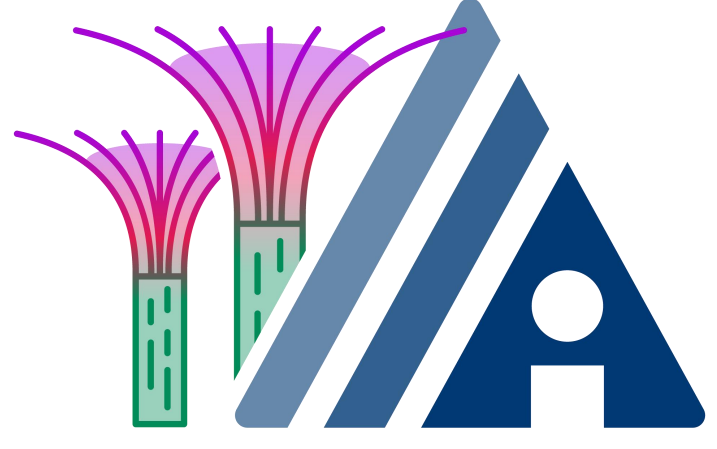


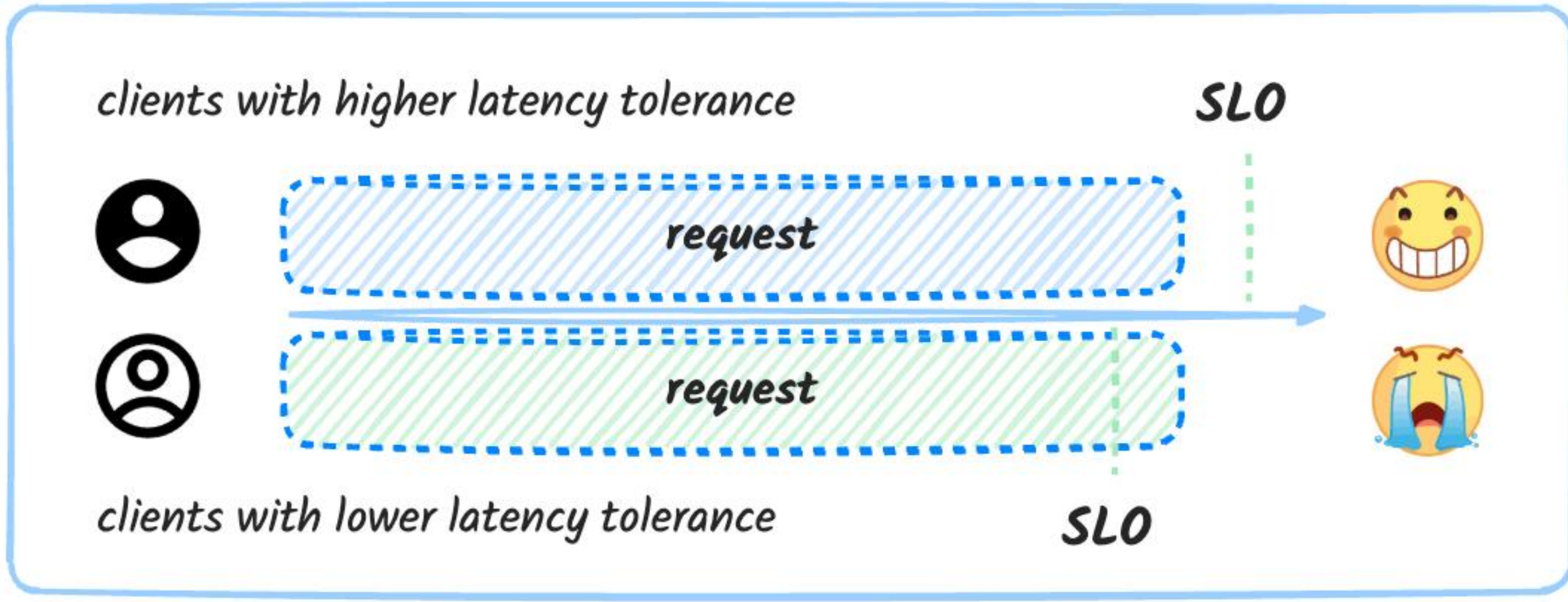
Experiential Fairness: Bridging the Gap Between User Experience and Resource-Centric Fairness in Online LLM Services



Jiahua Huang^{1,2}, Wentai Wu^{*3}, Yongheng Liu², Guozhi Liu¹, Yang Wang⁴, Weiwei Lin^{*1,2}

¹South China University of Technology ²Pengcheng Laboratory ³Jinan University ⁴Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

