Reinforced Workload Distribution Fairness

Zhiyuan Yao, Zihan Ding, Thomas Heide Clausen

Reinforced Workload Distribution Fairness

Zhiyuan Yao
Department of Computer Science
École Polytechnique
91120 Palaiseau, France
zhiyuan.yao@polytechnique.edu

Zihan Ding
Department of Electrical and Computer Engineering
Princeton University
08544 New Jersey, U.S.A.
zhding96@gmail.com

Thomas Clausen
Department of Computer Science
École Polytechnique
91120 Palaiseau, France
thomas.clausen@polytechnique.edu

Abstract

Network load balancers are central components in data centers, that distributes workloads across multiple servers and thereby contribute to offering scalable services. However, when load balancers operate in dynamic environments with limited monitoring of application server loads, they rely on heuristic algorithms that require manual configurations for fairness and performance. To alleviate that, this paper proposes a distributed asynchronous reinforcement learning mechanism to – with no active load balancer state monitoring and limited network observations – improve the fairness of the workload distribution achieved by a load balancer. The performance of proposed mechanism is evaluated and compared with state-of-the-art load balancing algorithms in a simulator, under configurations with progressively increasing complexities. Preliminary results show promise in RL-based load balancing algorithms, and identify additional challenges and future research directions, including reward function design and model scalability.

1 Introduction

In data centers and distributed systems, application servers are deployed on infrastructures with multiple servers, each with multiple processors to provide scalable services [1]. To optimize workload distribution and reduce additional queuing delay, load balancers (LBs) play a significant role in such systems, yet rely on heuristics for decisions on where to place each workload [2–5]. Reinforcement learning (RL) approaches [6–9] have shown performance gains in various system and networking problems. They help avoid error-prone manual configurations. This paper therefore studies the potential performance improvement when applying RL techniques on network load balancing problems in distributed systems.

Applications of RL techniques on network load balancing problems are challenging for several reasons. First, unlike traditional workload distribution problem [6, 7], network LBs have limited observations on task size and actual server load states. Being unaware of task size, application servers can be overloaded by collided "elephant tasks" and thus yield degraded quality of service. As network LBs operate on the Transport Layer, obtaining task size information before allocating workloads requires application-specific – and thus by definition non-universal – LB implementations [10, 11]. With services running on heterogeneous hardware or elastic data centers [12] where server capacities vary, "fairly" (uniformly) distributing workloads without considering actual server load-states may overload servers with limited resources and underload powerful servers. On the other hand, probing

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

actual server load states requires deploying agents on servers and incurs additional management traffic [4, 5]. Second, the interactive training procedure of RL models produces periodically and asynchronously updated actions in a slow pace (*e.g.*, every hundreds of milliseconds [9]) in the control plane, while system dynamics can change rapidly (*e.g.*, sub-millisecond in modern networked systems [13]) in the data plane.

This paper proposes an RL-based network load balancing mechanism (RLB-SAC) that is able to exploit asynchronous actions based only on local observations and inference. Performance is evaluated on an event-based simulator and compared with benchmark load balancing mechanisms.

2 RL for Network Load Balancing

Network load balancing can be defined as allocating a Poisson sequence of tasks with different workloads $\{w_k\} \in \mathcal{W}$[1] on a set of n servers, to achieve the maximal exploitation of the computational capacity of the servers. The task workload $w_k^j(t)$ assigned on the j-th server at time t usually follows an exponential distribution in practical experiments [14]. The load balancing method is a function $\pi \in \Pi$: $\mathcal{W} \to [n]$. The processing speed for each server is $s_j, j \in [n]$, *i.e.*, the amount of workloads that can be processed per unit of time. The load on the j-th server ($j \in [n]$) during a time interval $t \in [t_0, t_n]$ is thus defined as:

$$l_j = \frac{\sum_{t \in [t_0, t_n)} w_k^j(t)}{s_j}, \tag{1}$$

which represents the expected time to finish processing all the workloads on the j-th server.

