Queueing analysis of imbalance between multiple server pools with an application to 3-phase EV charging

Andres Ferragut
Universidad ORT Uruguay
Montevideo, Uruguay
ferragut@ort.edu.uy

Fernando Paganini [*]
Universidad ORT Uruguay
Montevideo, Uruguay
paganini@ort.edu.uy

ABSTRACT

We consider systems where multiple servers operate in parallel, with a particular feature: servers are classified into d classes, and we wish to keep approximate balance between the load allocated to each class. We introduce a relevant imbalance metric, and study its behavior under stochastic demands with different task routing policies. For random routing, we analyze two cases of interest, depending on whether capacity constraints are operative: we obtain expressions for the stationary distribution and analyze the scaling behavior of our metric as a function of system size. Subsequently, we analyze active routing to the least loaded class, obtaining sharp bounds for the imbalance metric. As a practical application, we study the problem of imbalance between $d = 3$ phases, for the service of electrical vehicle charging. We show the engineering relevance of our imbalance metric in this context, and validate the theoretical results with simulations and real traces from EV charging data.

Keywords

Load balancing, Lyapunov methods, Electrical Vehicles.

1. INTRODUCTION

The balancing of load between servers has been a long concern of queueing systems, present in classical references [24, 2], and also with substantial recent activity in the context of cloud computing infrastructures [19]. In this setting, the usual object of study is a process of arriving *computing tasks* which must be processed by the overall system; individual servers maintain *queues* of tasks awaiting service, and the job of the load balancer is to route incoming jobs in a way that ensures stability and reduces the resulting latency. The baseline load-balancing strategy in this context is Join the Shortest Queue [9]. However, due to the scale of the problem, simpler variants have been studied to reduce the information burden while maintaining performance, such as Power-of-d choices [20, 25] or Join-the-Idle-Queue [19]. This has sparked a lot of recent advances on load balancing, particularly through fluid and mean field limits. An excellent summary on this line of work is [8].

Far less attention has been given to load balancing questions in queueing models for *parallel* server systems, typical of circuit-switched networks. In these models that go back to the Erlang studies for telephony, resources are *reserved* and queues model their occupation state; of this nature are the $M/G/\infty$ queue and the $M/G/C/C$ for finite capacity with blocking. A notable exception is [21], where the authors analyze a system with parallel server pools and blocking, related to the one described below. In their paper, however, they resort to scaling limits and fluid approximation, with the main objective being to minimize the overall blocking probability of the system. Another relevant reference in this regard is [11], where the authors analyze balancing between *processor-sharing* systems using join the shortest queue policy, and provide optimality and near-insensitivity results. In such a system, each server works on their allocated tasks in parallel with no queueing, but within each server its full capacity can be pooled and used to serve a single task, and therefore each server behaves essentially as a single server queue.

In this paper, we would like to address a different question: are there any other roles for *balancing* in these kinds of systems? We describe two application areas in which this kind of questions are natural. A first, also related to cloud computing, is when one looks at the problem at the higher layer of server reservations. Consider for instance the case of a distributed or cloud based microservices architecture (e.g. Kubernetes [22]) that opens container replicas in different pods. Each pod may run in a different physical node (the server pool), up to a maximum capacity. Besides reducing blocking, keeping a *balanced* number of active pods across nodes prevents individual node utilization becoming too high, which has a performance impact due to other jobs concurrently running at the node. It also minimizes the number of running tasks that are brought down when a random node fails.

Another example, which we will analyze in detail in Section 6, is the case of electrical vehicle charging in a parking facility. These installations are increasingly being deployed in response to the growing EV adoption, and cause strain in the electrical distribution network due to their relatively large power demand. In this regard, a crucial concern of network operators is three-phase *balance*: large installations are fed by three-phase AC lines, but individual chargers have a single-phase AC connection. An asymmetric consumption of current between the three phases leads to a number of

[*] This work was partially supported by ANII Uruguay under grant FCE_1_2021_1_167301 and AFOSR US under grant #FA9550-23-1-0350.

IFIP Performance '23. November 13–17 2023, Evanston, USA

undesirable effects for the grid [3], [29], and for this reason imbalance must be measured and controlled [14]. In many proposals for load management via online optimization [7, 4, 13, 27, 17], imbalance is included as a relevant constraint.

In this paper we then propose to analyze the imbalance that appears between multiple classes of servers, when subject to stochastic load and different models of routing. Each class or pool operates as a parallel server queue with no waiting, i.e. servers are reserved up to a maximum capacity. The general model Markov chain model is introduced in Section 2, together with an imbalance metric which measures the expected departure from the perfectly balanced state in a stationary regime.

In Section 3 we analyze the resulting Markov chain for the case of unconstrained capacities and uniform random routing, quantifying the baseline value of our imbalance metric; we also provide approximations in distribution for the large load case. In Section 4, finite capacities per pool are included, and we study *free spaces* routing, the natural model of random assignment for the case of asymmetric occupations. The resulting queue admits a reversible closed form solution: building upon this, we derive a bound for the relevant imbalance metric valid for general job size distributions, and over all traffic conditions and system sizes. This fact is particularly important in the EV charging application, where system size is not large and job sizes are rarely exponential.

In Section 5 we consider finally the policy where arriving jobs are assigned to the least loaded pool (i.e. the equivalent of JSQ in this context). In this case, we resort to the Foster-Lyapunov methods developed by [10, 26, 28]. For this routing policy, we derive a bound on imbalance that is *valid for any system size and traffic load.* Again this is of utmost importance when designing *robust* practical systems with a finite capacity.

Section 6 covers the application to EV charging: we briefly describe the problem of 3-phase imbalance in electrical networks, and how it is usually quantified in the industry. Under natural assumptions for the EV charging case, we show it maps directly to our measure of imbalance. Illustrative numerical experiments are given with real-world data for EV charging from [16]. Conclusions are given in Section 7.

2. SYSTEM MODEL

We begin by laying out precisely the model under consideration. Our system is in charge of serving tasks of a homogeneous type, which arrive as a Poisson process of rate λ. Tasks have a random duration T, which initially we assume exponential of rate μ, but some of our conclusions are more general. The total *traffic demand* on the system is thus $A := \lambda E[T] = \lambda/\mu$, which would correspond to the average number of active tasks in an $M/G/\infty$ queue.

Each task requires a dedicated server, and there is no queueing: if no servers are available tasks are rejected or fail to start. Although individual servers have homogeneous characteristics, they are divided into d *classes* or *pools*: for reasons explained in the introduction, we strive for *balance* between the level of activity of such pools at any given time. In particular, we will analyze the level of imbalance present under different models of assignment of tasks to pools, either spontaneous random assignment or active routing by a *balancer*. The system is depicted in Figure 1.

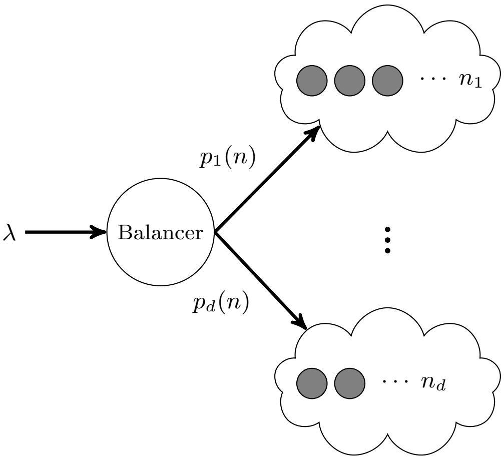


Figure 1: Balancing tasks across parallel server pools

Let X_i, $i = 1, \dots, d$ denote the number of tasks/active servers on pool i. Under the above assumptions, the state $X = (X_i)$ follows the continuous time Markov chain with $x \in \mathbb{N}^d$ and transition rates:

$$\begin{cases} x \mapsto x + e_i : & \lambda p_i(x), \\ x \mapsto x - e_i : & \mu x_i. \end{cases} \tag{1}$$

Here $p_i(x)$ is the probability that an incoming task is routed to pool i and e_i is the canonical vector. If there is a finite capacity C_i in pool i, we assume that $p_i(x) = 0$ whenever $x_i = C_i$.

