\partial\mathcal{V}^{(r)}(c_2+2\bar{\alpha}\lambda)$ was found under worst-case assumptions on particularly unfavourable switches, in practice, when the r-th switch is small, we expect the trajectories to remain close or inside the equilibrium set $\mathcal{W}_{\delta,\varepsilon}^{(r-1)}$.

IV. Numerical Simulations

We now detail the result of some numerical tests to validate the control law resulting from (12), (15), (17).

We consider a scenario with $|\mathcal{P}|=8$ providers, $|\mathcal{I}|=3$ commodities, each using $|\mathcal{P}^1|=6$, $|\mathcal{P}^2|=5$ and $|\mathcal{P}^3|=5$ providers, and a load vector $\boldsymbol{\lambda}=[0.55,0.40,0.35]$ that corresponds to a total load $\lambda=1$. Moreover, during each flow $r\in\mathcal{R}$, we assume that the provider latency functions are exponential, $l_p^{(r)}(x_p)=\alpha_p(\tau_r)e^{a_p x_p}-1$, with the parameters a_p in the set $[0.1,0.25]$ and $\alpha_p(\tau_1)=1,\forall p\in\mathcal{P}$.

For the first simulation, which has a duration of 1600s, we set the maximum latency value as $l_{\max}=0.2$, and assume that latencies' variation at each switch is linked to provider p either as $\alpha_p(\tau_{r+1})=\alpha_p(\tau_r)+0.005$ or as $\alpha_p(\tau_{r+1})=\alpha_p(\tau_r)-0.05$, therefore $\bar{\alpha}=0.01$. The algorithm parameters are set as $\delta=0.01$ and $\varepsilon=0.01$, $\sigma=5$ and the resulting minimum dwell-time is $H^{(r)}=H_{\min}=148.9$s. The latencies variations occur every $H_{\min}$ during the first 1280s of the simulation.

Simulation results are shown in Fig. 2, which reports the system evolution starting from random initial load vectors. The top and center rows display the load vector dynamics and the values of the provider latencies, respectively, of the three commodities during the entire simulation.
The plots in the bottom row of Fig. 2 report the maximum latency mismatches among the δ-loaded providers of the three commodities, that is $e_\delta^i(t):=\max_{p\in\mathcal{P}^i|x_p^i(t)\geq\delta\lambda^i}l_p\big(x_p(t)\big)-\max_{m\in\mathcal{P}^i}l_m\big(x_m(t)\big)$.

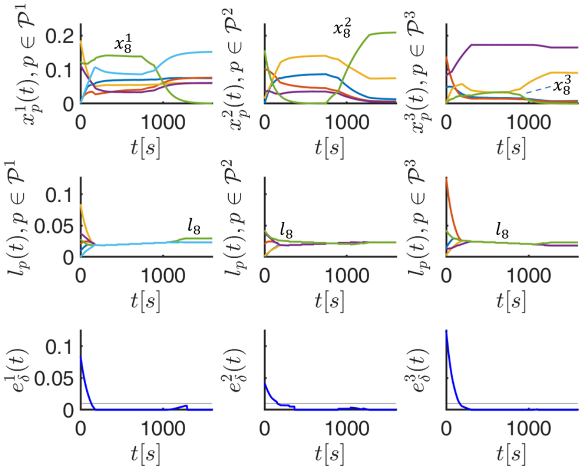

Fig. 2. Simulation 1: (top) load vectors; (center) provider latencies; (bottom) latency mismatches.

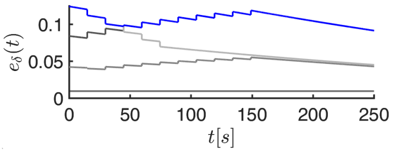

Fig. 3. Simulation 2: latency mismatches, $e_\delta^i(t)$, $i = 1, 2, 3$, and maximum latency mismatch $e_\delta(t) = \max_{i=1,2,3} e_\delta^i(t)$

The figure shows that e_δ^1, e_δ^2 and e_δ^3 fall below the threshold ε at times $t_1 = 131$s, $t_2 = 157$s and $t_3 = 165$s.

It is interesting to note how the policy of the commodities (upper plots) varies during the simulation in order to keep the error below ε (lower plots) in response to the variations of the latency functions (the switches are clearly identified by the steps in the latency values of the middle plots).

