Achieving fairness for EV charging in overload: a fluid approach

Martín Zeballos
Universidad ORT Uruguay
Montevideo, Uruguay
mzeballos@uni.ort.edu.uy

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

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

ABSTRACT

With the emergence of Electrical Vehicles (EVs), there is a growing investment in power infrastructure to provide charging stations. In an EV parking lot, typically not all vehicles can be charged simultaneously, and thus some scheduling must be performed, taking into account the time the users are willing to spend in the system.

In this paper, we analyze the performance of several common scheduling policies through a fluid model. We show that in overload, the amount of unfinished work is the same for all policies, but these can distribute the work performed unfairly across users. We also introduce a new policy called Least Laxity Ratio that achieves a suitable notion of fairness across jobs, and validate its performance by simulation.

Keywords

Scheduling, Deadlines, EV charging, Fluid model

1. INTRODUCTION

Electrical Vehicles (EVs) are becoming increasingly common in recent years; their penetration is prompted by the scarcity of fossil fuels and concerns on carbon emissions, and enabled by the lowering cost of energy storage technologies. A large scale deployment of EVs in the near term is now conceivable, posing new demands on the charging infrastructure [13,15]. A promising solution is to provide charging stations at a parking lot. Since the amount of power required to charge each EV is significant, it may not be practical to size the system capacity to provide simultaneous power to all chargers; note the utilization of the parking lot may not be that high. If a lower sizing is adopted, however, the charging capacity may be overloaded during busy hours, and thus some *scheduling* algorithm must be introduced to determine the order in which vehicles should be charged.

A typical characteristic of EV charging jobs is that they can be *deferred* to a certain extent. This means that a given vehicle can receive service immediately, or be delayed, because its sojourn time in the parking lot may be larger than the time required to provide a full charge. On the other hand, a feature of this problem is that if a vehicle leaves the system before full service, the partial amount of work performed is still useful. This sets the problem apart from classical scheduling of deadline constrained computing tasks [12].

IFIP WG 7.3 Performance 2018 Toulouse, France

In this paper, we start in Section 2 with a queueing model for an EV parking lot, where vehicles arrive with random demands and sojourn times. We assume that each charging station as well as the total installation have a power constraint, and analyze policies to schedule the jobs. A *fluid* model is then introduced for a large scale situation: it takes the form of an *advection* equation for the density of EVs flowing through the service-sojourn space. The service policy impacts the vector field of this flow, a formulation that includes several well known policies. We show that in underload all efficient policies have the same equilibrium. When the system is in *overload*, overall reneged work is the same for all efficient policies, but its distribution across jobs is highly dependent on the policy implemented. In Section 3, we analyze several policies and show that the reneged service may be unfairly distributed across users.

In Section 4 we propose a new policy called *least-laxity-ratio*, that is efficient in underload and achieves a suitable notion of fairness between users when the system is in overload. Finally, in Section 5 we provide empirical validation of the fluid approximation by simulating the discrete system in detail. Conclusions are given in Section 6.

1.1 Related work

The vehicle charging problem is a current concern for system operators. The increasing penetration rate of EVs leads to the development of optimization techniques to fully utilize resources. In [14], the authors discuss the use of future demand estimation to maximize the state of charge of vehicles upon departure. In [4, 17, 18], the authors analyze the scheduling problem by formulating a dynamic programming approach, and prove properties of the optimal algorithm to minimize reneged work. Similar ideas in a time-varying environment with real data were analyzed in [11]. More relevant to our work, a queueing approach is proposed in [2], where a simple model for parallel chargers is analyzed. The authors in [3] also extend this model to include network considerations for the grid.

In this paper, we seek to extend the results of [2] by considering more general scheduling policies when subject to deadlines. In this regard [6, 8] laid the foundations for the analysis of reneged work in deadline systems with a single server. Deadline based policies such as earliest-deadline-first (EDF) were thoroughly analyzed in [5, 10] in a fluid and diffusion scale. A more comprehensive analysis of fluid limits for earliest deadline policies and many server queues is done in [1, 9], and constitutes an ongoing research subject. Here we focus on extending this analysis by finding a common fluid description and extract results for system behavior.