For a given random state X, we will use $N := \sum_{i=1}^d X_i$ to denote the total task population across pools, and $\bar{X} := N/d$ the average occupation. Let us now define our imbalance measure. Perfectly balanced states are those with equal x_i, i.e. with x in the span of $\mathbf{1} = (1, \dots, 1)^T$. Consider the transformation:

$$Px := x - \bar{x}\mathbf{1} = \left(I - \frac{1}{d}\mathbf{1}\mathbf{1}^T\right)x, \tag{2}$$

projection of x onto the orthogonal complement of $\mathbf{1}$. The Euclidean norm $||Px||$ is then a suitable measure of the amount of system imbalance at state x, being null in the span of $\mathbf{1}$ as desired.

Considering the random process $X(t)$ in steady-state, we may use the expectations

$$J^1_{imb} := E[||PX||] \quad \text{or} \quad J^2_{imb} := E[||PX||^2] \tag{3}$$

as performance metrics for the process imbalance.

REMARK 2.1. *While we are concerned with balance between server pools and not fairness in the allocation, we would like to remark that our imbalance metric $||Px||$ is closely related to the classical Jain fairness index, by the following relationship:*

$$J(x) := \frac{< x, \mathbf{1} >^2}{d||x||^2} = 1 - \frac{||Px||^2}{||x||^2},$$

which can be deduced directly from (2). *Thus, a system with low $||Px||$ will have a higher fairness measure in Jain's sense.*

We can now formulate the main objective of the paper: to estimate the above imbalance metrics under different routing decisions and capacity constraints.

3. UNCONSTRAINED SERVER POOLS

As a baseline, we begin by analyzing the situation where pools are large enough so that they never fill up, which can serve as a model for cases when total utilization is well below capacity. Moreover, we assume here that tasks are routed uniformly at random between the d server pools ($p_i(x) \equiv 1/d$).

In this case, each pool behaves as an infinite server queue with total arrival rate λ/d, and the steady state distribution of the Markov chain (1) has the following product form:

$$\pi(x) = e^{-A} \prod_{i=1}^{d} \frac{(A/d)^{x_i}}{x_i!} \quad , x_i \in \mathbb{N}, i = 1, \dots, d, \tag{4}$$

i.e. d independent Poisson random variables with mean A/d.

Remark 3.1. *Since the $M/G/\infty$ queue is* insensitive *[6], the above steady-state occupation formula holds beyond the exponential assumption on the service time distribution.*

We compute the imbalance metric for this baseline case.

Proposition 3.2. *For the system with d unconstrained server pools and uniform random routing, we have:*

$$J_{imb}^2 = \frac{d-1}{d} A. \tag{5}$$

Proof. We first recall the following property of the Poisson distribution. If a random vector X is distributed as (4), then conditioned on the total number of tasks $N = n$, X has a multinomial distribution with total sample size n and probability $1/d$ on each of the d classes. In particular:

$$E[X_i \mid N = n] = \frac{n}{d},$$
$$E[(X_i - \bar{X})^2 \mid N = k] = k(1/d)(1 - 1/d) = k\frac{d-1}{d^2}.$$

Computing the imbalance conditioned on the total number of tasks we get:

$$\begin{aligned} E\left[||PX||^2 \mid N = n\right] &= E\left[||X - \bar{X}\mathbf{1}||^2 \mid N = n\right] \\ &= E\left[\sum_{i=1}^{d} (X_i - n/d)^2 \middle| N = n\right] \\ &= \sum_{i=1}^{d} n\frac{d-1}{d^2} = n\frac{d-1}{d}. \end{aligned}$$

Using now that the total number of tasks N has a Poisson distribution with mean A, we get:

$$E\left[||PX||^2\right] = E\left[E\left[||PX||^2 \mid N\right]\right] = \frac{d-1}{d}E[N] = \frac{d-1}{d}A.$$

□

While the above result concerns only the mean, an interesting approximation for the distribution arises in the case of $A \to \infty$.

Proposition 3.3. *For the system with d unconstrained server pools and random routing, as $A \to \infty$:*

$$\frac{d}{A}||PX||^2 \Longrightarrow^w ||(Z_1, \dots, Z_{d-1})||^2 \sim \chi^2_{d-1},$$

where Z_j are independent standard Gaussian random variables and χ^2_{d-1} is the Chi-square distribution with $d-1$ degrees of freedom. Here, $\Longrightarrow^w$ denotes convergence in law.

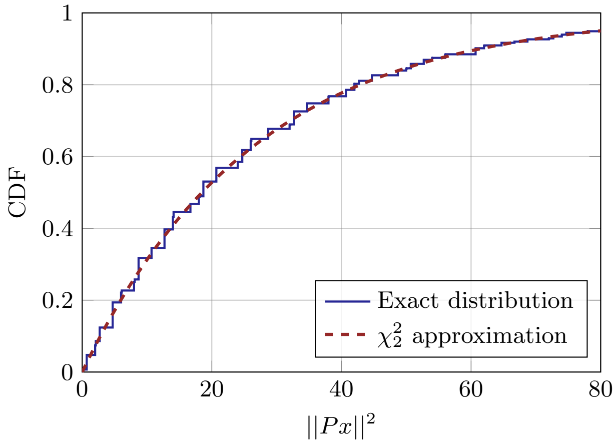


Figure 2: Cumulative distribution function of $||PX||^2$ for $d = 3$, $A = 40$ and the χ^2_2 approximation of Proposition 3.3.

Proof. The result stems from the Gaussian approximation to the Poisson distribution. If $X_i \sim Poisson(A/d)$ then:

$$\frac{X_i - A/d}{\sqrt{A/d}} \Longrightarrow^w \mathcal{N}(0, 1).$$

If $X = (X_1, \dots, X_d)$ are independent Poisson distributed random variables with parameter A/d, then:

$$PX = \sqrt{A/d} \cdot P\left[\frac{X - (A/d)\mathbf{1}}{\sqrt{A/d}}\right],$$

since $P\mathbf{1} = 0$. The term in brackets converges in distribution to a standard d-dimensional Gaussian random vector, of covariance the identity matrix. Multiplying by the orthogonal projection matrix P, satisfying $P = P^T = P^2$, yields:

$$\frac{1}{\sqrt{A/d}}PX \Longrightarrow^w \mathcal{N}(0, P).$$

This is a multivariate Gaussian supported in the $d-1$-dimensional span of P, so its norm-squared gives rise to a Chi-squared distribution with $d-1$ degrees of freedom. More explicitly: diagonalize the projection P by an orthogonal transformation U, i.e. $P = U^T DU$ with $D = \text{diag}(1, \dots, 1, 0)$. Then

$$\frac{1}{\sqrt{A/d}}UPX \Longrightarrow^w Z \sim \mathcal{N}(0, \Sigma)$$

with $\Sigma = UPU^T = D$ and thus $Z = (Z_1, \dots, Z_{d-1}, 0)$, where $Z_1, \dots, Z_{d-1}$ are independent standard Gaussians. Therefore

$$\frac{d}{A}||UPX||^2 \Longrightarrow^w ||(Z_1, \dots, Z_{d-1})||^2 \sim \chi^2_{d-1};$$

note finally that $||UPX|| = ||PX||$ since U is orthogonal. □

While the result is for $A \to \infty$, the approximation is valid for moderate values of A. As an example, we plot in Fig. 2 the case where $A = 30$ and $d = 3$, showing good fit.

