Designing Equitable Scheduling Systems

Sahana Rangarajan
Xuesi Chen, Pratyush Patel, Sara Mahdizadeh Shahri, Jaylen Wang
Akshitha Sriraman
Where do we see web services day to day?
Is latency a big deal?

Yes!

Result: Stringent latency constraint (300ms)

100 ms delay → 1% drop in sales

0.5 s delay → 20% drop in traffic
Where do we stand today?

Task prioritization by schedulers increasingly ML-based

Examples:
- Decima
- Quasar
- DeepRM
- Paragon
What’s a possible pitfall with ML?

Bias

Example: COMPAS Algorithm for Recidivism Risk

“High Risk” Score: 10

“Low Risk” Score: 3

Source: ProPublica
How’s this relevant to us?

Can I optimize performance by scheduling user 1’s task before user 2’s?

**User 1**
Younger, less patient
Strict latency requirement

**User 2**
Older, more patient
Relaxed latency requirement
How could this go wrong?

**Case study:** Varying perceptions of Wikipedia QoS

- **Higher user satisfaction:** Belarus
  - #90 global GDP (nominal)

- **Lower user satisfaction:** Germany
  - #4 global GDP (nominal)

Optimizing on user satisfaction is a slippery slope

**Source:** Analyzing Wikipedia Users’ Perceived Quality of Experience: A Large-Scale Study (Salutari et al.)
Proposal: A Bias-Free Scheduling Framework

Do different demographics exhibit different latency tolerances?

Can a scheduler capitalize on demographic differences?

Build a bias-free scheduling framework with under-the-hood real-time auditing
Designing Equitable Scheduling Systems

Sahana Rangarajan
Xuesi Chen, Pratyush Patel, Sara Mahdizadeh Shahri, Jaylen Wang
Akshitha Sriraman