2. SYSTEM MODEL

We consider an EV parking lot where each parking spot has an associated charging station. We assume that the size of the parking lot is large (infinite), but the total power consumed at any given time by the installation is limited by a finite capacity. The scheduling policy must allocate these limited resources among the EV clients currently present, taking into account their energy needs and their planned departure times.

In addition to the overall capacity limit, each charging station has a nominal power rating (maximum charging rate) that can be delivered to each EV. This quantity is assumed uniform across the parking lot, and typically much smaller than capacity.

2.1 Discrete queueing model

To motivate the fluid model which will be the main analytical tool of this paper, we consider first a discrete, stochastic counterpart.

Here, vehicles arrive as a Poisson process of intensity λ, and arriving vehicles choose two random characteristics in i.i.d. fashion: a required *service time* S_k, i.e. the energy requested divided by the nominal rate of charge of the station, and a *sojourn time* T_k, which is the time until the car leaves the parking lot. We assume that S_k and T_k follow general distributions, and $T_k \geqslant S_k$ with probability 1, which amounts to assuming that the demand of each EV is a priori feasible at the charging station. We denote by $f(\sigma, \tau)$ the joint density of (S, T).

The allocation decision in the hands of the garage operator is to assign to each vehicle a *charging rate* $r_k(t)$; we will normalize to unity the maximum individual charging rate, and thus require that

$$0 \leqslant r_k(t) \leqslant 1 \quad \text{for every } k, t. \tag{1}$$

Also, the total capacity of the installation is bounded, so we impose

$$\sum_{k=1}^{n(t)} r_k(t) \leqslant C, \tag{2}$$

where $n(t)$ is the number of EVs present in the garage which still require service. C can be interpreted as the maximum number of chargers that could be simultaneously turned on at full rate; we could, however, choose to activate more than C chargers at a reduced rate.

We will consider charging policies that take into account the current population of EVs, and their *residual* times; for these, the system state can be represented as a counting measure on the service - sojourn space, as in [8], namely:

$$\Phi_t = \sum_{k=1}^{n(t)} \delta_{(\sigma_k(t), \tau_k(t))}.$$

Here $\delta_{(\sigma_k(t),\tau_k(t))}$ is a point-mass measure in $\mathbb{R}^2$ located at the point $(\sigma_k(t), \tau_k(t))$, where $\sigma_k(t)$ is the remaining service time of each task and $\tau_k(t)$ is the remaining time until departure.

The system dynamics is as follows: each point k in the system consumes service time at a rate $r_k(t)$ and its lead time or time-to-deadline at rate 1, as depicted in Figure 1. The scheduling policy can thus be represented by a (possibly state-dependent) vector field on $\mathbb{R}^2_+$ given by:

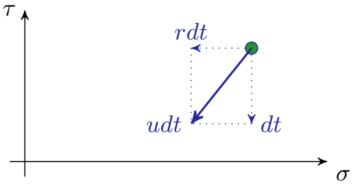

Figure 1: Dynamics for each job.

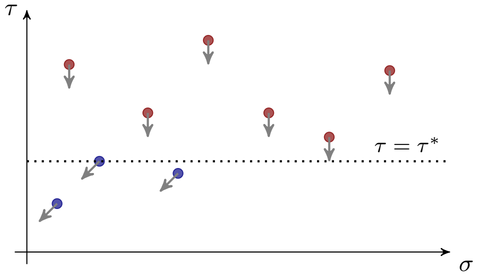

Figure 2: EDF policy behavior for $n(t) = 9$ and $C = 3$.

$$u(\sigma, \tau, \Phi) = -\left(r(\sigma, \tau, \Phi),\ 1\right). \tag{3}$$

Points follow this vector field up to reaching $\sigma = 0$ (charge completed), or their time expires at $\tau = 0$. Vehicles that are completely charged may stay in the parking lot, but we consider them served and out of our system. Vehicles that exhaust their time constraint depart the system with partial charge (reneged service).

