Summary of Research Accomplishments

Ziv Scully

Accomplishment 1: Unifying Theory of Single-Server Scheduling

Problem There is a huge design space of scheduling policies, even for single-server queueing systems. Different policies have different tradeoffs: mean latency, tail latency, simplicity of implementation, etc. *How can we rigorously evaluate different proposals for scheduling policies?*

My Contribution I developed a new technique, called "SOAP" (Schedule Ordered by Age-based Priority), which provides a *universal rigorous analysis* of a broad spectrum of scheduling policies [7].

Impact SOAP is a powerful new tool that can be used in a variety of ways:

- *Evaluating design tradeoffs*, e.g. for systems with scheduling constraints [6].
- *Proving guarantees on mean latency*, e.g. studying simple heuristics for scheduling in settings where job sizes (i.e. service/processing times) are unknown [8] or noisily estimated [4].
- *Proving guarantees on tail latency*, e.g. showing that policies that provably minimize mean latency can sometimes also have optimal tail latency for very high percentiles, particularly under heavy-tailed job size distributions [9, 10].

Awards My work on SOAP was a finalist for the 2018 INFORMS APS Best Student Paper Prize [7]. My talk on simple scheduling with unknown job sizes won the SIGMETRICS 2020 Best Video Award [8].

Accomplishment 2: First Analysis of Scheduling in Multiserver Systems

Problem Multiserver systems are ubiquitous, but there is very little queueing theory on scheduling in them. *How should we schedule in multiserver systems? Can we analyze them queueing theoretically?*

My Contribution We provided the *first latency bounds* for scheduling policies in multiserver systems, including Shortest Remaining Processing Time (SRPT) [1, 2] and the Gittins index policy [5].

Impact Our work opens up the field of scheduling in multiserver systems. Our findings so far suggest that policies that perform well in single-server systems transfer well to multiserver systems.

Awards Our work on multiserver SRPT won two Best Student Paper Awards: one at PERFORMACE 2018 [1] and one at SIGMETRICS 2019 [2]. The latter paper was also featured at STOC 2021's TheoryFest.

Accomplishment 3: Solving Several Open Problems in Fundamental Queueing Theory

Problem Several fundamental questions in single-server queueing are open, such as the following:

(a) Many scheduling policies have better mean latency than First-Come, First-Served (FCFS), but this sometimes comes at the cost of worse tail latency. *Can we improve upon FCFS's mean and tail latency simultaneously?*

(b) The Gittins index yields the scheduling policy that minimizes mean weighted latency. However, the Gittins index assumes job weights are static and known to the scheduler. *How should we schedule with variable or unknown weights?*

(c) The busy period length for a single-server queue (i.e. the amount of time between a job arriving to an empty system and the system becoming empty again) is a fundamental concept in queueing theory that has numerous applications. Simple formulas are known for integer moments of the busy period, but not fractional moments. *What are the fractional moments of busy periods?*

My Contribution I have contributed to answering *all three questions above.*

(a) We proposed a new scheduling policy and proved it has *strictly better latency distribution* than FCFS, improving the mean and all percentiles of latency [3].

(b) I *generalized the Gittins index* to work with variable and unknown job weights [5].

(c) We derived a *simple formula for bounding fractional moments* of busy periods [10].

Awards Our work on improving upon FCFS won the SIGMETRICS 2021 Best Paper Award [3]. My work on the Gittins index for variable and unknown weights was an invited paper at WiOpt 2021 [5].

References

[1] Isaac Grosof, Ziv Scully, and Mor Harchol-Balter. 2018. SRPT for Multiserver Systems. *Perform. Eval.* 127–128 (Nov. 2018), 154–175. Winner of **PERFORMANCE 2018 Best Student Paper Award**.

[2] Isaac Grosof, Ziv Scully, and Mor Harchol-Balter. 2019. Load Balancing Guardrails: Keeping Your Heavy Traffic on the Road to Low Response Times. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 2, Article 42 (June 2019), 31 pages. Winner of **SIGMETRICS 2019 Best Student Paper Award** and featured as a **STOC 2021 TheoryFest Mini-Plenary**.

[3] Isaac Grosof, Kunhe Yang, Ziv Scully, and Mor Harchol-Balter. 2021. Nudge: Stochastically Improving upon FCFS. *Proc. ACM Meas. Anal. Comput. Syst.* 5, 2, Article 21 (June 2021), 29 pages. Winner of **SIGMETRICS 2021 Best Paper Award**.

[4] Ziv Scully, Isaac Grosof, and Michael Mitzenmacher. 2022. Uniform Bounds for Scheduling with Job Size Estimates. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022) (Leibniz International Proceedings in Informatics (LIPIcs))*. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Berkeley, CA, Article 41, 30 pages.

[5] Ziv Scully and Mor Harchol-Balter. 2021. The Gittins Policy in the M/G/1 Queue. In *19th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2021)*. IEEE, Philadelphia, PA, 8 pages.

[6] Ziv Scully and Mor Harchol-Balter. 2021. How to Schedule Near-Optimally under Real-World Constraints. *arXiv* (Oct. 2021), 25 pages. arXiv:1805.06865 In submission.

[7] Ziv Scully, Mor Harchol-Balter, and Alan Scheller-Wolf. 2018. SOAP: One Clean Analysis of All Age-Based Scheduling Policies. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1, Article 16 (April 2018), 30 pages. Finalist of **2019 INFORMS APS Best Student Paper Prize**.

[8] Ziv Scully, Mor Harchol-Balter, and Alan Scheller-Wolf. 2020. Simple Near-Optimal Scheduling for the M/G/1. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 1, Article 11 (May 2020), 29 pages. Winner of **SIGMETRICS 2020 Best Video Award**.

[9] Ziv Scully and Lucas van Kreveld. 2021. When Does the Gittins Policy Have Asymptotically Optimal Response Time Tail? *ACM SIGMETRICS Performance Evaluation Review* (2021). To appear. Full version in submission. arXiv:2110.06326.

[10] Ziv Scully, Lucas van Kreveld, Onno J. Boxma, Jan-Pieter Dorsman, and Adam Wierman. 2020. Characterizing Policies with Optimal Response Time Tails under Heavy-Tailed Job Sizes. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 2, Article 30 (June 2020), 33 pages.