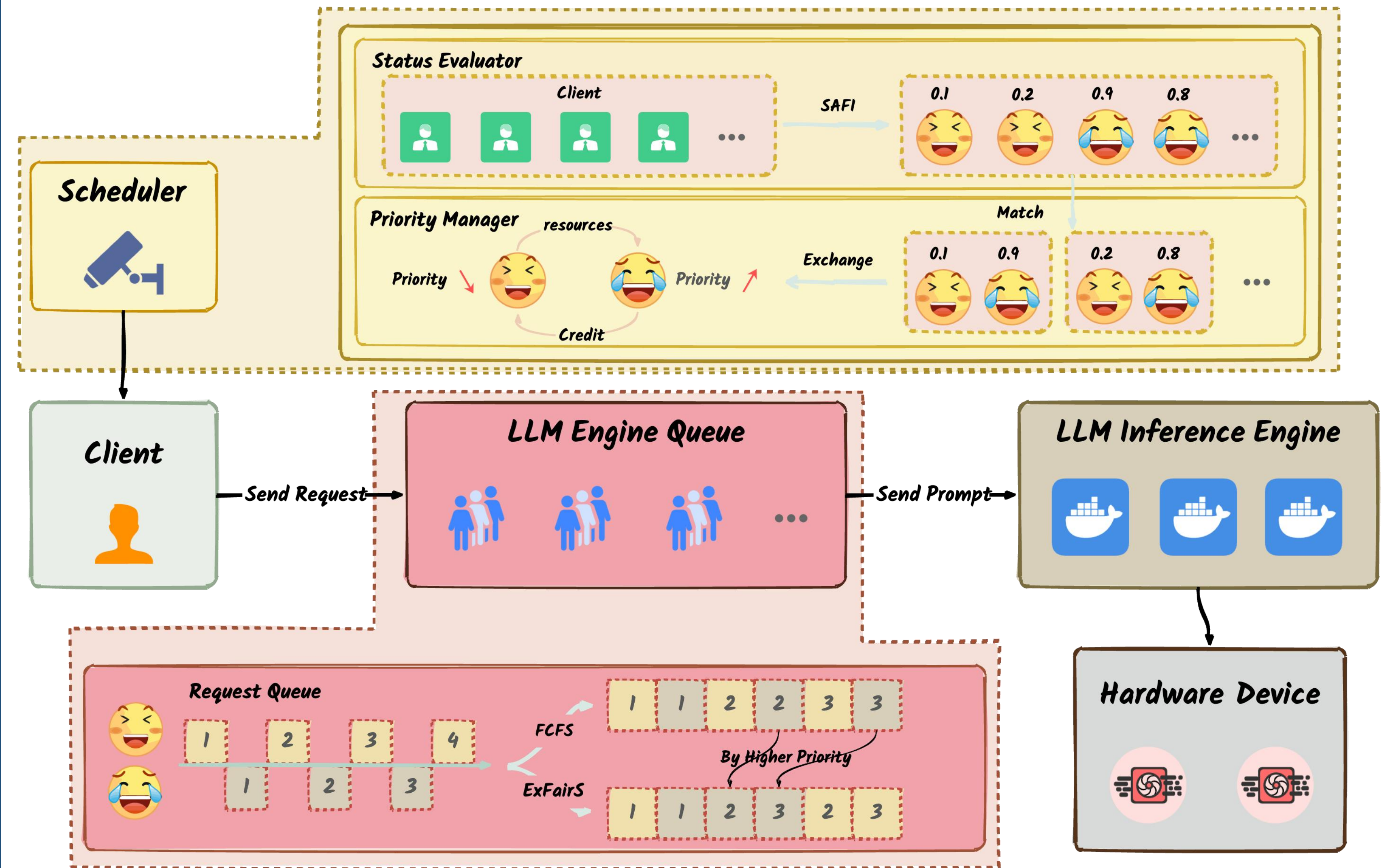
1. Introduction



- ◆ **The Trap--"Deceptive Fairness":** Conventional schedulers (e.g., FCFS, VTC) prioritize **System-Centric Fairness** (e.g., equalizing resource usage). We argue this creates an **"Experience-Blindness"** gap, leading to **"Deceptive Fairness"** where system metrics look good, but users are unhappy.
- ◆ **The Reality--Diverse Sensitivities:** As shown in the figure, users have different latency tolerances (SLOs). A **3s delay** might be acceptable for a complex reasoning task but is disastrous for an interactive chat. Equal resource allocation does **not** mean equal satisfaction.
- ◆ **Our Solution--Equity of Outcome:** We propose **Experiential Fairness**, shifting the scheduling objective from **Equality of Opportunity** (access to resources) to **Equity of Outcome** (SLO fulfillment). ExFairS is the first framework to explicitly quantify and optimize this user-centric metric.