We restrict our attention to policies that do not waste charging opportunities; specifically, we require the following:

Definition 1. *A charging policy is called* efficient *if at every time t, either* (2) *is satisfied with equality, or* (1) *is at its upper bound for every $k = 1, \ldots, n(t)$.*

As an example, consider that vehicles are served under the *earliest deadline first (EDF)* policy, where the first C vehicles with closer deadlines are served at rate $r = 1$. Then the system dynamics has the form depicted in Figure 2, and the rate r in equation (3) is:

$$r(\sigma, \tau, \Phi) = \mathbf{1}_{\{\tau \leqslant \tau^*(\Phi)\}} \tag{4}$$

with $\tau^*(\Phi) = \inf\{\tau : \Phi(\mathbb{R}^+ \times [0, \tau]) \geqslant C\}$.

The process Φ_t, determined by the arrival distributions and the scheduling policy r, is a measure-valued Markov process. Such a detailed system description is in general difficult to analyze. In a large scale situation, it is more tractable to consider a macroscopic, fluid description of the dynamics.

2.2 Fluid model

In the fluid scale, the number of points in the system is large and we replace Φ by a density function over $\mathbb{R}^2_+$. Let $g(t, \sigma, \tau)$ be the density of points that at time t have remaining work $\in [\sigma, \sigma + d\sigma]$ and remaining time $\in [\tau, \tau + d\tau]$. New mass arrives into the system at rate $\lambda f(\sigma, \tau)$, where f is the joint density of service and sojourn times, as before. Mass is transported along the vector field $u = -(r(\sigma, \tau, g), 1)$ defined by the scheduling policy.

This implies that the density g should satisfy the following *advection equation*:

$$\frac{\partial g}{\partial t} + \nabla \cdot (gu) = \lambda f \tag{5}$$

where $\nabla \cdot (\cdot)$ is the divergence operator on $\mathbb{R}^2_+$, i.e. on the variables σ, τ.

To explain this model, consider a region $\mathcal{R}$ of the (σ, τ) plane, with boundary $\partial\mathcal{R}$. The total mass of particles (EVs) within this region at time t is given by

$$\Phi_t(\mathcal{R}) = \iint_{\mathcal{R}} g(t, \sigma, \tau) d\sigma d\tau.$$

The variation of this quantity over time is due to arriving mass minus flow across the boundary. We therefore have:

$$\frac{d\Phi_t(\mathcal{R})}{dt} = \lambda \iint_{\mathcal{R}} f(\sigma, \tau) d\sigma d\tau - \int_{\partial\mathcal{R}} g(t, \sigma, \tau) u(\sigma, \tau, g) \cdot d\vec{n}.$$

Here $\vec{n}$ denotes the direction normal to the boundary; this second term can be transformed using the divergence theorem, leading to

$$\iint_{\mathcal{R}} \frac{\partial g}{\partial t} d\sigma d\tau = \lambda \iint_{\mathcal{R}} f d\sigma d\tau - \iint_{\mathcal{R}} \nabla \cdot (gu) d\sigma d\tau,$$

which is the integral version of (5).

Remark 1. *The formal relationship between stochastic and fluid models is usually framed in terms of scaling limits; for these measure-valued processes it is a very technical subject, beyond our scope here. Relevant references are [1, 5, 6, 8, 10].*

We are interested in steady-state solutions of eq. (5) so we set $\frac{\partial g}{\partial t} = 0$ and substitute the vector field, obtaining the equilibrium condition:

$$\frac{\partial (rg)}{\partial \sigma} + \frac{\partial g}{\partial \tau} + \lambda f = 0. \tag{6}$$

In the fluid model, the steady-state population of EVs is given by

$$n = \iint g(\sigma, \tau) d\sigma d\tau.$$

Finally, the constraints on the scheduling policy are now:

$$0 \leqslant r(\sigma, \tau, g) \leqslant 1;$$

$$\iint r(\sigma, \tau, g) g(\sigma, \tau) d\sigma d\tau \leqslant C.$$

In the next section we will analyze the steady state behavior of several scheduling policies by solving eq. (6) in each case.

2.3 Underload and Overload

The system *load* is defined as the product of the arrival rate and the mean service requirement,

$$\rho := \lambda E[S_k] = \lambda \int_0^\infty \int_0^\infty \sigma f(\sigma, \tau) d\sigma d\tau.$$

Remark 2. *The load represents the average power requested to the garage; however, since we are representing service in units of time, then load is a dimensionless quantity. ρ represents the mean number of chargers needed to fully satisfy the demand.*

We now state some general results that apply to all efficient policies, depending only on the load conditions. Proofs are omitted due to space limitations.