The objective of load balancing can be defined as finding the optimal policy:

$$\pi^* = \min_{\pi \in \Pi} \max_j l_j \tag{2}$$

This problem is multi-commodity flow problems and is NP-hard, which makes it hard to solve with trivial algorithmic solution within micro-second level [15].

2.1 RL Methodology

To apply standard RL methods, the network load balancing problem can be formulated as a Markov Decision Process (MDP), which can be represented as $(\mathcal{S}, \mathcal{A}, R, \mathcal{T}, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ are the state space and the action space, and R is a reward function $R(s, a)$: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. The state-transition probability from current state and action to a next state $s' \in \mathcal{S}$ is defined by $\mathcal{T}(s'|s, a)$: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$. $\gamma \in (0, 1)$ is a reward discount factor. The goal of an RL algorithm is optimizing the policy to maximize the agent's expected cumulative rewards: $\mathbb{E}[\sum_t \gamma^t r_t]$.

Soft Actor-Critic (SAC) [16] follows the maximum entropy reinforcement learning framework, which optimizes the objective $\mathbb{E}[\sum_t \gamma^t r_t + \alpha \mathcal{H}(\pi_\theta)]$ to encourage the exploration $\mathcal{H}(\cdot)$ of the policy π_θ during the learning process. More concretely, the critic Q network is updated with the gradient $\nabla_\phi \mathbb{E}_{s,a}[Q_\phi(s, a) - [R(s, a) + \gamma \mathbb{E}_{s'}[V_{\tilde{\phi}}(s')]]]$, where $V_{\tilde{\phi}}(s') = \mathbb{E}_{a'}[Q_{\tilde{\phi}}(s', a') - \alpha \log \pi_\theta(a'|s')]$ and $Q_{\tilde{\phi}}$ is the target Q network; the actor policy π_θ is updated with the gradient $\nabla_\theta \mathbb{E}_s[\mathbb{E}_{a \sim \pi_\theta}[\alpha \log \pi_\theta(a|s) - Q_\phi(s, a)]]$.

Other key elements of RL methods involve the observation, action and reward function, which are detailed as following:

Observation. Based on the limited information that is visible to network LBs, *i.e.* task ID k, arrival and finish time of each task, the number of ongoing tasks $|w^j(t)|$ on the j-th server can be computed and the following 3 observations are sampled, *i.e.*, task inter-arrival time, task duration, and task completion time (TCT). Each sampled time-related feature channel is reduced to 5 scalars, *e.g.*, average, 90th-percentile, standard deviation, discounted average and weighted discounted average[2].

[1] The unit of workload can be, *e.g.*, amount of time to process.

[2] Discounted average weights are computed as $0.9^{t'-t}$, where t is the sample timestamp and t' is the moment of calculating the reduced scalar.