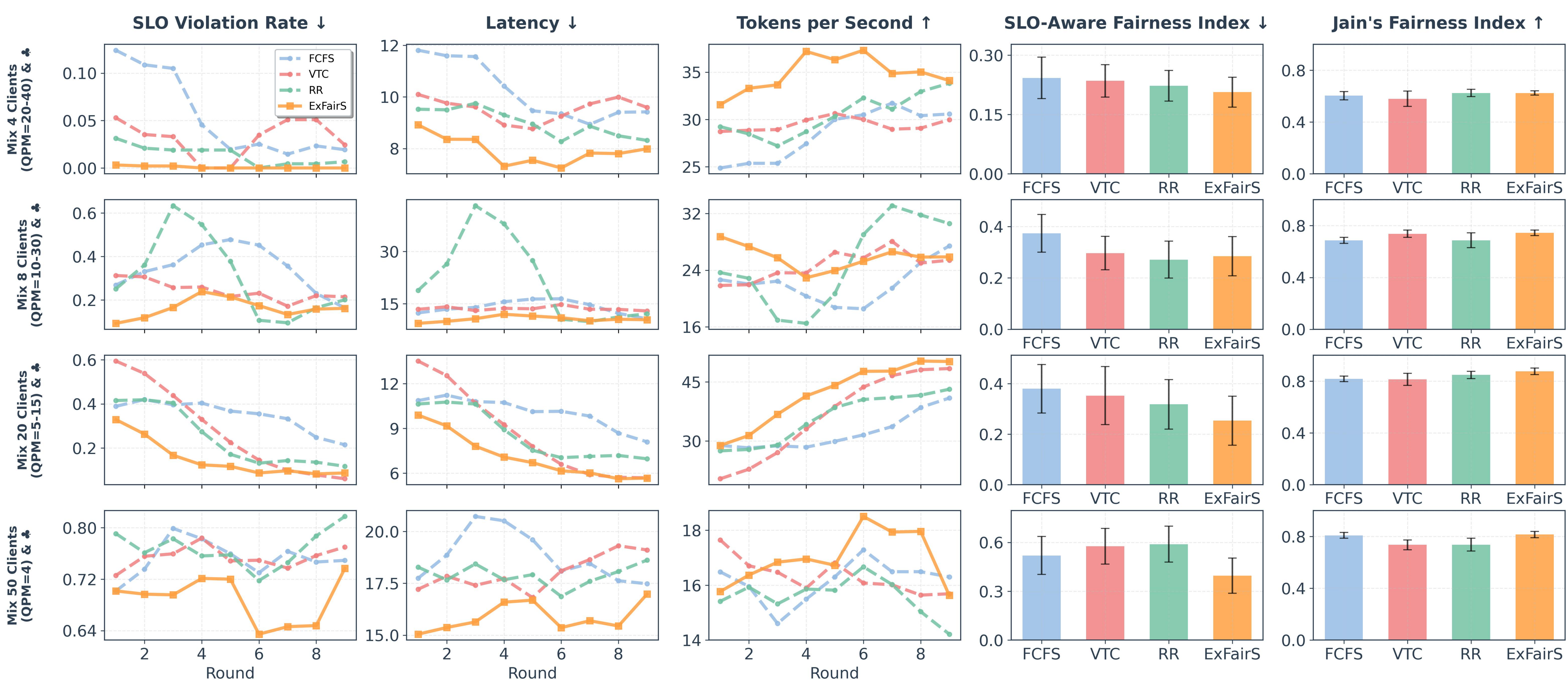
2. Method: ExFairS

$$\text{SLO-Aware Fairness Index}(\text{SAFI})_i = \alpha(\text{SLO}_i) + (1 - \alpha)(\text{ServiceUsage}_i)$$



- ◆ **Status Evaluator (Quantify):** Unifies **SLO Compliance** and **Resource Usage** into a single fairness score (SAFI).
- ◆ **Credit Exchange (Balance):** **"Happy"** clients yield resources to earn credits; **"Sad"** clients spend credits to gain priority.
- ◆ **Priority Queue (Schedule):** Uses **Proportional Insertion** to re-order requests, strictly guaranteeing **Starvation Avoidance**.