Remark 3.4. *If we are interested instead in the first-moment imbalance criterion, observe that $\sqrt{\frac{d}{A}}||PX||$ follows (for large A) a chi-distribution χ_{d-1} (square root of* *chi-squared). From its known moments we obtain the approximation:*

$$J^1_{imb} := E\left[||PX||\right] \approx \sqrt{\frac{2A}{d}}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}.$$

Here $\Gamma(\cdot)$ is the Gamma function. In particular for $d = 3$, case of interest for the EV application, the right-hand side evaluates to $\sqrt{\frac{\pi A}{6}}$.

4. CONSTRAINED SERVER POOLS

We now turn our attention to the more realistic case where the server pools have finite capacity. Let C_i be the capacity of pool i, and $C = \sum_i C_i$ the total system capacity, i.e. the maximum number of simultaneous tasks that the system can handle. Consistently with our objective of *balance*, we will mostly focus on the case of homogeneous pool capacities, $C_i = C/d$. Nevertheless, our first result will be stated for general C_i.

When capacities are limited, routing uniformly at random is undesirable, since the balancer may choose a pool that is already working at full capacity. A more reasonable version of random routing is for the balancer to keep a list of available spaces at each of the server pools (i.e. inactive server instances), and then choose one at random out of this list for an arriving task. Appropriately, we call this the *free spaces routing* policy.

Under this assumption, the Markov chain (1) for the system occupation becomes:

$$\begin{cases} x \mapsto x + e_i : & \lambda\left[\dfrac{C_i - x_i}{C - n}\right] =: \lambda_i(x), \\ x \mapsto x - e_i : & \mu x_i =: \mu_i(x), \end{cases} \tag{6}$$

with $n := \sum_{i=1}^d x_i$. The arrival rates take into account that the probability $p_i(x)$ of choosing pool i is equal to the fraction of free spaces available that belong to that pool. Note that blocking will only occur if all pools are full.

The Markov chain in (6) is irreducible and has a finite state space, and thus it always has a steady-state distribution. Interestingly, this chain belongs to the class of *balanced routing and allocations* studied in [6]. The arrival and departure rates are said to be *balanced* (unfortunately, this is a different meaning to our use elsewhere in the paper) if there exists functions defined in the state space $\Lambda(x)$ and $\Phi(x)$ such that:

$$\Lambda(x + e_i) = \lambda_i(x)\Lambda(x), \tag{7a}$$

$$\Phi(x - e_i) = \mu_i(x)\Phi(x). \tag{7b}$$

By [6, Eq. (10)] the steady-state distribution of a Markov chain in this class admits a product form given by:

$$\pi(x) = \pi(0)\Lambda(x)\Phi(x). \tag{8}$$

We can now state the main result of this Section:

THEOREM 4.1. *The equilibrium occupancy distribution of the Markov chain* (6) *for the free spaces routing policy is given by:*

$$\pi(x) = \pi(0)\frac{A^n}{n!}\frac{\prod_{i=1}^d \binom{C_i}{x_i}}{\binom{C}{n}}, \tag{9}$$

for $0 \leq x_i \leq C_i$, $i = 1, \ldots, d$ and with

$$\pi(0) = \frac{1}{\sum_{j=0}^C A^j/j!}. \tag{10}$$

The result can be interpreted as follows: the total number of tasks N follows the steady-state distribution of an $M/M/C/C$ (Erlang) queue, with $C = \sum_{i=1}^d C_i$. This is expected since all arrivals are admitted until the system is completely full. Given a value $N = n$ for the total occupation, in steady-state all possible subsets of size n become equiprobable; the binomial coefficients compute how many ways there are of assigning x_i tasks (balls) to C_i servers (bins) available in each pool; this amounts to a multivariate Hypergeometric distribution for fixed n.

PROOF. It is trivial to show that the departure rates in (6) satisfy (7b) with

$$\Phi(x) := \frac{1}{\mu^n \prod_{i=1}^d x_i!}.$$

Also, defining:

$$\Lambda(x) := \lambda^n \frac{(C-n)!}{C!}\prod_{i=1}^d \frac{C_i!}{(C_i - x_i)!},$$

we can verify (7a) from the arrival transitions in (6). In fact, this same routing was applied in [15] to analyze a many-server cloud load balancing system with queueing.

We combine the above with (8) to yield:

$$\begin{aligned} \Lambda(x)\Phi(x) &= \lambda^n \frac{(C-n)!}{C!}\left[\prod_{i=1}^d \frac{C_i!}{(C_i - x_i)!}\right]\frac{1}{\mu^n \prod_{i=1}^d x_i!} \\ &= A^n \frac{(C-n)!}{C!}\prod_{i=1}^d \binom{C_i}{x_i} \\ &= \frac{A^n}{n!}\frac{\prod_{i=1}^d \binom{C_i}{x_i}}{\binom{C}{n}}. \end{aligned}$$

To obtain the normalization factor, we rely on the following identity, which stems from the multivariate Hypergeometric distribution:

$$\sum_{x_1+\ldots+x_d=n} \frac{\prod_{i=1}^d \binom{C_i}{x_i}}{\binom{C}{n}} = 1.$$

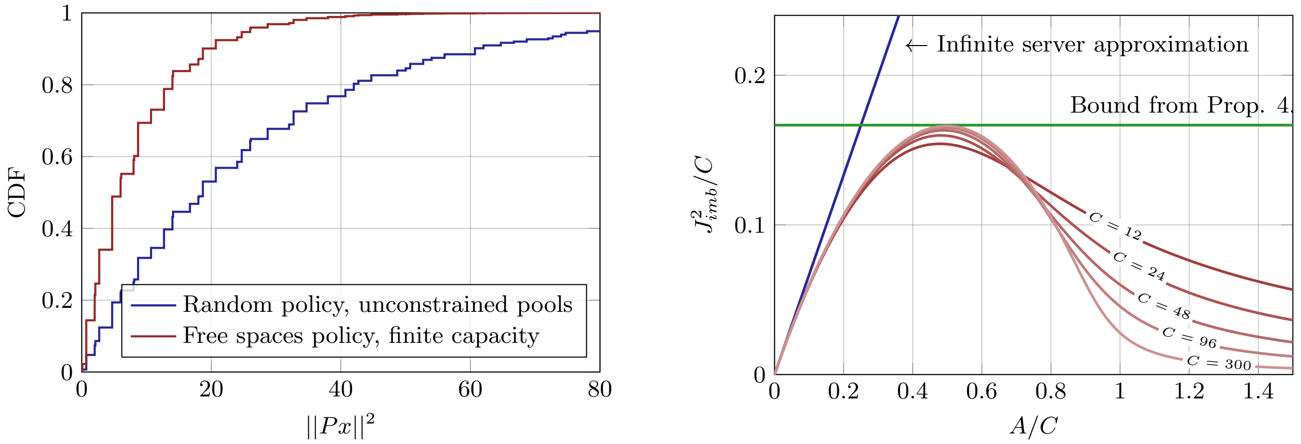
Summing now over $n = 0, \ldots, C$ we conclude the proof. □

An important property of balanced allocations is that the underlying Markov chain process is *reversible* [6]. This in turn implies that the system is *insensitive* to the exact sojourn time distribution, provided that arrivals are Poisson. Therefore, we have the following:

REMARK 4.2. *The exponential hypothesis on the sojourn times in Theorem 4.1 can be relaxed and the steady state distribution* (9) holds for general sojourn times.

We now return to the case when pools are homogeneous, i.e. $C_i = C/d$, on which we focus on the sequel. First note that if we let $C \to \infty$ while keeping $C_i/C = 1/d$, the multivariate Hypergeometric distribution converges to a Multinomial distribution. Specifically, for fixed $x_i \leq C_i$ and $n = \sum_i x_i$ we have the limit

$$\frac{\prod_{i=1}^d \binom{C_i}{x_i}}{\binom{C}{n}} \xrightarrow[C\to\infty]{} \binom{n}{x_1 \ldots, x_d}\frac{1}{d^n}.$$



(a) Cumulative distribution function of $||PX||^2$ for $d = 3$, $A = 40$ and the unconstrained random policy case for comparison.