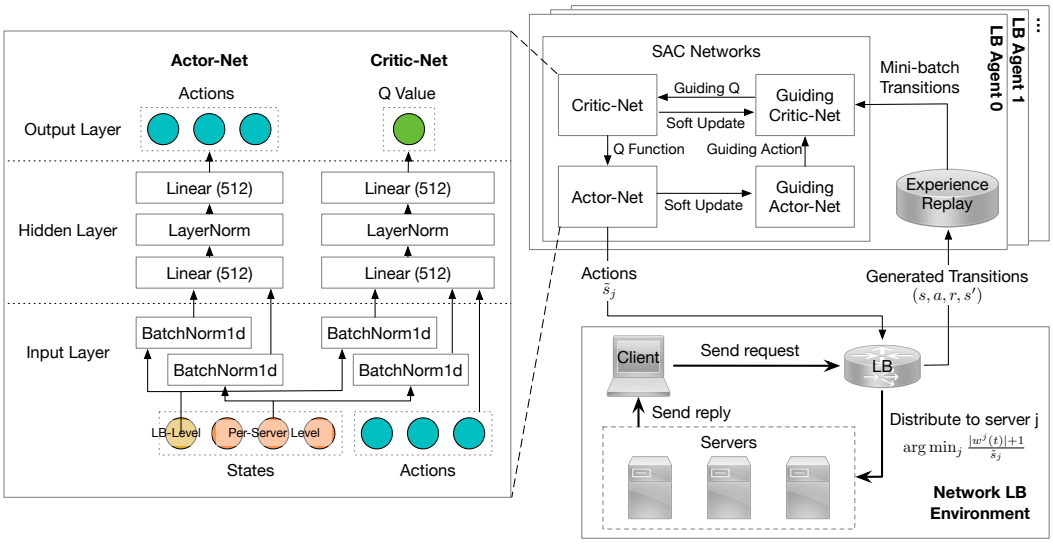

Figure 1: Overview of the proposed RL framework for network LB. A distributed learning framework with multiple LB agents is implemented to interact with the simulated LB and allocate tasks on different servers according to a server assignment function. Each LB agent contains a replay buffer and a SAC model with two pairs of actor and critic networks. Network architectures are shown on the left block. Linear (n) indicates the number of hidden units is n.

Task inter-arrival time is computed at LB level, while task duration and TCT are gathered at per-server level.

Action. To bridge the difference between slow-paced action updates and high-speed network packets arrivals, the RL agent assign the j-th server to newly arrived task using the ratio of two factors:

$$\arg\min_{j\in[n]} \frac{|w^j(t)|+1}{\tilde{s}_j}, \tag{3}$$

where the number of on-going tasks $|w^j(t)|$ helps track dynamic system server occupation and $\tilde{s}_j$ is the periodically updated RL-inferred server processing speed given observations, as the direct outputs of the LB agents. Eq. (3) is the server assignment function.

Reward. The reward is chosen as $1 - F(\tilde{\mathbf{w}})$, where $\tilde{\mathbf{w}}$ is a list of discounted average of actual TCT on each server, and F is a fairness measurement. In this paper, 3 fairness indices are studied, namely, Jain's, G's, and Bossaer's fairness index, which have the following forms respectively, $F_J(\tilde{\mathbf{w}}) = \frac{\overline{\tilde{\mathbf{w}}}^2}{\overline{\tilde{\mathbf{w}}^2}}$, $F_G = \prod_{j\in[n]} \sin(\frac{\pi \tilde{w}^j}{2\max(\tilde{\mathbf{w}})})$, $F_B = \prod_{j\in[n]} \frac{\tilde{w}^j}{\max(\tilde{\mathbf{w}})}$.

Model. The architecture of the proposed RL framework is depicted in figure 1. The SAC model within each LB agent contains a pair of actor and critic networks, as well as a pair of guiding actor and guiding critic networks, with the same network architectures as the former but updated in a delayed and soft manner. As shown in the left block of figure 1, the actor and critic networks take the batch-normalized features from LB-level states and per-server-level states and use the same Linear-LayerNorm-Linear layers to process the data independently. The critic additionally takes the actions as inputs to generate the estimated Q values.

3 Implementation

In order to compare and contrast performance of different load balancing methods in various scenarios, in particular those that are difficult to evaluate in testbeds, such as large-scale data center networks, an event-driven simulator is implemented.