In the *underload* case, independently of the specific policy all vehicles receive full service in steady state, hence there is a common equilibrium.

Proposition 1. *Assume $\rho < C$. The steady state for any efficient policy is such that $r \equiv 1$, $n = \rho$, and*

$$g(\sigma, \tau) = \lambda \int_0^\infty f(\sigma + x, \tau + x) dx.$$

In the *overload* case, where the system load exceeds capacity, the steady-state distribution will depend on the specific policy. However the total amount of *reneged work* in the system, i.e. requested energy not delivered, is the same for all efficient policies. Again, reneged work is expressed here in dimensionless units.

Proposition 2. *Assume that $\rho > C$, and that the scheduling policy is efficient. Then the amount of reneged work in steady state is*

$$W = \int_0^\infty \sigma g(\sigma, 0) d\sigma = \rho - C. \tag{7}$$

An important question is, however: how is this reneged service *distributed* between individual vehicles in the system? In the next section we will find that in this aspect the various scheduling policies differ, and not always in an intuitive way. In particular, policies such as EDF which take into account lead time, do *not* end up discriminating reneged service based on users' time in the system.

3. SCHEDULING POLICIES IN OVERLOAD

We have seen that when the system is in underload ($\rho < C$), all efficient policies in the fluid limit behave as an infinite server queue. We now analyze in more detail the overload case ($\rho > C$) for different policies, with particular focus on the distribution of reneged work.

3.1 Earliest deadline first

We begin by considering the EDF policy already described, depicted in Figure 2. The rate function for this policy is the fluid counterpart of (4):

$$r(\sigma, \tau, g) = \mathbf{1}_{\{\tau \leqslant \tau^*(g)\}}. \tag{8}$$

Eq. (8) says there is a threshold τ^*, dependent on the state g such that loads with remaining sojourn time than τ^* do not receive service. At equilibrium, this value is fixed. Hence, the typical service profile is to wait until the time-to-deadline is τ^* and then be served up to completion or deadline expiration.

Using the method of characteristic curves [7] we can construct explicit solutions for the PDE (6) under rate function (8), and prove the following result:

Proposition 3. *Under the EDF policy in overload, the reneged work S_r per user is distributed as $(S - \tau^*)^+$, where the threshold τ^* satisfies*

$$\lambda E[\min\{S, \tau^*\}] = C. \tag{9}$$

Note that the above equation has a single solution $0 < \tau^* < \infty$ provided that $\rho > C$.

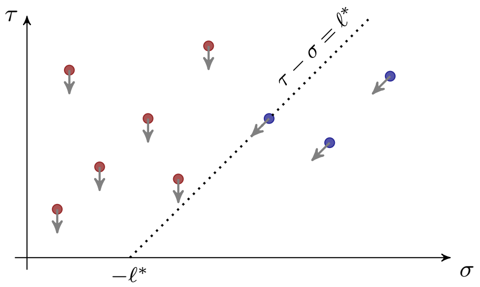

Figure 3: LLF policy behavior for $n(t) = 9$ and $C = 3$.

When the system is in overload, EDF finds a threshold τ^* given by (9): jobs with service time $S < \tau^*$ are then served to completion. Jobs with service time greater that τ^* are only delayed and served for a time τ^* and depart with reneged work $S_r = (S - \tau^*)^+$. Therefore, the policy is unfair towards large jobs, which get their service chopped to the threshold τ^*.

Moreover, the service received is independent of the sojourn time T: jobs that offer more flexibility are only delayed and not served at all until their remaining time is τ^*.