(b) Scaled behavior of the imbalance measure J^2_{imb} showing the effect of the finite capacity in the imbalance.

Figure 3: Free-spaces routing in steady-state for $d = 3$.

Also, from (10) we have $\pi(0) \to e^{-A}$ as $C \to \infty$. Substitution into (9) implies the limit:

$$\pi(x) \xrightarrow[C\to\infty]{} e^{-A}\frac{A^n}{n!}\frac{n!}{\prod_{i=1}^d x_i!}\frac{1}{d^n} = e^{-A}\prod_{i=1}^{d}\frac{(A/d)^{x_i}}{x_i!},$$

pointwise for fixed x. Now, as $C \to \infty$ with $C_i = C/d$, the region of validity of this limit covers the entire positive orthant $\mathbb{N}^d$, and we have recovered the distribution in (4). Namely, if capacity is large and evenly distributed, the free spaces routing policy will behave exactly as the uniform random policy, as expected.

However, when capacity is not large, an interesting feedback effect occurs. As an example, in Figure 3a we plot the exact CDF of the imbalance measure $||PX||^2$ computed numerically averaging over the distribution in eq. (9), and compare it with the unconstrained server case of eq. (4). We can observe a *smaller* value of imbalance in the constrained case: this is because although routing is random, the fact that a pool is already loaded reduces the probability of a new arrival being routed to it, and thus the distribution *naturally* concentrates more along the diagonal. We refer to Figure 4a for a scatter plot of this distribution.

In practice, imbalance will depend on total capacity C and offered traffic A. For $A \ll C$ it will behave as in the random policy with infinite capacity. For $A \gg C$, the system will operate completely full, and thus naturally balanced albeit with a very high blocking probability. The more interesting cases are the intermediate values of A. We now obtain an explicit bound on the worst-case imbalance one may encounter across all regimes.

Proposition 4.3. *For the free spaces routing policy* (6) *with homogeneous pool capacity $C_i = C/d$, in steady state we have:*

$$J^2_{imb} \leqslant \frac{d-1}{4d}\frac{C^2}{C-1}. \qquad (11)$$

Proof. The key point is again to observe that, given the total occupation N, X follows a multivariate Hypergeometric distribution, whose moments are known. In particular, from (2) we have:

$$E[\|PX\|^2|N] = \sum_{i=1}^{d} E\left[\left(X_i - \frac{N}{d}\right)^2 \middle| N\right].$$

Conditioned on N, the vector X follows a multivariate Hypergeometric with parameters $C_i = C/d$ and N. This implies $E[X_i \mid N] = N/d$ and thus the terms inside the last sum are just the variances of the Hypergeometric components. Substituting the corresponding formulas for the variances we have:

$$\begin{aligned} E[\|PX\|^2|N] &= \sum_{i=1}^{d} N\frac{C_i}{C}\left(1-\frac{C_i}{C}\right)\frac{C-N}{C-1} \\ &= \sum_{i=1}^{d}\frac{1}{d}\frac{d-1}{d}\frac{N(C-N)}{C-1} \\ &= \frac{d-1}{d}\frac{N(C-N)}{C-1}. \end{aligned}$$

Using that $N \leqslant C$, and thus $N(C-N) \leqslant C^2/4$ we obtain:

$$\begin{aligned} E[\|PX\|^2] = E\left[E[\|PX\|^2 \mid N]\right] &= E\left[\frac{d-1}{d}\frac{N(C-N)}{C-1}\right] \\ &\leqslant \frac{d-1}{d}\frac{1}{C-1}\frac{C^2}{4} = \frac{d-1}{4d}\frac{C^2}{C-1}. \end{aligned}$$

□

In fact, by computing the exact expressions for the moments of N involved using the equilibrium distribution, it can be shown that the above bound is tight for large C, and that the worst case is attained when the offered traffic $A = \lambda/\mu \approx C/2$. We omit the computations.

Instead, we illustrate the behavior through an example in Figure 3b, for $d = 3$. The x-axis is the offered traffic A, normalized by C; the y-axis is our imbalance metric, normalized also by C. Different system sizes are explored. For small A relative to C, the imbalance metric J^2_{imb} follows the linear increase in A consistent with the formula (5) for the unconstrained case. Imbalance peaks around $C/2$ as mentioned above, and it is very close to the bound provided in the previous proposition. Then it decreases, approaching zero in the limit of large A/C, in which the pools saturate and therefore tend to operate in balance.

As a conclusion of this Section, the *free spaces* random routing policy will have less imbalance than random uniform routing in the unconstrained case, however it will still increase with system size.

5. ROUTE TO THE LEAST LOADED POOL

To further reduce the imbalance of load among pools, a more proactive routing policy is required. A natural choice would be an analog of the classical Join the Shortest Queue (JSQ) policy from traditional load balancing. In particular, the balancer may route each incoming task to the pool that has fewer tasks running on it, i.e. the least loaded one, with ties broken at random. Accordingly, we call our policy Least-Loaded-Pool (LLP).

For a precise description we return to (1), where arrival transitions $x \mapsto x + e_i$ had intensity $\lambda p_i(x)$, and specify the routing as follows: for a vector $x \in \mathbb{N}^d$, let $k(x)$ denote the number of components that achieve its minimum, i.e. $k(x) := \#\{\arg\min(x_i)\}$, and take

$$p_i(x) = \begin{cases} \frac{1}{k(x)}: & \text{if } \ i \in \arg\min(x_i), \quad \min(x_i) < C/d; \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In other words, the arrival rate λ is split equally among the pools which have a minimum occupation. Blocking is included in the above model, for the case of finite capacity; it occurs only at the state $x = \frac{C}{d}\mathbf{1}$, i.e. when the entire system is full.

The following properties of $p(x)$ hold in any non-blocking state:

$$\sum_i p_i(x) = 1, \qquad \sum_i x_i p_i(x) = \min(x_i). \tag{13}$$

Departures $x \mapsto x - e_i$ have rate μx_i as before. The Markov chain in (1) with the routing in (12) cannot, unfortunately, be solved explicitly as in previous sections. Note that the total occupation N still behaves as an $M/M/\infty$ queue for the unconstrained case, or as an Erlang queue for $C < \infty$, but the full distribution does not admit a closed form solution. Nevertheless, we can bound the amount of imbalance in the routing system applying Foster-Lyapunov methods to the Markov chain. In particular, we need the following moment bound from [12]:

LEMMA 5.1 (CF. [12] CHAPTER 6). *Let $X(t)$ be a continuous time Markov chain in the state space $\mathcal{X}$ and V, f, and g are nonnegative functions on $\mathcal{X}$. Assume that:*

$$QV(x) \leqslant -f(x) + g(x) \quad \forall x \in \mathcal{X}$$

where QV is the drift of the Lyapunov function V over the chain dynamics. If $X(t)$ is positive recurrent, so that the means are well-defined, then in steady state $E[f(X)] \leqslant E[g(X)]$.