For all the commodities, the top row of plots shows that not all the provider latencies converge within a small neighbourhood; however, the providers of the latencies which do not converge are not δ-loaded. For instance, let us consider the provider 8, which is shared by all three commodities. Initially, the provider is used by commodities 1 and 2 but, as the simulation time grows, it becomes used by commodities 1 and 3 (e.g., at time 500s, $x_8^1 = 0.14$, $x_8^2 \approx 0$, $x_8^3 = 0.03$). Therefore, at time 500s, the latency value of provider 8, $l_8 = 0.023$, has not to converge to the latency values of the other providers used by commodity 2, with $l_p = 0.019$ for all $p \in \mathcal{P}^2 \setminus \{8\}$. At time 748s, the variations of the latency functions is such that x_8^2 starts growing and x_8^1 and x_8^3 start decreasing, and, at the end of the simulation, provider 8 becomes used by commodity 2 only. Then, the latency l_8 converges to the latency values of the providers used by commodity 2 and diverges from the latency values of the providers used by commodities 1 and 3.

Repeating the test in a setting in which the condition expressed by (26) is not violated, as is the case if we set $\bar{\alpha} = 0.02$, $\sigma = 1$ and $H^{(r)} = 15$s for all $r \in \mathcal{R}$. Fig. 3 shows that violating the constraint on the dwell-time, the controller is unable to recover the increase of the latency mismatches caused by the switches. In fact, the latency mismatch increases for all the switches for the commodities 1 and 3 (while it decreases for commodity 2), causing e_δ to diverge. The figure also shows that, as the switches terminate at time $t = 150$s, the latency mismatches of the commodities start annihilating.

V. Conclusion

This paper analysed the convergence of non-cooperative load balancing over providers characterized by time-varying latency functions. Leveraging on switched systems and set stability theory, it was possible to determine the minimum dwell-time under which the system state is guaranteed to converge in finite-time into a set of approximated Wardrop equilibria, in which the provider latencies are equalized with an arbitrarily small tolerance.

A possible future research direction involves the explicit inclusion of capacity constraints in the system dynamics.

References

[1] A. M. Alakeel *et al.*, "A guide to dynamic load balancing in distributed computer systems," *International journal of computer science and information security*, vol. 10, no. 6, pp. 153–160, 2010.

[2] M. R. Vuluvala and L. M. Saini, "Load balancing of electrical power distribution system: An overview," in *2018 International Conference on Power, Instrumentation, Control and Computing*. IEEE, 2018.

[3] D. Nace and M. Pioro, "Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 5–17, 2008.

[4] T. Roughgarden and E. Tardos, "How bad is selfish routing?" in *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc, 2000.

[5] R. M. Thrall, M. Beckmann, C. B. McGuire, and C. B. Winsten, "Studies in the economics of transportation," *Econometrica*, 1958.

[6] S. Fischer, H. Räcke, and B. Vöcking, "Fast convergence to wardrop equilibria by adaptive sampling methods," *SIAM Journal on Computing*, vol. 39, no. 8, pp. 3700–3735, 2010.

[7] A. Giuseppi and A. Pietrabissa, "Wardrop equilibrium in discrete-time selfish routing with time-varying bounded delays," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 526–537, 2021.

[8] F. Delli Priscoli, E. De Santis, A. Giuseppi, and A. Pietrabissa, "Capacity-constrained wardrop equilibria and application to multi-connectivity in 5g networks," *Journal of the Franklin Institute*, 2021.

[9] A. Giuseppi and A. Pietrabissa, "Stability and wardrop equilibria of non-cooperative routing with time-varying load," *IEEE Transactions on Automatic Control*, 2022.

[10] A. Pietrabissa and V. Suraci, "Wardrop equilibrium on time-varying graphs," *Automatica*, vol. 84, pp. 159–165, 2017.

[11] J. Anselmi, U. Ayesta, and A. Wierman, "Competition yields efficiency in load balancing games," *Performance Evaluation*, vol. 68, no. 11, pp. 986–1001, 2011.

[12] S. Fischer, L. Olbrich, and B. Vöcking, "Approximating wardrop equilibria with finitely many agents," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008, pp. 238–252.

[13] J. G. Wardrop, "Road paper. Some theoretical aspects of road traffic research." *Proceedings of the Institution of Civil Engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[14] M. Beckmann, C. B. McGuire, and C. B. Winsten, "Studies in the economics of transportation," *Econometrica*, vol. 26, no. 1, 1958.

[15] D. Liberzon and A. Morse, "Basic problems in stability and design of switched systems," *IEEE Control Systems Magazine*, vol. 19, no. 5, pp. 59–70, 1999.

[16] T. Alpcan and T. Basar, "A stability result for switched systems with multiple equilibria," *Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis*, vol. 17, no. 4, pp. 949–958, 2010.

[17] G. A. Shanholt, "Set stability for difference equations," *International Journal of Control*, vol. 19, no. 2, pp. 309–314, 1974.

[18] H. K. Khalil, *Nonlinear control*. Pearson New York, 2015, vol. 406.