Fairness-Aware Scheduling for Bursty HPC Workloads

Karaen Senthilkumar, **Suman Raj**, Maxime Gonthier

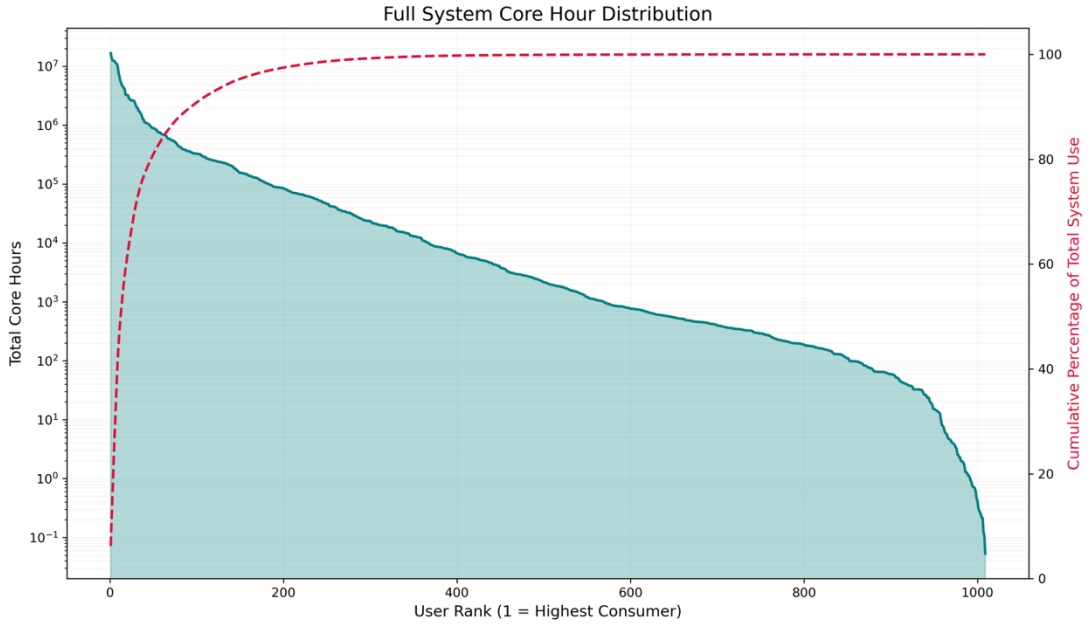
APEX SUMMIT
CHICAGO 2026

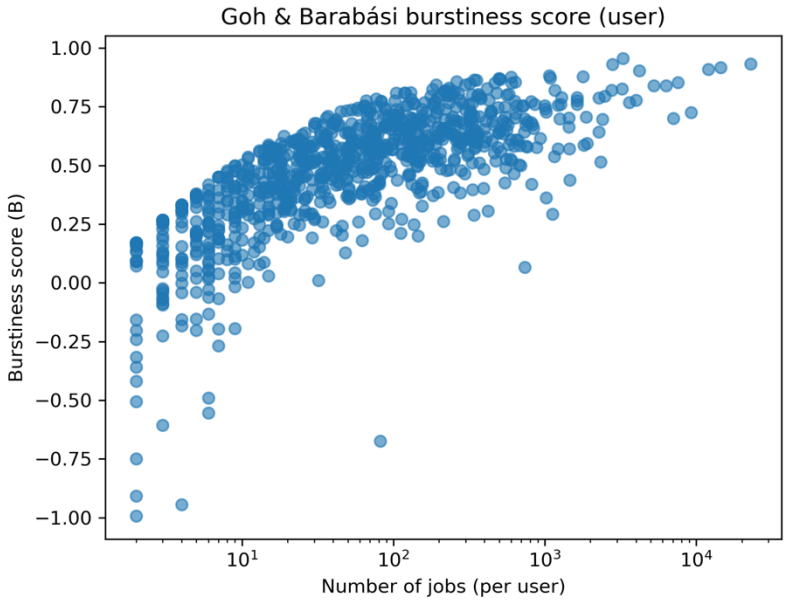
THE UNIVERSITY OF CHICAGO

Introduction

APEX SUMMIT
CHICAGO 2026

- The heaviest users of ALCF Polaris are also the burstiest.
- Roughly 20% of users accounting for over 95% of resources, and the burstiness score rises with submission volume.
- The same small group of users therefore monopolizes capacity.