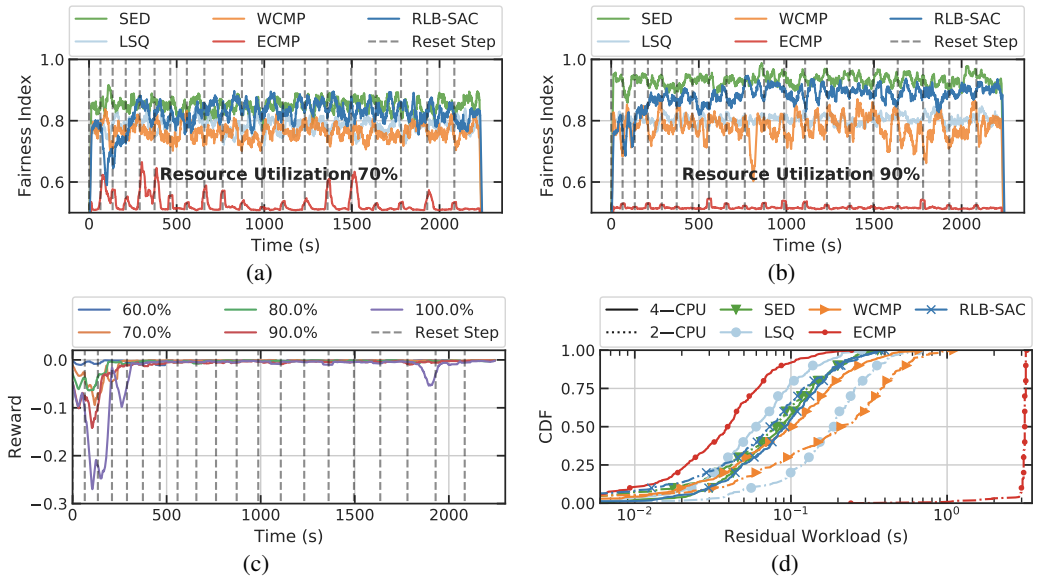

Figure 2: Comparison of load balancing performance (during training) under different traffic rates in a simplest setup where all tasks are the same. Figure(a) and (b) show Jain's fairness index of residual workloads l_j between 2 servers over time. Figure(c) shows the reward over 20 episodes under different traffic rates. Figure(d) compares the CDF of residual workloads l_j on 2 servers in the last episode.

3.1 Server Processing Model

Realistic network applications feature blocked processor sharing model [17, 18], in which the instantaneous processing speed $s_j(t)$ at time t on the j-th server is:

$$s_j(t) = \begin{cases} 1 & |w^j(t)| \leq p_j, \\ \frac{p_j}{\min(\hat{p}_j, |w^j(t)|)} & |w^j(t)| > p_j, \end{cases} \tag{4}$$

where $|w^j(t)|$ denotes the number of on-going tasks, and p_j denotes the number of processors on the j-th server. At any given moment, the maximum number of tasks that can be processed is $\hat{p}_j$. Tasks that arrive when $|w^j(t)| \geq \hat{p}_j$ will be blocked in a queue (similar to the backlog in *e.g.*, Apache), and will not be processed until there is an available slot in the CPU processing queue.

3.2 Benchmark Load Balancing Methods

To quantify the performance of the proposed LB method (RLB-SAC), 4 baseline workload distribution algorithms are implemented. Equal-cost multi-path (ECMP) randomly assigns servers to tasks with a server assignment function $\mathbb{P}(j) = \frac{1}{n}$, where $\mathbb{P}(j)$ denotes the probability of assigning the j-th server [19]. Weighted-cost multi-path (WCMP) assigns servers based on their weights derived, and has an assignment function as $\mathbb{P}(j) = \frac{p_j}{\sum p_j}$ [3]. Local shortest queue (LSQ) assigns the server with the shortest queue, *i.e.*, $\arg\min_{j\in[n]} |w^j(t)|$ [20]. And finally, shortest expected delay (SED) assigns the server the shortest queue normalized by the number of processors, *i.e.*, $\arg\min_{j\in[n]} \frac{|w^j(t)|+1}{p_j}$ [2], and is expected to have the best performance among baseline LB methods.

4 Experimental Evaluation

This section investigates the impacts of several key components for our proposed RL framework, including (i) the system dynamics properties like traffic rate and TCT (workload $w_k^j(t)$) distribution, (ii) the reward function in RL and (iii) the scalability of the system. Hyperparameters for training see Appendix. A.

Traffic Rate. Following the approach in [21], a simplest scenario, *i.e.*, all the tasks are identical (each has 100ms TCT), is created to verify the learning ability of RLB-SAC. Different traffic rates

Table 1: Comparison under different traffic rates ($60\%, 70\%, 80\%, 90\%, 100\%$) with the same TCT. "FI" and "RW" are "Jain's fairness index" and "average residual workloads" respectively in the last episode.