3.2 Least laxity first

The second policy we analyze is Least Laxity First (LLF) from the real-time scheduling literature [16], and discussed in the EV context in [17]. Here, the *laxity* or *spare time* of each job is considered, defined by $\ell_k = \tau_k - \sigma_k$, i.e. the amount of time that the job can be delayed and still be able to meet its deadline. The LLF policy, depicted in Figure 3, fills capacity with the jobs of lowest laxity, so the rate function is

$$r(\sigma, \tau, g) = \mathbf{1}_{\{\tau - \sigma \leqslant \ell^*(g)\}}. \tag{10}$$

In this case, the system serves loads with laxity $\ell_k \leqslant \ell^*$ at full rate, while the remaining consume their spare time up to reaching this level. In equilibrium, this laxity level ℓ^* is fixed, and in overload this laxity level becomes negative, implying that all jobs depart with reneged work. The equilibrium of (6) with rate function (10) is again computed using characteristic curves, and we can prove:

Proposition 4. *Under the LLF policy in overload, the reneged work S_r per user is distributed as* $\min\{S, \sigma^*\}$*, where the threshold $\sigma^* = -\ell^*$ satisfies*

$$\lambda E[(S - \sigma^*)^+] = C. \tag{11}$$

When the system is in overload, it finds a threshold $\ell^* < 0$ and all EVs with initial laxity $\ell > \ell^*$ consume their spare time up to reaching ℓ^* and leave the system with $S_r = \sigma^* := -\ell^*$ when their deadlines expire. However, if a given job arrives with service request $S < \sigma^*$, it never attains the required laxity level for service and departs the system without being charged at all, leaving with $S_r = S$. Thus, an LLF system in overload discriminates against small jobs. Again, the system ends up not discriminating by T (time in the system), only job size.

3.3 Processor sharing

Finally, we discuss a policy that does not consider deadlines (or indeed, service times) in allocation decisions. The processor sharing allocation is as follows: if the total number of vehicles n is less than C, then the rate at which each vehicle is served is $r = 1$. If $n > C$, then available power is equally shared by all chargers, i.e. $r = C/n$. This policy was analyzed in [8] for the single server queue, where constraint (1) is not present, and discussed in the EV context in [2].

In equilibrium, a PS system will reach a rate r^*, homogeneous across EVs. The corresponding equilibrium PDE (6) is once more solved for this situation, yielding the corresponding result for the reneged work:

Proposition 5. *Under the PS policy in overload, the reneged work S_r per user is distributed as $(S - r^*T)^+$, where the equilibrium rate r^* satisfies*

$$\lambda E[\min\{S, r^*T\}] = C. \tag{12}$$

Once more, this equation has a single solution $0 < r^* < 1$ under the overload condition. Equation (12) is analogous to the fixed point equation derived in [8] for the single server PS queue, and in [2] for the EV case under exponential assumptions.

We find here that the reneged work depends on *both* the service requirement S and the sojourn time T offered to the system, rewarding EVs that offer more flexibility. This occurs despite the fact that deadlines are not explicitly considered. In contrast, deadline-based policies like EDF and LLF in overload do not offer any service differentiation based on T.

Note however that, sojourn times being equal, the PS policy favors small jobs, which are served to completion, while large jobs only get partial service.

4. FAIR SCHEDULING: LEAST LAXITY RATIO

The policies analyzed in Section 3 do not show a fair behavior in overload. Both EDF and PS discriminate against large jobs, by chopping their service, and LLF discriminates against small jobs, by not giving service at all. In this Section we propose a new policy that we call *Least-Laxity-Ratio* (LLR). We show that this policy has similar macroscopic behavior to EDF or LLF in overload, but the amount of service received by each job is proportional to their requested service.

The LLR policy works as follows: given the current set of jobs with remaining service and deadlines (σ_k, τ_k), construct the following index called *laxity ratio*:

$$\theta_k := \frac{\tau_k}{\sigma_k} = 1 + \frac{\ell_k}{\sigma_k}.$$

Then serve the C jobs with smallest θ_k in the system at full rate. The policy serves the most urgent loads, i.e. those with more urgent deadlines, *relative* to their residual service time.

The behavior of the policy is depicted in Figure 4. The rate allocated to each job is given by:

$$r(\sigma, \tau, g) = \mathbf{1}_{\{\frac{\tau}{\sigma} \leqslant \theta(g)\}}, \tag{13}$$

where $\theta(g)$ is the threshold laxity ratio. In equilibrium, this ratio reaches a value θ^* so that

$$\iint_{\{\tau \leqslant \theta^*(g)\sigma\}} g(\sigma, \tau) d\sigma d\tau = C.$$

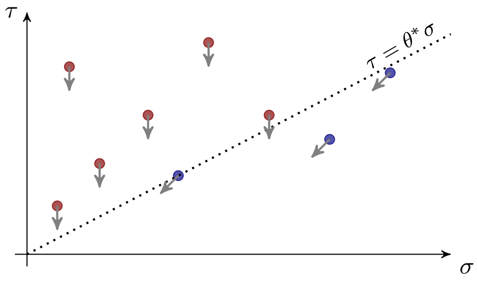

Figure 4: Least laxity ratio policy behavior for $n(t) = 9$ and $C = 3$.