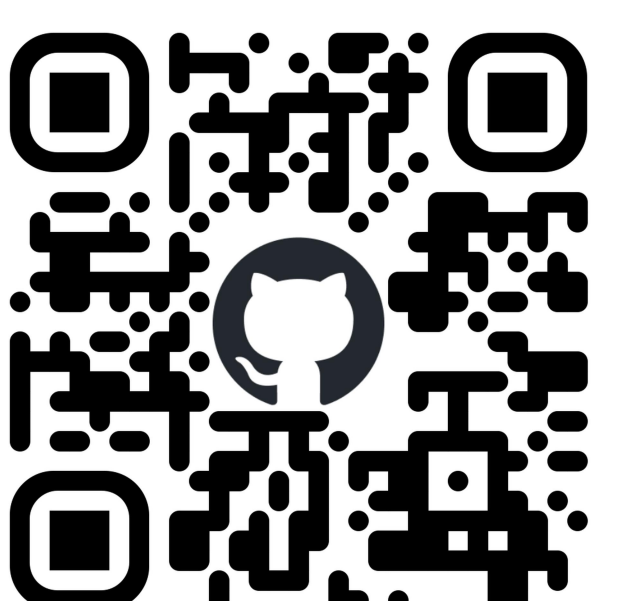
3. Main Results



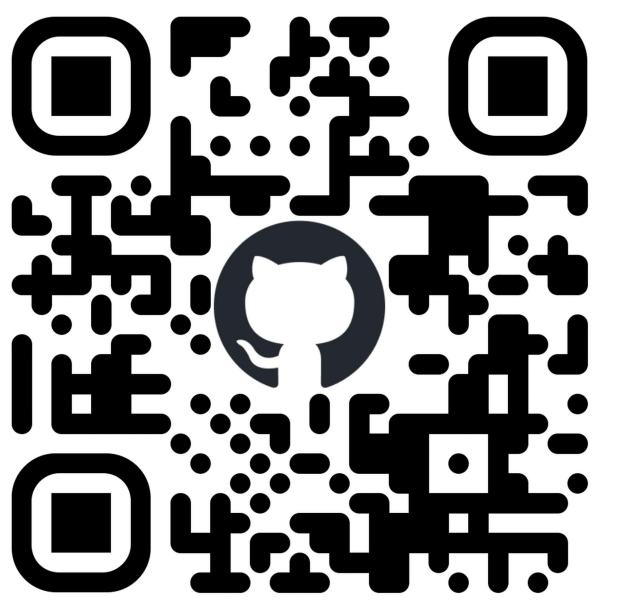
	<div><div></div><div>0.00.20.40.60.81.0</div><div>Performance Score (Darker = Better)</div></div>																																		
	S+L Clients (QPM=50) & ♣					S+L Clients (QPM=10-90) & ♠					Mix 4 Clients (QPM=20-40) & ♣					Mix 8 Clients (QPM=10-30) & ♣					Mix 20 Clients (QPM=5-15) & ♣					Mix 50 Clients (QPM=4) & ♣									
	SLO	Lat.	T/s	SAFI	Jain	SLO	Lat.	T/s	SAFI	Jain	SLO	Lat.	T/s	SAFI	Jain	SLO	Lat.	T/s	SAFI	Jain	SLO	Lat.	T/s	SAFI	Jain	SLO	Lat.	T/s	SAFI	Jain					
FCFS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.896	0.000	0.000	0.000	0.000	0.538	0.000	0.696	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.066	0.267	0.000	0.460	0.361	0.923
VTC	0.604	0.364	0.302	0.479	0.579	0.383	0.353	0.000	1.000	0.935	0.438	0.288	0.148	0.200	0.000	0.541	0.726	0.678	0.750	0.869	0.340	0.467	0.447	0.215	0.000	0.189	0.247	0.556	0.068	0.000	0.000	0.000	0.000	0.000	
RR	0.628	0.448	0.886	1.000	1.000	0.553	0.213	0.218	0.303	1.000	0.719	0.521	0.316	0.550	0.997	0.247	0.000	0.842	1.000	0.002	0.540	0.533	0.387	0.488	0.549	0.000	0.000	0.284	0.000	0.000	0.000	0.019	0.000		
ExFairS	1.000	1.000	1.000	0.636	0.334	1.000	1.000	1.000	0.816	0.048	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.865	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

4. Resources

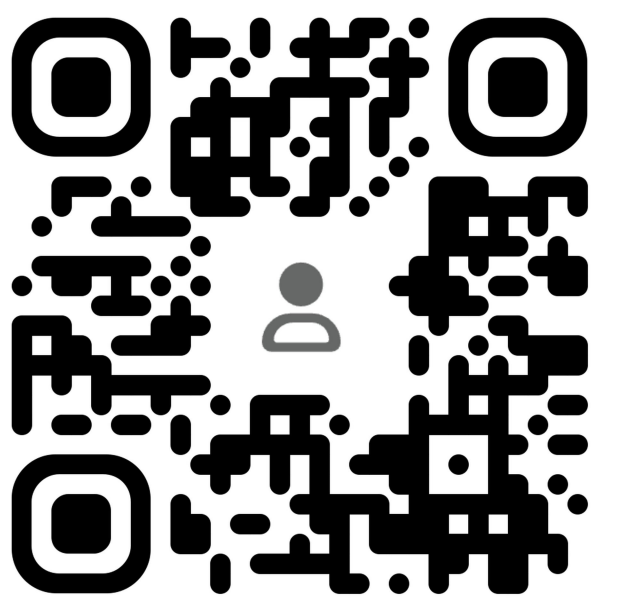
Code



Our Team



About



Wechat