We also need the following bound:

LEMMA 5.2. *For $x \in \mathbb{R}^d$,*

$$\|Px\| \leqslant \sqrt{d^2 - d}\left(\bar{x} - \min_i x_i\right). \tag{14}$$

PROOF. First note that both sides of (14) are invariant if we shift the coordinates by $c \in \mathbb{R}$, i.e. by adding a vector $c\mathbf{1}$ to x. Taking $c = -\min_i x_i$ we can assume without loss of generality that $\min_i x_i = 0$. Also, the inequality is scale invariant, so we can also assume that x is normalized such that $\bar{x} = 1$, or equivalently $\mathbf{1}^T x = d$. We must therefore prove the bound $\|Px\| \leqslant \sqrt{d^2 - d}$ for such vectors. We first observe that

$$\max_{\{\mathbf{1}^T x = d, \ \min(x_i) = 0\}} \|Px\| \leqslant \max_{\{\mathbf{1}^T x = d, \ x_i \geq 0\}} \|Px\| \tag{15}$$

and the right-hand side problem involves the *maximization* of a convex function over a convex set, its optimum must be an extreme point. For the simplex in the right-hand side the extreme points are of the form de_i, e_i being the coordinate vectors. Evaluating the function at these points we have:

$$\|Pde_i\| = \|de_i - \mathbf{1}\| = \sqrt{d^2 - d}.$$

Thus, the right-hand side of (15) evaluates to $\sqrt{d^2 - d}$ and the result follows. □

We are now ready to prove the main result of this Section:

THEOREM 5.3. *Consider the Markov chain* (1) *under the LLP policy described by* (12) *with d server pools, either with infinite capacity of with a total capacity C equally distributed among them. Then for any value of λ, μ, in steady state:*

$$J^1_{imb} = E[||PX||] \leqslant (d-1)\sqrt{1 - \frac{1}{d}}.$$

PROOF. We first establish positive recurrence of the chain for the case[1] $C = \infty$, introducing the Lyapunov function $V_1(x) = \bar{x}$. Computing its drift we obtain:

$$QV_1(x) = \sum_i \lambda p_i(x)\frac{1}{d} - \sum_i \mu x_i \frac{1}{d} = \frac{1}{d}\lambda - \mu\bar{x},$$

where we have invoked (13). In particular, $QV_1(x) \leqslant -\varepsilon < 0$ outside a compact set, so the Foster criterion implies the positive recurrence of the chain.

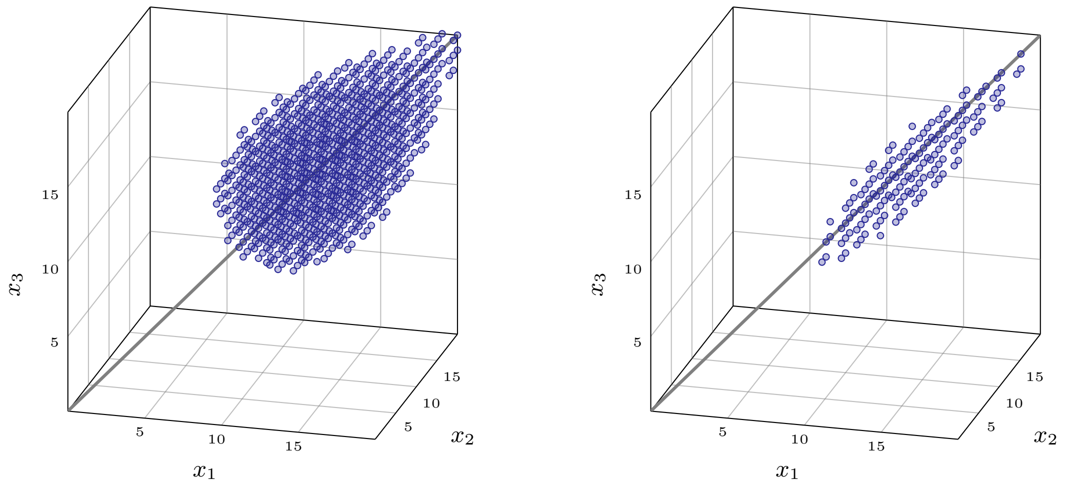
We now introduce a second Lyapunov function, $V_2(x) = ||Px||^2$. To compute the drift of V_2, we first write the identity:

$$\begin{aligned} ||P(x \pm e_i)||^2 &= (x \pm e_i)^T P^T P (x \pm e_i) \\ &= (x \pm e_i)^T P (x \pm e_i) \\ &= ||Px||^2 \pm 2e_i^T Px + e_i^T Pe_i \\ &= |Px||^2 \pm 2(x_i - \bar{x}) + \frac{d-1}{d}; \end{aligned} \tag{16}$$

the last step uses the expression (2) for the projection.

We can now compute the drift of $V_2(x)$ at a non-blocking

[1] The case $C < \infty$ with a finite state-space is also positive recurrent.



(a) Free spaces routing.

(b) Least loaded pool routing.

Figure 4: Scatter plot representing the 90% most probable points from the steady-state distribution, showing the state space collapse of the LLP policy, $A = 40, C = 60$.

state (again, invoking (13)):

$$\begin{aligned} QV_2(x) &= \sum_i \lambda p_i(x) 2(x_i - \bar{x}) - \sum_i \mu x_i 2(x_i - \bar{x}) \\ &\quad + \frac{d-1}{d}\left[\sum_i \lambda p_i(x) + \mu x_i\right] \\ &= 2\lambda\left[\min_i x_i - \bar{x}\right] - 2\mu\left[\sum_i x_i^2 - d\bar{x}^2\right] \\ &\quad + \frac{d-1}{d}\lambda + (d-1)\mu\bar{x}. \end{aligned}$$

Define now the Lyapunov function $V(x) = (d-1)V_1(x) + V_2(x)$. Using the above expressions, the total drift of V at a non-blocking state is:

$$\begin{aligned} QV(x) &= 2\lambda\left[\min_i x_i - \bar{x}\right] - 2\mu\left[\sum_i x_i^2 - d\bar{x}^2\right] + 2\frac{d-1}{d}\lambda \\ &\leqslant 2\lambda\left(\min_i x_i - \bar{x}\right) + 2\lambda\frac{d-1}{d}, \end{aligned} \tag{17}$$

where the term multiplying 2μ is non-negative from a direct application of Cauchy-Schwarz inequality.

We note that the drift bound (17) also holds at the blocking state $x = \frac{C}{d}\mathbf{1}$ (when $C < \infty$). At this point $V_1(x) = \frac{C}{d}$, $V_2(x) = 0$, and only downward transitions are allowed, with:

$$\begin{aligned} V(x - e_i) &= (d-1)\overline{x - e_i} + \|P(x - e_i)\|^2 \\ &= (d-1)\frac{C-1}{d} + \frac{d-1}{d} = (d-1)\frac{C}{d} = V(x); \end{aligned}$$

note the first two terms in (16) vanish at this state. Thus, the drift $QV(x)$ is zero; and so is the first term on the right of (17), satisfying the bound.

We are now ready to apply Lemma 5.1 by choosing $f(x) = 2\lambda\left(\bar{x} - \min_i x_i\right)$ and $g(x) \equiv 2\lambda\frac{d-1}{d}$, to conclude that in steady-state:

$$E\left[\bar{X} - \min_i X_i\right] \leqslant \frac{d-1}{d}. \tag{18}$$

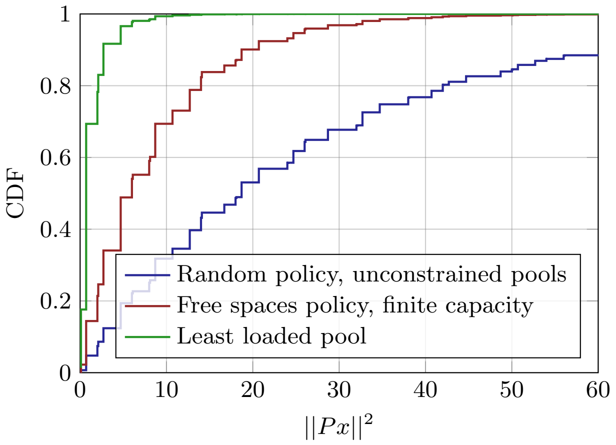


Figure 5: Cumulative distribution function of $||PX||^2$ for $d = 3$, $A = 40$ under the Least Loaded Pool policy and previous policies for comparison.