Policy	Threshold equation	Reneged work (S_r)	Att. service $(S - S_r)$
EDF	$\lambda E[\min\{S, \tau^*\}] = C$	$(S - \tau^*)^+$	$\min\{S, \tau^*\}$
LLF	$\lambda E[(S - \sigma^*)^+] = C$	$\min\{S, \sigma^*\}$	$(S - \sigma^*)^+$
PS	$\lambda E[\min\{S, r^* T\}] = C$	$(S - r^* T)^+$	$\min\{S, r^* T\}$
LLR	$\lambda \theta^* E[S] = C$	$(1 - \theta^*) S$	$\theta^* S$

Table 1: Summary of performance metrics for the different algorithms.

In an equilibrium condition, an EV will receive no service until τ/σ falls below the threshold θ^*, and receive unit rate after that. In overload, $\theta^* < 1$ meaning that jobs must be lagging behind their deadline to get service, and will always have some reneged service. We have thus have the following result for this policy:

Proposition 6. *Under the LLR policy in overload, the reneged work S_r per user is distributed as $S(1 - \theta^*)$, where the threshold θ^* satisfies*

$$\theta^* = C/\rho. \tag{14}$$

Therefore, by using the LLR policy, received service is simply $\theta^* S$, a uniform downscaling of the service request S. This amounts to a notion of *proportional fairness* between jobs: all EVs will receive the same fraction of their required charge. This proportionality is arguably a fair way to distribute resources in the case of overload.

Finally, we summarize in Table 1 the properties derived for all the policies analyzed in the paper.

5. SIMULATIONS

We now validate the predictions of the fluid model by simulating the discrete system under the different policies analyzed before and compare its results with the fluid model predictions. The arrival rate is λ and sets the scale of the system. The job size S_k is exponentially distributed with rate μ. Each job arrives with an initial laxity L_k also exponentially distributed with rate γ, independent of S_k, and we set $T_k = S_k + L_k$. This corresponds to the joint density:

$$f(\sigma, \tau) = \mu\gamma e^{-(\mu-\gamma)\sigma - \gamma\tau} \quad 0 \leqslant \sigma \leqslant \tau.$$

The load of the system is thus $\rho = \lambda/\mu$ and we set $\rho > C$. By integrating this exponential distribution we can compute the different thresholds given in Table 1 yielding:

$$\tau^* = -\frac{1}{\mu} \log\left(1 - \frac{C}{\rho}\right) \qquad \text{EDF}$$

$$\sigma^* = -\frac{1}{\mu} \log\left(\frac{C}{\rho}\right) \qquad \text{LLF}$$

$$r^* : \frac{\gamma(1 - r^*)^2}{\mu r^* + \gamma(1 - r^*)} = 1 - \frac{C}{\rho} \qquad \text{PS}$$