Method	Traffic Rate (%)									
	60		70		80		90		100	
	FI	RW	FI	RW	FI	RW	FI	RW	FI	RW
ECMP	0.646	0.130	0.512	1.225	0.512	1.504	0.515	1.553	0.517	1.568
WCMP	0.753	0.046	0.765	0.060	0.788	0.088	0.787	0.192	0.835	0.809
LSQ	0.768	0.046	0.791	0.062	0.789	0.090	0.806	0.137	0.864	0.607
SED	**0.841**	**0.033**	**0.855**	0.046	**0.902**	0.067	**0.929**	**0.099**	**0.989**	**0.417**
RLB-SAC	0.782	**0.033**	0.807	**0.040**	0.874	**0.064**	0.904	0.100	0.969	0.520

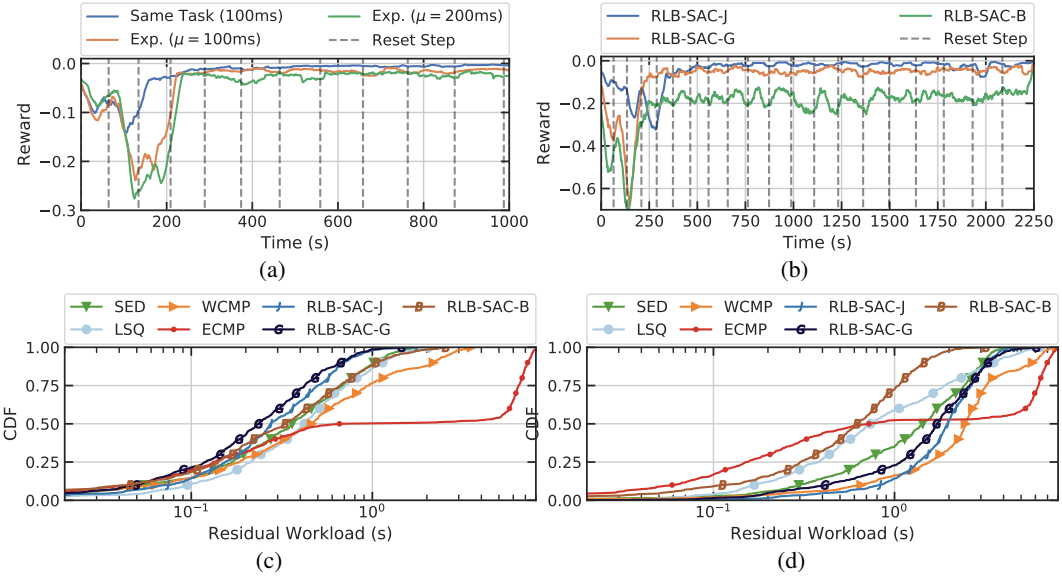

Figure 3: Figure(a) shows Jain's fairness index reward over 20 episodes using different TCT distributions under 90% traffic rate. Figure(b) shows different reward options under 100% traffic rate and exponential TCT distributions ($\mu = 200$ms). Figure(c) and (d) shows CDF of residual workloads using different LB mechanisms in the last episode, under 90% and 100% traffic of tasks with exponentially distributed TCT ($\mu = 200$ms) respectively. All results are obtained during the training process.

are applied on a system with 1 LB node and 2 servers, where one server has 4 CPUs and the other has 2 CPUs. The traffic rates are normalized to $[0, 1]$ by the processing capacity of configured server clusters so that 100% traffic rate creates a stable system state (equal average arrival and departure rate). Jain's fairness index is used to compute reward and evaluate workload distribution performance. The result of 20-episode training in simulation[3] is depicted in figure 2 and listed in table 1. After 5 episodes, the proposed SAC-based LB (RLB-SAC) improves its performance in terms of workload distribution fairness. RLB-SAC learns to behave similar to SED and assigns similar workloads l_j across the two servers. Though unlike SED, RLB-SAC requires no manual configuration.