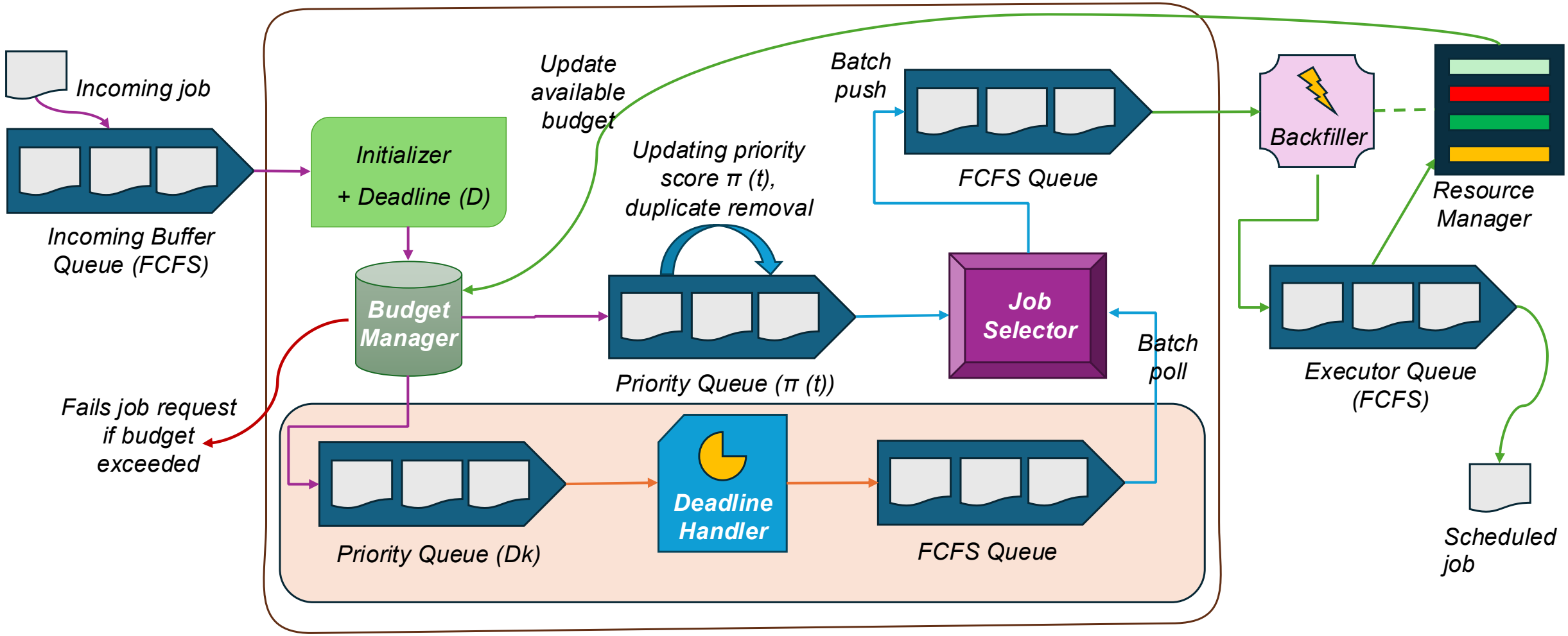




Gaps

- Existing HPC schedulers typically optimize one objective at a time:
  - Simple and efficient, but ignore project-level fairness and urgency (**FIFO / SJF / SDF**)
  - Favor important projects, but may starve lower-priority work (**Priority-based methods**)
  - Improve fairness, but can sacrifice latency or utilization (**WFQ / THEMIS**)
  - Flexible, but depends heavily on manual weighting and **does not directly enforce bounded waiting** (**SLURM-style multi factor scoring**)
- HyDRA is a hybrid deadline and ranking-aware scheduler to ensure **fairness**.
- Dynamic score balances:
  - Project / job priority, waiting age (starvation), resource demand
- Deadline mechanism provides a *bounded-wait / urgency signal* for jobs at risk of excessive delay.