$$\theta^* = \frac{C}{\rho} \qquad \text{LLR}$$

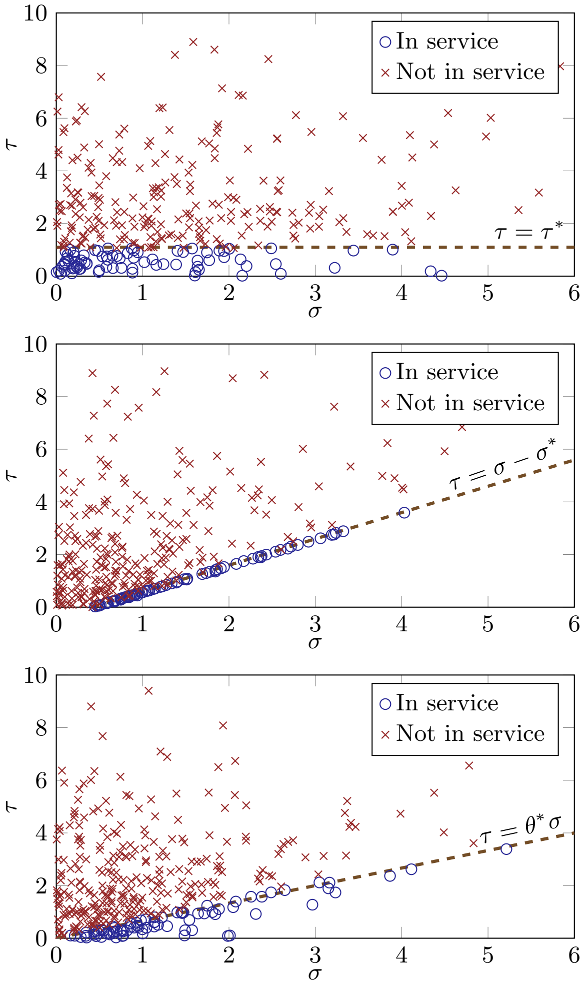

Figure 5: Steady-state snapshot of a system under EDF (above), LLF (center) and LLR (below), with in service and not in service loads. The dotted line indicates the corresponding thresholds predicted by the fluid model.

We simulate the system with $\lambda = 120$, $\mu = 1$, $\gamma = 0.5$ and $C = 80$, so the load is $\rho = 120 > C$, using the classical policies EDF and LLF as well as our proposal LLR. In Figure 5 we plot a snapshot of the system in steady-state, and the thresholds computed using the fluid model. As we can see, the fluid model predicts correctly the transition for all three policies. The number of loads in the system is on the scale of ρ, i.e. few hundred vehicles, highlighting the power of the

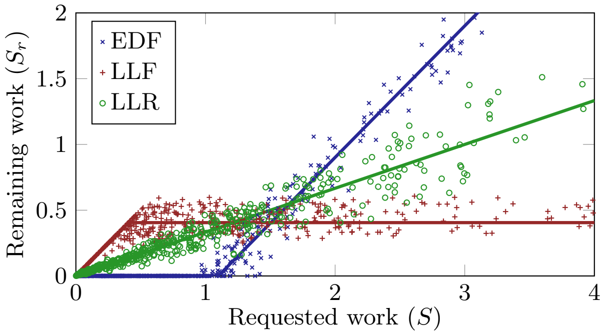

Figure 6: Requested and reneged work for each job under the different policies and fluid prediction.

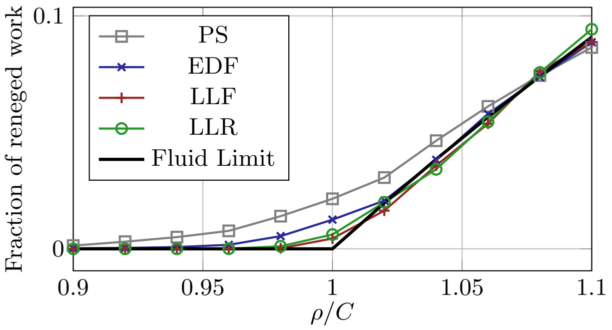

Figure 7: Reneged work for the different policies around the critical load $\rho = C$ and fluid limit.

fluid approximation, which even for a medium sized station captures the correct behavior.

More importantly, in Figure 6 we plot the initial demand S and final reneged work S_r achieved by the jobs under the three policies, as well as the fluid model predictions. As we can see, the empirical observations follow the predicted behavior closely, and the EDF and LLF policies discriminate against large and small jobs respectively. On the other hand, our proposed LLR policy achieves the desired linear relationship, imposing proportional fairness across jobs.

Finally, we explore the performance of the policies as we approach the overload condition. As discussed before, in the fluid model the rate of reneged work is 0 whenever $\rho < C$, and is $\rho - C$ when $\rho > C$, given by Proposition 2. Therefore, the fraction of reneged work is $(1 - C/\rho)^+$.