TCT Distribution. In a more realistic setup, tasks are different and follows long-tail distributions [14]. As is depicted in figure 3a, the additional variance induced by exponential distribution TCT increases the difficulty of learning. Since the μ of exponential distributions is equivalent to its standard deviation, increasing the average TCT (from 100ms to 200ms) also increases the standard deviation of TCT. The additional variance in the TCT observation thus further reduces rewards obtained using RLB-SAC over time.

Reward Function. This section further compares different choices of the reward function in the RL procedure, including the Jain's (RLB-SAC-J), G's (RLB-SAC-G), and Bossaer's (RLB-SAC-B) fairness index as introduced in section 2.1. Compared against the baseline LB methods, depicted in

[3]The first episode lasts 60s and each following episode has 5s incremental duration than the former one. The time interval between two consecutive steps is 0.5s. The same configuration applies throughout the paper.

Table 2: Comparison of the last episode performance under different traffic rates ($60\%, 70\%, 80\%, 90\%, 100\%$) with the exponential TCT distribution ($\mu = 200$ms). "Max" and "Avg" stand for the maximum and average residual workload in the last episode.

Method	Traffic Rate (%)									
	60		70		80		90		100	
	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
SED	**0.127**	**0.776**	0.183	1.055	**0.238**	**0.951**	0.463	2.058	1.571	4.762
LSQ	0.173	1.306	0.225	1.246	0.304	1.770	0.522	2.038	1.352	6.408
WCMP	0.159	1.013	0.214	1.360	0.346	2.999	0.739	3.531	2.705	7.896
ECMP	0.662	3.085	1.365	5.565	3.088	9.129	3.233	8.158	3.032	8.048
RLB-SAC-J	0.145	0.952	0.173	1.188	0.299	2.208	0.350	1.747	2.004	4.608
RLB-SAC-G	0.162	0.994	**0.162**	**0.766**	0.245	1.573	**0.301**	**1.451**	1.846	6.072
RLB-SAC-B	0.146	1.072	0.174	1.082	0.265	1.239	0.464	2.538	**0.722**	**3.194**

Table 3: Comparison of the last episode performance under 70% traffic rate with the exponential TCT distribution ($\mu = 200$ms). "Max" and "Avg" stand for the maximum and average residual workload in the last episode.

Method	Network Topology							
	1LB-4S		2LB-4S		1LB-8S		2LB-8S	
	Avg	Max	Avg	Max	Avg	Max	Avg	Max
SED	**0.146**	**1.202**	**0.149**	**0.938**	**0.140**	**1.110**	**0.141**	**1.035**
LSQ	0.184	1.661	0.193	1.302	0.187	1.518	0.194	1.300
WCMP	0.232	2.190	0.223	2.803	0.236	2.822	0.236	2.736
ECMP	2.195	8.520	2.256	8.604	2.085	9.483	2.192	8.863
RLB-SAC-J	0.177	1.271	0.186	1.385	0.172	1.741	0.172	1.270
RLB-SAC-G	0.167	1.426	0.188	1.725	0.162	1.334	0.182	1.700
RLB-SAC-B	0.167	1.639	0.175	1.695	0.166	1.653	0.174	1.611

figure 3c and 3d, the SAC method with these three types of reward is tested under different traffic rates and exponential TCT distribution ($\mu = 200$ms). More complete results for the episodic maximum and average residual workload are listed in table 4. It shows that, without any prior knowledge of server processing capacities, the SAC methods work significantly better than other methods for cases with high traffic rates ($90\% - 100\%$), with close performance to the best SED method for relatively low traffic rates. Moreover, RLB-SAC-G and RLB-SAC-B fairness index have superior performance over RLB-SAC-J for most cases. This can be explained by the fact that Jain's fairness index provides values that are close to 1 unless "severe" unfairness is present, which also makes the reward close to 0. Bossaer's fairness index yields lower values and is more sensitive to unfairness (as in figure 3b), which makes it better when evaluating system state than G's fairness index especially when the traffic rate is higher (*e.g.*, 100% as in figure 3d).