Combining this last inequality with the bound (14) from Lemma 5.2, we arrive at:

$$\begin{aligned} E[\|PX\|] &\leqslant \left(\sqrt{d^2 - d}\right) E\left[\bar{X} - \min_i X_i\right] \\ &\leqslant \left(\sqrt{d^2 - d}\right)\frac{d-1}{d} = (d-1)\sqrt{1 - \frac{1}{d}}, \end{aligned}$$

which concludes the proof. □

A remarkable fact is that the bound of Theorem 5.3 only depends on the number of pools d. As a consequence, the imbalance metric J^1_{imb} remains *uniformly bounded* for *any system load and any system size*. This can be interpreted in terms of the state-space collapse results of [10]: routing to the least loaded pool collapses the state space towards the diagonal (perfect balance), and large excursions are not allowed. An example of this collapse is shown in Figure 4.

We also depict in Figure 5 the CDF of the imbalance measure J^2_{imb} for this example and compare it to the previous policies, showing that for small d (as in the EV case) the imbalance is greatly reduced.

REMARK 5.4. *Let us briefly discuss the tightness of the bound in Theorem 5.3: the key part of the proof is eq.* (18)*: this first bound shows that, in steady state, the average pool occupation is at most* 1 *off the minimal occupation in expectation. This part of the bound is quite tight and suggests that the imbalance should not grow with dimension. But in order to bound the relevant imbalance metric* $||PX||$*, we resort to Lemma 5.2, which is a worst case bound that grows with d. There could be room for improvement in this second step.*

As a side note, remark that another relevant measure for imbalance in the system may be the *maximum pool occupation*, i.e. $\max_i x_i = ||x||_\infty$. From Theorem 5.3 we can directly prove:

COROLLARY 5.5. *In the conditions of Theorem 5.3, we also have:*

$$E[\max_i X_i - \bar{X}] \leqslant J_1^{imb},$$

and thus the bound of Theorem 5.3 also holds for the spread between the maximum pool occupation and the average occupation.

PROOF. The proof follows from the triangular inequality:

$$\begin{aligned}\max_i x_i &= ||x||_\infty = ||\bar{x}\mathbf{1} + Px||_\infty \leqslant ||\bar{x}\mathbf{1}||_\infty + ||Px||_\infty \\ &= \bar{x} + ||Px||_\infty \leqslant \bar{x} + ||Px||.\end{aligned}$$

In the last step we used that $||\mathbf{1}||_\infty = 1$ and that $||Px||_\infty \leqslant ||Px||$. Taking expectations on both sides we conclude the proof. □

REMARK 5.6. *Finally, we remark that the least-loaded-pool policy may not be easy to implement in large scale systems, where the number of pools and capacities are high. In this case, the load balancer may not be able to keep track pool usage in real time without an expensive exchange of information and messages. Thus, approximate policies based on sampling, such as power-of-d choices [20], become an attractive alternative. Pursuing the study of such policies in the large scale limit is outside our scope in this paper, since we focus on steady state finite size results, but represents an interesting line of future work.*

6. APPLICATION TO THREE-PHASE ELECTRICAL VEHICLE CHARGING

We now apply our results to a relevant problem for electrical vehicle charging: three-phase imbalance. First we analyze how an EV parking lot can be modeled as a parallel server system with finite capacity and $d = 3$ pools, namely the AC phases. We then show how to cast the relevant measure of imbalance to our projection matrix P, and how the analyzed routing policies map naturally to this setting. We then provide analytical results for the electrical imbalance that can be expected in the system in the stochastic load setting, and also illustrate real-world behavior with traces from the Caltech Adaptive Charging Network [18].

6.1 Parking lot model

Consider a parking lot facility providing EV charging capabilities. This will be provided by multiple *Electrical Vehicle Supply Equipments (EVSEs)* connected to the network, as in Figure 6a. The grid provides three-phase alternating current (AC), but EVSEs are single-phase, meaning that they draw their power from two of the three AC lines (thus, there are three classes or phases for the EVSEs). A typical Level 2 EVSE consumes 7.2kW, and they are among the largest single-phase loads connected to the grid [23].

Assume that vehicles requiring charge arrive into the system as a Poisson process of rate λ and, provided there are chargers available, remain in the system for an exponential time of rate μ. The parking lot has a total of C homogeneous chargers, with C_i of them connected to phase i, $i = 1, 2, 3$. By design typically $C_i = C/3$ in order to keep the phases balanced. Vehicles present in the system are charged simultaneously, and the total power drawn from each phase depends on its current occupation.

Under the above assumptions, the EV charging system behaves as the general model from eq. (1) and Figure 1, with $d = 3$ pools and total capacity C equally split among the pools. Now $X_i(t)$ is the number of vehicles charging in phase i at time t, and thus the total phase current will be

$$\rho_i(t) := I_0 X_i(t), \quad i = 1, 2, 3, \tag{19}$$

where I_0 is the current of a single EVSE.[2]

An arriving vehicle may choose to park at any of the free available spaces at random, without knowledge of the phase it is connected to. Therefore, traffic will split naturally among the phases according to the *free spaces* routing policy analyzed before. If, instead, the station operator proactively routes vehicles to the least loaded phase, the system behaves according to the LLP policy. In what follows, we analyze electrical imbalance in both systems and show that important gains can be obtained by applying a more proactive routing.

6.2 Modeling three-phase imbalance

It is beyond our scope to provide a self-contained background on three-phase alternating current (AC) circuits (see, e.g. [5]), but we briefly define our terminology: the sinusoidal current in each AC line is represented by a *phasor*, i.e. a complex number representing its amplitude and phase. In this setting, an ideal, *balanced three-phase* current system (I_a, I_b, I_c) must have equal amplitude and 120^o relative phase, that is:

$$I_a = \alpha I_b = \alpha^2 I_c, \tag{20}$$

where we have introduced the notation $\alpha := e^{j2\pi/3}$, i.e. a 120^o rotation; j is the imaginary unit. Some useful identities for α are:

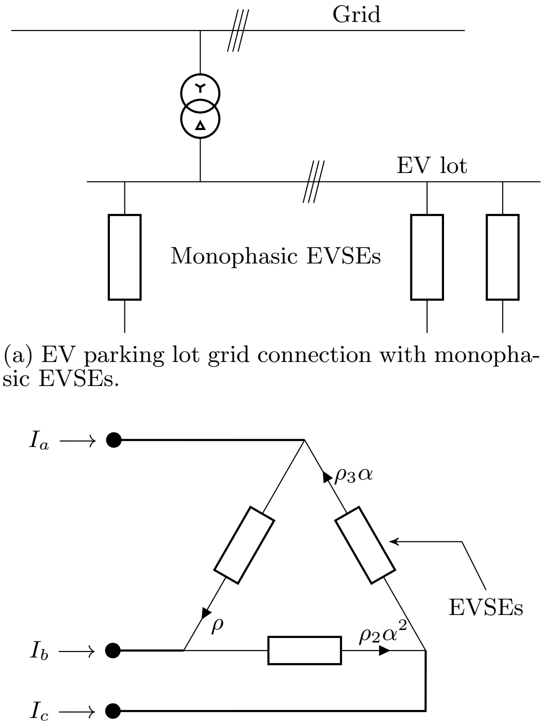
$$\alpha^3 = 1, \quad \overline{\alpha} = \alpha^2, \quad \alpha + \alpha^2 = -1.$$

Moreover, an orientation choice, named *positive sequence* has been made in definition (20). An asymmetric consumption will deviate the system from perfect balance, either by differences in magnitude or relative angle. However, as shown in [1], for EV loads differences in magnitude prevail, so we can make the following:

ASSUMPTION 6.1. *The currents* I_i *drawn by each EVSE pool satisfy:*

$$I_1 = \rho_1; \quad I_2 = \rho_2\alpha^2; \quad I_3 = \rho_3\alpha, \tag{21}$$

[2] This is equivalent to having a constant power consumption at the given EVSE voltage. A typical value for I_0 is 30A, corresponding to 7.2kW at 240V.



(a) EV parking lot grid connection with monophasic EVSEs.

(b) Δ configuration of chargers and current phasors involved.