HyDRA Scheduling Framework



Scheduler Formulation

APEX SUMMIT
CHICAGO 2026

- **Job:** $J_k = (a_k, R_k, n_k, \rho_k)$
  a_k = arrival time, R_k = runtime, n_k = nodes, ρ_k = priority
- **Service demand:**
  $q_k = R_k \times n_k$
- **Job age at time t:**
  $A_k(t) = t - a_k$
- **Dynamic priority score:**
  $\pi_k(t) = \rho_k \, (1 + \lambda A_k(t)) \, / \, q_k^{\gamma}$
- **Bounded waiting time:**
  $B_w(k) = \kappa \, q_k^{\gamma} \, / \, \rho_k$
- **Deadline-critical definition:**
  $D_k = a_k + B_w(k)$; $D_k - t \leq \delta$

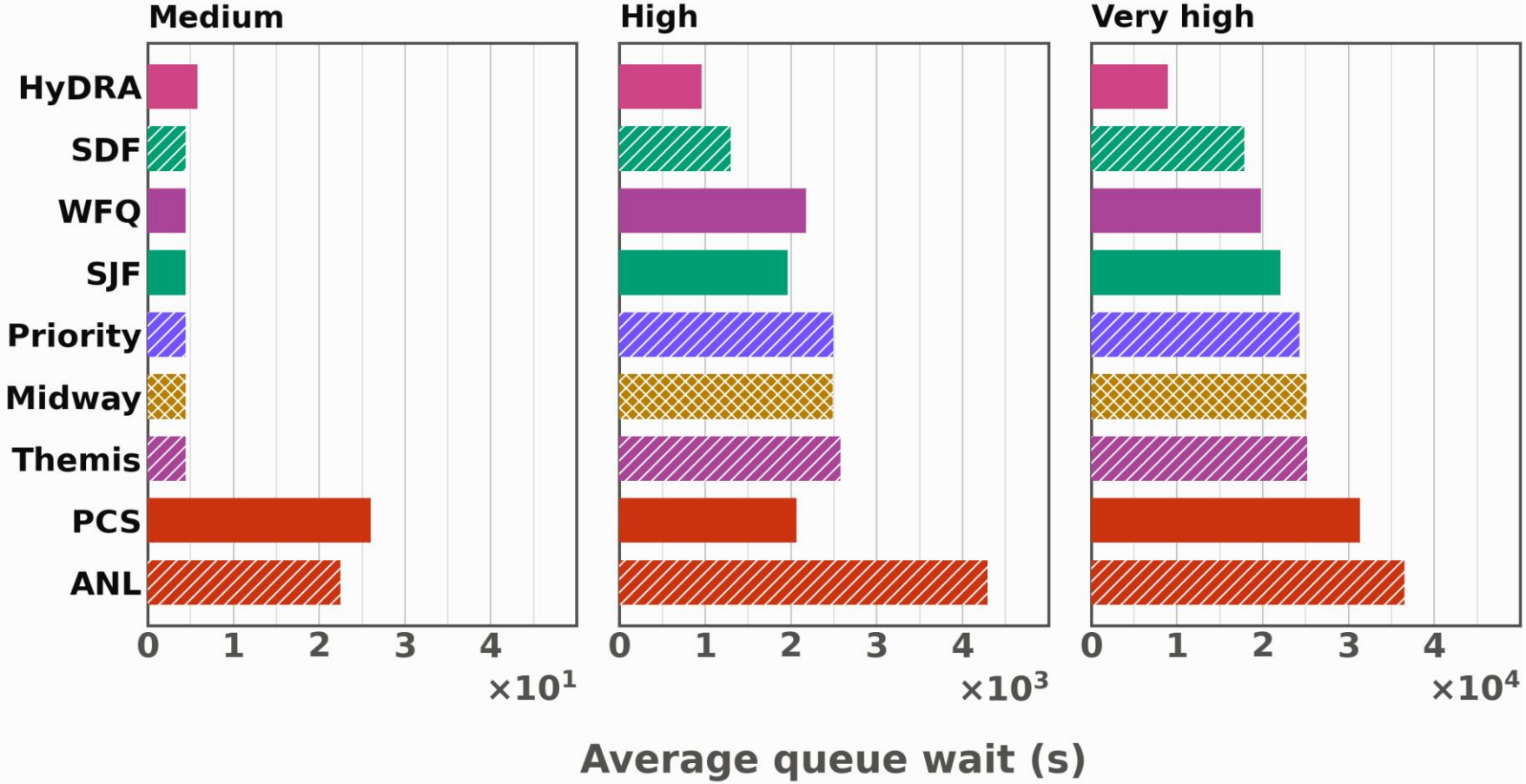
- **Hyperparameters**
  - Kappa – 10, Lambda - 1.0, Gamma - 0.75, Deadline-threshold - 30