Scalability. To study the scalability of the RLB-SAC model, 4 types of topology are applied where 1 or 2 LB agents distribute workloads across 4 or 8 servers, and half of all the servers have 4 CPUs while the other half have 2 CPUs. As the action space grows, the task becomes more challenging and RLB-SAC fails to surpass SED after 40 episodes of training. Though RLB-SAC achieves the second place among all methods under moderate traffic rate (70%).

5 Conclusion and Future Work

Promising as RL techniques are, it is challenging to adapt them to realistic systems, where, in the context of network load balancing, agents have limited observation over the environment and where periodic action updates may not be able to catch up with system dynamics. This paper takes a first step in studying the application of RL on network load balancing problems starting from the simplest assumptions that tasks are identical, towards more sophisticated (and perhaps more realistic) setups. The preliminary results of the proposed RLB-SAC LB show promise in the proposed RL-based load balancing framework comparing with baseline LB approaches (ECMP, WCMP, LSQ, SED). The design of the reward function requires further investigations so that RLB-SAC achieves better performance and improves generalization ability even under more realistic configurations. In presence of multiple LB agents and multiple application servers, the scalability of RLB-SAC is limited when the action space and the number of agents grow, which makes it an interesting direction for further research on adapting different RL models and algorithms.

Table 4: Comparison of the last episode performance under 90% traffic rate with the exponential TCT distribution ($\mu = 200$ms). "Max" and "Avg" stand for the maximum and average residual workload in the last episode.

Method	Network Topology							
	1LB-4S		2LB-4S		1LB-8S		2LB-8S	
	Avg	Max	Avg	Max	Avg	Max	Avg	Max
SED	**0.304**	**1.865**	**0.287**	**1.633**	**0.219**	**1.428**	**0.238**	**1.467**
LSQ	0.362	2.261	0.384	2.059	0.293	1.666	0.326	1.549
WCMP	0.704	6.698	0.816	5.110	0.684	4.512	0.695	4.983
ECMP	3.048	8.832	3.086	9.219	3.096	9.337	3.085	9.074
RLB-SAC-J	0.420	2.617	0.427	2.484	0.351	2.630	0.355	2.147
RLB-SAC-G	0.382	2.684	0.396	2.534	0.340	2.748	0.322	2.136
RLB-SAC-B	0.369	2.791	0.380	2.787	0.325	3.230	0.345	2.197

References

[1] Nicola Dragoni, Saverio Giallorenzo, Alberto Lluch Lafuente, Manuel Mazzara, Fabrizio Montesi, Ruslan Mustafin, and Larisa Safina. Microservices: yesterday, today, and tomorrow. In *Present and Ulterior Software Engineering*, pages 195–216. Springer, 2017.

[2] The Linux Virtual Server Project - Linux Server Cluster for Load Balancing. `http://www.linuxvirtualserver.org/`.

[3] Daniel E Eisenbud, Cheng Yi, Carlo Contavalli, Cody Smith, Roman Kononov, Eric Mann-Hielscher, Ardas Cilingiroglu, Bin Cheyney, Wentao Shang, and Jinnah Dylan Hosein. Maglev: A fast and reliable software network load balancer. In *NSDI*, pages 523–535, 2016.

[4] Yoann Desmouceaux, Pierre Pfister, Jérôme Tollet, Mark Townsley, and Thomas Clausen. 6lb: Scalable and application-aware load balancing with segment routing. *IEEE/ACM Transactions on Networking*, 26(2):819–834, 2018.

[5] Ashkan Aghdai, Michael I-C Wang, Yang Xu, Charles H-P Wen, and H Jonathan Chao. In-network congestion-aware load balancing at transport layer. *arXiv preprint arXiv:1811.09731*, 2018.