Figure 6: Network diagram of a typical EV charging facility

where the $\rho_i = \rho_i(t)$ depend on the system occupation as in (19).

A final consideration is that, for a given installation, engineers have a choice on how to connect the phases: the Δ or Y (wye) configuration. We focus on the Δ case depicted in Figure 6b, common in practice. From eq. (19) and the Kirchoff laws we have that:

$$\begin{pmatrix} I_a \\ I_b \\ I_c \end{pmatrix} = \begin{pmatrix} I_1 - I_3 \\ I_2 - I_1 \\ I_3 - I_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -\alpha \\ -1 & \alpha^2 & 0 \\ 0 & -\alpha^2 & \alpha \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix}. \qquad (22)$$

The standard way to measure imbalance in such a system is called *symmetrical components* analysis (see e.g. [5], Ch 12.): this amounts to a change of coordinates from the original phasors defined by the *Fortescue transformation*:

$$\begin{pmatrix} I^0 \\ I^+ \\ I^- \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{pmatrix} \begin{pmatrix} I_a \\ I_b \\ I_c \end{pmatrix}. \qquad (23)$$

In fact, eq. (23) amounts to a three-point Discrete Fourier Transform of the original phasor sequence (I_a, I_b, I_c). The symmetrical components are then the new phasors I^+ (*positive sequence*), I^- (*negative sequence*), and I^0 (*zero sequence*). In a perfectly balanced system satisfying (20), the positive sequence is exactly $I^+ = I_a = \alpha I_b = \alpha^2 I_c$ and both $I^- = I_0 = 0$. Therefore, the *magnitude* of the negative and null sequence phasors serve as a measure of imbalance.

Under Assumption 6.1 we can combine (22) and (23) to get:

$$\begin{aligned} \begin{pmatrix} I^0 \\ I^+ \\ I^- \end{pmatrix} &= \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{pmatrix} \begin{pmatrix} 1 & 0 & -\alpha \\ -1 & \alpha^2 & 0 \\ 0 & -\alpha^2 & \alpha \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 0 & 0 & 0 \\ 1-\alpha & 1-\alpha & 1-\alpha \\ 1-\alpha^2 & \alpha - 1 & -\alpha + \alpha^2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix}. \qquad (24) \end{aligned}$$

The zero-sequence component disappears, this is a consequence of the Δ configuration which forces the sum of the sum of the line currents to be 0. Define the vector of magnitudes $\rho := (\rho_1, \rho_2, \rho_3)$. Then the positive sequence magnitude is

$$|I^+| = \frac{|1-\alpha|}{3} \mathbf{1}^T \rho = \frac{1}{\sqrt{3}} \mathbf{1}^T \rho, \qquad (25)$$

and the negative sequence magnitude satisfies:

$$I^- = \frac{1-\alpha^2}{3} \begin{pmatrix} 1 & \alpha & \alpha^2 \end{pmatrix} \rho,$$

where we have used the identities for α. Computing its magnitude squared and using that $\rho_i \in \mathbb{R}$ we get:

$$|I^-|^2 = \frac{1}{3}(\rho_1^2 + \rho_2^2 + \rho_3^2 - \rho_1\rho_2 - \rho_2\rho_3 - \rho_3\rho_1).$$

The above quadratic form can be readily expressed as:

$$|I^-|^2 = \frac{1}{2} \rho^T P \rho = \frac{1}{2} \|P\rho\|^2, \qquad (26)$$

where P is the matrix:

$$P = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}. \qquad (27)$$

Now, P is exactly the projection onto the orthogonal complement of $\mathbf{1} = (1,1,1)^T$ that we introduced in (2), for the case $d = 3$. In particular substituting (19), in steady state we should have:

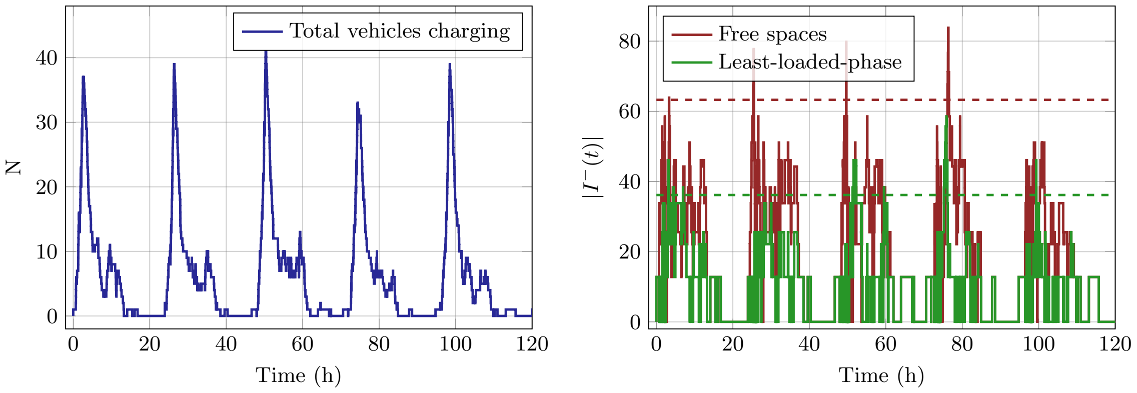
$$E\left[|I^-|\right] = \frac{I_0}{\sqrt{2}} E\left[\|PX\|\right] = \frac{I_0}{\sqrt{2}} J^1_{imb}. \qquad (28)$$

To summarize: in a Δ configuration, electrical imbalance is measured by the magnitude $|I^-|$ of the negative sequence component of the line current phasors (I_a, I_b, I_c); this is in fact defined in [14] as the industry standard. For an EV parking lot, where Assumption (6.1) holds, this metric is proportional to the norm of a projection under P in (2) of the vector of EVSE occupations per phase. Thus, the expected imbalance under stochastic load for the parking lot is expressed by (28), enabling us to apply the previous analysis for $d = 3$, taking as our server pools the EVSEs in each phase.

6.3 Numerical experiments

We now apply our previous results to bound the steady state electrical imbalance. Consider an EV parking lot with a finite number C of EVSEs evenly distributed among the phases and let $X_i(t)$, $i = 1,2,3$ be the number of active users in phase i, and $N = X_1 + X_2 + X_3$ the total occupation.

If users choose the parking space at random, then the probability of choosing a free spot from phase i is exactly $(C_i - X_i)/(C - N)$, i.e. the free spaces routing policy analyzed earlier. We can thus prove:



(a) Parking occupation.

(b) Imbalance metrics and estimated bounds.

Figure 7: Total number of charging EVs and imbalance evolution for the discussed policies in the JPL system.

Proposition 6.2. *For a parking lot with the random free spaces routing policy, in steady state we have:*

$$E\left[|I^-|\right] = \frac{I_0}{\sqrt{2}} J^1_{imb} \leqslant \frac{I_0}{\sqrt{2}}\sqrt{J^2_{imb}} \leqslant \frac{I_0}{2\sqrt{3}}\frac{C}{\sqrt{C-1}},$$

for any traffic load A.

The proof follows from expression (28), Jensen's inequality to bound $J^1_{imb} \leqslant \sqrt{J^2_{imb}}$ and the bound for J^2_{imb} in Proposition 4.3. From this result, the relative amount of imbalance satisfies:

$$\frac{1}{I_0 C}E[|I^-|] \leqslant \frac{1}{2\sqrt{3}}\frac{1}{\sqrt{C-1}} \underset{C\to\infty}{\sim} O\left(\frac{1}{\sqrt{C}}\right).$$

Not surprisingly, relative imbalance vanishes with system size, but decay is slow, which means smaller installations may suffer from relatively high imbalance currents, tripping circuit protections. In that case, minimizing imbalance through vehicle routing can be a useful and simple to implement approach. From our analysis in Section 5, the LLP policy will achieve better results. Explicitly, applying the bound in Theorem 5.3 we have:

Proposition 6.3. *For a parking lot actively routing vehicles to the least loaded phase, in steady state we have:*

$$E\left[|I^-|\right] = \frac{I_0}{\sqrt{2}}J^1_{imb} \leqslant \frac{2}{\sqrt{3}}I_0$$

for any traffic load A and system size C.