In Figure 7 we plot the empirical fraction of reneged work as ρ moves from underload to overload, and using the same demand parameters as before. In the discrete system around the critical point, the fraction of reneged work is highest for PS and lowest for LLF and LLR. Therefore, while our proposed policy was derived to perform well in overload, it empirically performs as well as the other policies across all load conditions, while maintaining fairness. Note that the fluid limit correctly captures the reneged work, however differences within policies remain, and around the critical load not all policies perform equally. Therefore, it would be interesting to enhance the analysis around criticality by using diffusion approximations to better differentiate across policies.

6. CONCLUSIONS

In this work, we analyze the performance of several common scheduling policies for EV charging by means of a fluid model, showing that in overload the amount of unfinished work is invariant across work conserving policies, but can be unfairly distributed across jobs. We introduced a new policy called Least Laxity Ratio that achieves fairness across jobs when in overload, while showing good behavior in underload and critical scenarios. We validated its performance by simulation experiments on the discrete system. In future work, we plan to address more refined approximations for the system to analyze its behavior around the critical load, as well as considering time-varying scenarios.

7. REFERENCES

[1] R. Atar, A. Biswas, and H. Kaspi. Fluid limits of G/G/1+G queues under the nonpreemptive earliest-deadline-first discipline. *Mathematics of Operations Research*, 40(3):683–702, 2014.

[2] A. Aveklouris, Y. Nakahira, M. Vlasiou, and B. Zwart. Electric vehicle charging: a queueing approach. *ACM SIGMETRICS Performance Evaluation Review*, 45(2):33–35, 2017.

[3] A. Aveklouris, M. Vlasiou, and B. Zwart. A stochastic resource-sharing network for electric vehicle charging. *arXiv preprint arXiv:1711.05561*, 2017.

[4] S. Chen, Y. Ji, and L. Tong. Large scale charging of electric vehicles. In *Power and Energy Society General Meeting, 2012 IEEE*, pages 1–9. IEEE, 2012.

[5] L. Decreusefond and P. Moyal. Fluid limit of a heavily loaded edf queue with impatient customers. *arXiv preprint math/0512660*, 2005.

[6] B. Doytchinov, J. Lehoczky, and S. Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Annals of Applied Probability*, pages 332–378, 2001.

[7] L. C. Evans. *Partial Differential Equations*. AMS, 1998.

[8] H. C. Gromoll, P. Robert, B. Zwart, and R. Bakker. The impact of reneging in processor sharing queues. *ACM SIGMETRICS Performance Evaluation Review*, 34(1), 2006.

[9] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Double skorokhod map and reneging real-time queues. In *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz*, pages 169–193. Institute of Mathematical Statistics, 2008.

[10] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for edf queues with reneging. *The Annals of Applied Probability*, 21(2):484–545, 2011.

[11] Y. Nakahira, N. Chen, L. Chen, and S. H. Low. Smoothed least-laxity-first algorithm for ev charging. In *Proceedings of the Eighth International Conference on Future Energy Systems*, pages 242–251. ACM, 2017.

[12] S. S. Panwar, D. Towsley, and J. K. Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *Journal of the ACM (JACM)*, 35(4):832–844, 1988.

[13] D. B. Richardson. Electric vehicles and the electric grid: A review of modeling approaches, impacts, and

renewable energy integration. *Renewable and Sustainable Energy Reviews*, 19:247–254, 2013.

[14] W. Su and M.-Y. Chow. Performance evaluation of an eda-based large-scale plug-in hybrid electric vehicle charging algorithm. *IEEE Transactions on Smart Grid*, 3(1):308–315, 2012.

[15] W. Su, H. Eichi, W. Zeng, and M.-Y. Chow. A survey on the electrification of transportation in a smart grid environment. *IEEE Transactions on Industrial Informatics*, 8(1):1–10, 2012.

[16] D. Towsley and S. S. Panwar. On the optimality of minimum laxity and earliest deadline scheduling for real-time multiprocessors. In *Real Time, 1990. Proceedings., Euromicro'90 Workshop on*, pages 17–24. IEEE, 1990.

[17] Y. Xu, F. Pan, and L. Tong. Dynamic scheduling for charging electric vehicles: A priority rule. *IEEE Transactions on Automatic Control*, 61(12):4094–4099, 2016.

[18] Z. Yu, Y. Xu, and L. Tong. Large scale charging of electric vehicles: A multi-armed bandit approach. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 389–395. IEEE, 2015.