[6] Li Chen, Justinas Lingys, Kai Chen, and Feng Liu. Auto: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 191–205. ACM, 2018.

[7] Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh. Learning scheduling algorithms for data processing clusters. *arXiv preprint arXiv:1810.01963*, 2018.

[8] Yue Xu, Wenjun Xu, Zhi Wang, Jiaru Lin, and Shuguang Cui. Load balancing for ultra-dense networks: A deep reinforcement learning based approach. *IEEE Internet of Things Journal*, 6(6):9399–9412, Dec 2019. arXiv: 1906.00767.

[9] Viswanath Sivakumar, Tim Rocktäschel, Alexander H Miller, Heinrich Küttler, Nantas Nardelli, Mike Rabbat, Joelle Pineau, and Sebastian Riedel. Mvfst-rl: An asynchronous rl framework for congestion control with delayed actions. *arXiv preprint arXiv:1910.04054*, 2019.

[10] Willy Tarreau et al. Haproxy-the reliable, high-performance tcp/http load balancer, 2012.

[11] Will Reese. Nginx: the high-performance web server and reverse proxy. *Linux Journal*, 2008(173):2, 2008.

[12] Adithya Kumar, Iyswarya Narayanan, Timothy Zhu, and Anand Sivasubramaniam. The fast and the frugal: Tail latency aware provisioning for coping with load variations. In *Proceedings of The Web Conference 2020*, pages 314–326, 2020.

[13] Chuanxiong Guo, Lihua Yuan, Dong Xiang, Yingnong Dang, Ray Huang, Dave Maltz, Zhaoyi Liu, Vin Wang, Bin Pang, Hua Chen, et al. Pingmesh: A large-scale system for data center network latency measurement and analysis. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pages 139–152, 2015.

[14] Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C. Snoeren. Inside the social network's (datacenter) network. In *Proceedings of the 2015 ACM Conference on Special Interest*

Group on Data Communication, SIGCOMM '15, page 123–137. ACM, 2015. event-place: London, United Kingdom.

[15] Siddhartha Sen, David Shue, Sunghwan Ihm, and Michael J Freedman. Scalable, optimal flow routing in datacenters via local link balancing. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 151–162, 2013.

[16] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[17] Apache Hadoop. Apache hadoop. *URL http://hadoop. apache. org*, 2011.

[18] Apache Spark. Apache spark. *Retrieved January*, 17:2018, 2018.

[19] João Taveira Araújo, Lorenzo Saino, Lennert Buytenhek, and Raul Landa. Balancing on the edge: Transport affinity without network state. page 111–124, 2018.

[20] Guy Goren, Shay Vargaftik, and Yoram Moses. Distributed dispatching in the parallel server model. *arXiv:2008.00793 [cs]*, Aug 2020. arXiv: 2008.00793.

[21] Silvery Fu, Saurabh Gupta, Radhika Mittal, and Sylvia Ratnasamy. On the use of ml for blackbox system performance prediction. In *NSDI*, pages 763–784, 2021.

A Hyperparameters

Table 5: Hyperparameters in RL-based LB.

	Hyperparameter	Experiments			
		Traffic Rate	TCT Distribution	Reward Option	Scalability
SAC	Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
	Batch size	64	64	64	64
	Replay Buffer Size	3000	3000	3000	3000
	Episodes	20	20	20	40
	Step interval	0.5s	0.5s	0.5s	0.5s
	Target entropy	$-\|\mathcal{A}\|$	$-\|\mathcal{A}\|$	$-\|\mathcal{A}\|$	$-\|\mathcal{A}\|$
LB System	TCT Distribution	Identical Tasks	Exponential	Exponential	Exponential
	Average TCT	100ms	{100ms,200ms}	200ms	200ms
	First episode Duration	60s	60s	60s	60s
	Incremental Duration	5s	5s	5s	5s
	Last episode Duration	160s	160s	160s	260s