Workloads

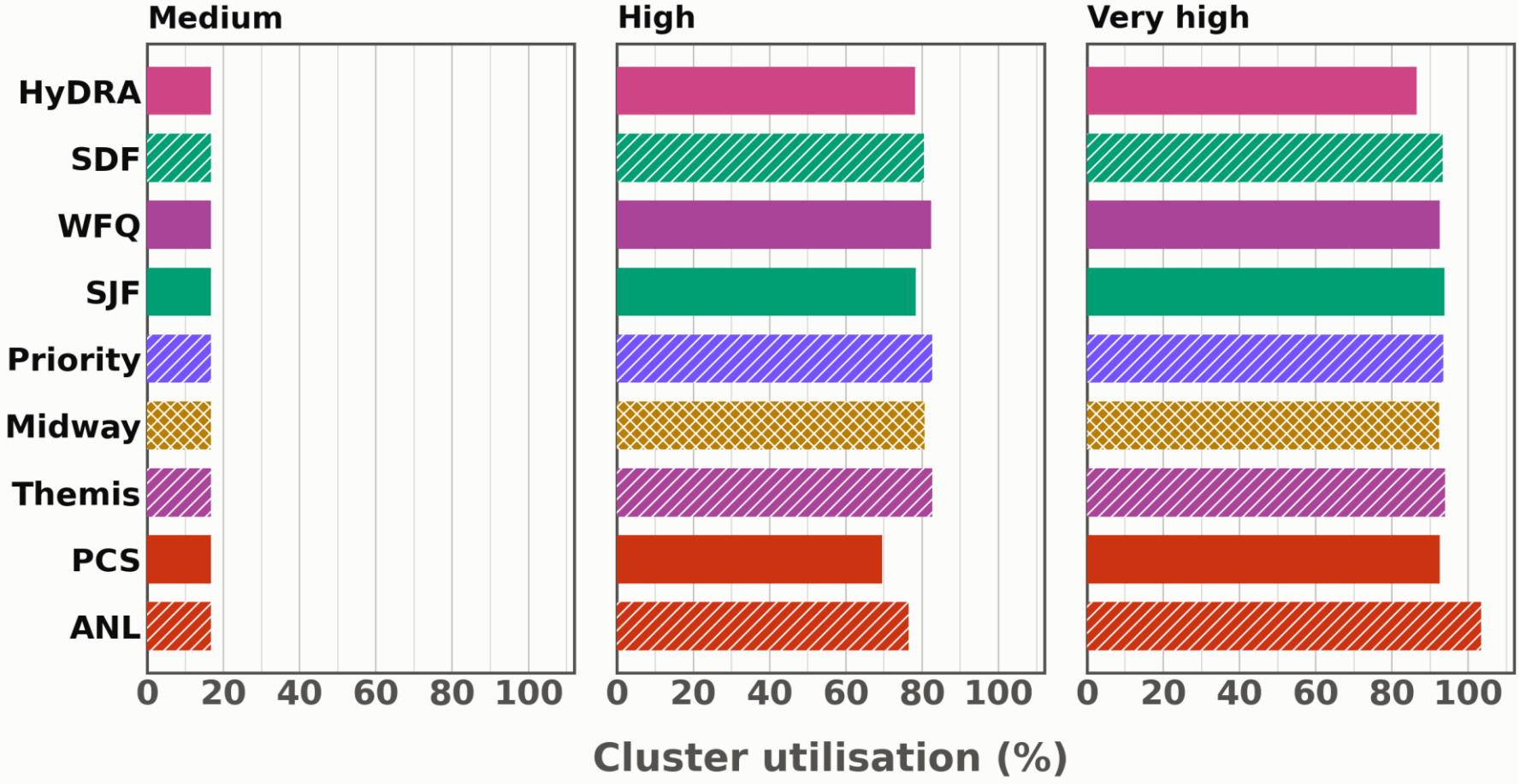
APEX SUMMIT
CHICAGO 2026

- Synthetic workload sampled from ALCF Polaris
  - https://reports.alcf.anl.gov/data/polaris.html#DIM_JOB_COMPOSITE
  - Actual period: 1 year (01/01/2025 - 12/31/2025)
  - Simulated period: 600 seconds
- 3 tiers of workloads were synthesized
  - Medium 30 — 30 project / 146 users / 275 jobs
  - High 100 — 100 project / 347 users / 635 jobs
  - Very High 1000 — 1000 project / 4045 users / 7367 jobs
- Project priority sampled from range 1 to 5
- Numbers of cores requested sampled from 1 to 35776