In this case, the relative imbalance decays much faster:

$$\frac{1}{I_0 C}E[|I^-|] \leqslant \frac{2}{\sqrt{3}C},$$

and thus further reducing the strain on the installation.

The preceding analysis assumes a stationary demand, however in practice, real world installations will handle a time varying traffic intensity due to natural daily usage cycles. Since our imbalance estimates are valid *for any traffic load*, provided the arrival rates change slowly in time, the estimates should approximately hold for time-varying scenarios. Therefore, we can compare them to the imbalance found in real world traces. To do so, we resort to measured traffic traces from Caltech Adaptive Charging Network installations, publicly available in [18].

The trace under consideration comes from an EV parking lot at NASA Jet Propulsion Laboratory in Pasdena, California. We took a typical work week (Mon-Fri) to capture the daily cycles, with a total of 385 charging sessions (77 per day), each having an average demand of 15kWh per EV, which amounts to 2.1h charging time at 7.2kW. The total number of chargers is $C = 48$, with 16 chargers per phase. The time evolution of the total number of charging vehicles $N(t)$ is depicted in Figure 7a; we observe the daily cycle.

We then simulate the system using the discussed policies and compute the time-varying imbalance metric $|I^-(t)|$ for the resulting occupation trajectory $X(t)$. The time evolution using both policies is depicted in Figure 7b, where we can see that LLP halves the amount of imbalance current in the system on average. Moreover, the bounds from Propositions 6.2 and 6.3 are also shown in the graph. As we can see, the real time imbalance measure remains most of the time below the computed bounds, with some deviations occurring during transient situations.

One of the key takeaways of these numerical experiments is that random free spaces routing alone in a parking lot may be *insufficient* to control imbalance currents in the system, therefore straining the infrastructure and possibly tripping system protections due to this imbalance, leading to heavy performance penalties. There is an incentive for operators to actively route users to the least loaded phase in order to reduce this imbalance, and our bounds help design the system in a robust way.

7. CONCLUSIONS

In this paper, we have analyzed load balancing between server pools that serve tasks in parallel and may be subject to capacity constraints. We showed how a suitable measure of system imbalance behaves under random routing in the unconstrained case, and how it deviates from this behavior under capacity constraints. We also obtained sharp bounds on average imbalance for the case where least-loaded-pool routing is used. Finally, we applied our results to the problem of EV charging, where imbalance is a practical limitation, and showed how the bounds derived can be used to estimate it independently of the traffic load.

In future work, we would like to analyze sampling based policies such as a Power-of-d version of LLP. Such a policy would be more amenable to implementation in large scale systems where information passing is a concern. The case of heterogeneous pools would also be an interesting line of work, where imbalance metrics should be defined accordingly and the results in this paper are not straightforward to generalize. Finally, in the case of EV charging, it would be interesting to compare the behavior of the simple LLP policy against other proposals based on online optimization algorithms, that use more information from the system.

8. REFERENCES

[1] D. Acuña, A. Ferragut, F. Paganini, and E. Briglia. Symmetrical components analysis for managing phase imbalance in EV charge scheduling. In *Proc. of the 13th ACM Conference on Intelligent Energy Systems (ACM e-Energy)*, pages 401–405, June 2022.

[2] M. Alanyali and B. Hajek. Analysis of simple algorithms for dynamic load balancing. *Mathematics of Operations Research*, 22(4):840–871, 1997.

[3] L. R. Araujo, D. Penido, S. Carneiro, and J. L. R. Pereira. A three-phase optimal power-flow algorithm to mitigate voltage unbalance. *IEEE Transactions on Power Delivery*, 28(4):2394–2402, 2013.

[4] N. B. Arias, J. C. López, M. J. Rider, and J. Fredy Franco. Adaptive robust linear programming model for the charging scheduling and reactive power control of EV fleets. In *IEEE Madrid PowerTech*, pages 1–6, 2021.

[5] A. R. Bergen and V. Vittal. *Power systems analysis.* Prentice Hall, 2000.

[6] T. Bonald, M. Jonckheere, and A. Proutiére. Insensitive load balancing. In *Proc. of ACM SIGMETRICS/Performance*, page 367–377, 2004.

[7] J. De Hoog, T. Alpcan, M. Brazil, D. A. Thomas, and I. Mareels. Optimal charging of electric vehicles taking distribution network constraints into account. *IEEE Transactions on Power Systems*, 30(1):365–375, 2014.

[8] M. V. der Boor, S. C. Borst, J. S. Van Leeuwaarden, and D. Mukherjee. Scalable load balancing in networked systems: A survey of recent advances. *SIAM Review*, 64(3):554–622, 2022.

[9] A. Ephremides, P. Varaiya, and J. Walrand. A simple dynamic routing problem. *IEEE transactions on Automatic Control*, 25(4):690–693, 1980.

[10] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72:311–359, 2012.

[11] V. Gupta, M. Harchol Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

[12] B. Hajek. *Random processes for engineers.* Cambridge University Press, 2015.

[13] L. Huang, D. Chen, C. S. Lai, Z. Huang, A. F. Zobaa, and L. L. Lai. A distributed optimization model for mitigating three-phase power imbalance with electric vehicles and grid battery. *Electric Power Systems Research*, 210:108080, 2022.

[14] IEEE. IEEE Std 1159-2019, recommended practice for monitoring electric power quality - Redline. Technical report, IEEE, 2019.

[15] M. Jonckheere and B. Prabhu. Asymptotics of insensitive load balancing and blocking phases. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, page 311–322, 2016.

[16] Z. Lee, T. Li, and S. H. Low. ACN-Data: Analysis and applications of an open EV charging dataset. In *Proc. of the 10th International conference on Future energy systems (ACM e-Energy 2019)*, 2019.

[17] Z. J. Lee, D. Chang, C. Jin, G. S. Lee, R. Lee, T. Lee, and S. H. Low. Large-scale adaptive electric vehicle charging. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7, 2018.

[18] Z. J. Lee, S. Sharma, D. Johansson, and S. H. Low. ACN-Sim: An open-source simulator for data-driven electric vehicle charging research. *IEEE Transactions on Smart Grid*, 12(6):5113–5123, 2021.

[19] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 68(11):1056–1071, 2011.

[20] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, 2001.

[21] D. Mukherjee, S. C. Borst, J. S. Van Leeuwaarden, and P. A. Whiting. Asymptotic optimality of power-of-d load balancing in large-scale systems. *Mathematics of Operations Research*, 45(4):1535–1571, 2020.

[22] N. Nguyen and T. Kim. Toward highly scalable load balancing in kubernetes clusters. *IEEE Communications Magazine*, 58(7):78–83, 2020.

[23] International Electrotechnical Comission. Electric vehicle conductive charging system - Part 1: General requirements. Standard, IEC, Geneva, Switzerland, 2017.

[24] A. N. Tantawi and D. Towsley. Optimal static load balancing in distributed computer systems. *Journal of the ACM (JACM)*, 32(2):445–465, 1985.

[25] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[26] W. Weng and W. Wang. Achieving zero asymptotic queueing delay for parallel jobs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(3):1–36, 2020.

[27] Z. Ye, T. Li, and S. Low. Towards balanced three-phase charging: Phase optimization in adaptive charging networks. *Electric Power Systems Research*, 212:108322, 2022.

[28] L. Ying, R. Srikant, and X. Kang. The power of slightly more than one sample in randomized load balancing. *Math. of Op. Research*, 42(3):692–722, 2017.

[29] J. Zhao, X. Liu, C. Lin, and W. Wei. Three-phase unbalanced voltage/var optimization for active distribution networks. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2016.