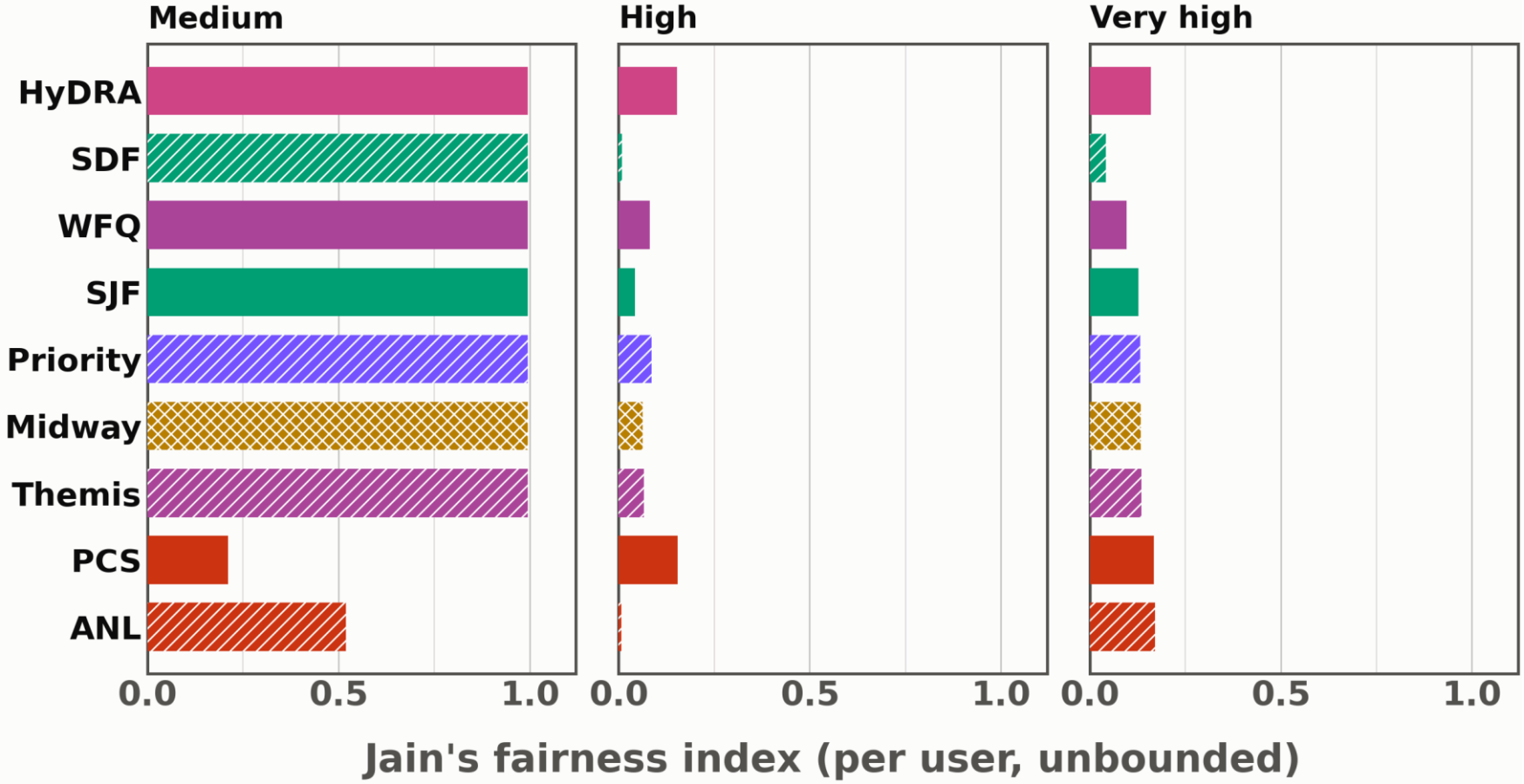
Results 1/3


Medium
High
Very high
HyDRA
SDF
WFQ
SJF
Priority
Midway
Themis
PCS
ANL
0 1 2 3 4
×10¹
0 1 2 3 4
×10³
0 1 2 3 4
×10⁴
Average queue wait (s)

Results 2/3


Medium
High
Very high
HyDRA
SDF
WFQ
SJF
Priority
Midway
Themis
PCS
ANL
0 20 40 60 80 100
0 20 40 60 80 100
0 20 40 60 80 100
Cluster utilisation (%)

Results 3/3



Higher is better.
J = 1 is perfectly fair.

Thank You!

Contact:
karaen@uchicago.edu
sumanraj@uchicago.edu