SRPT for Multi Server Systems Under Cellular Batching

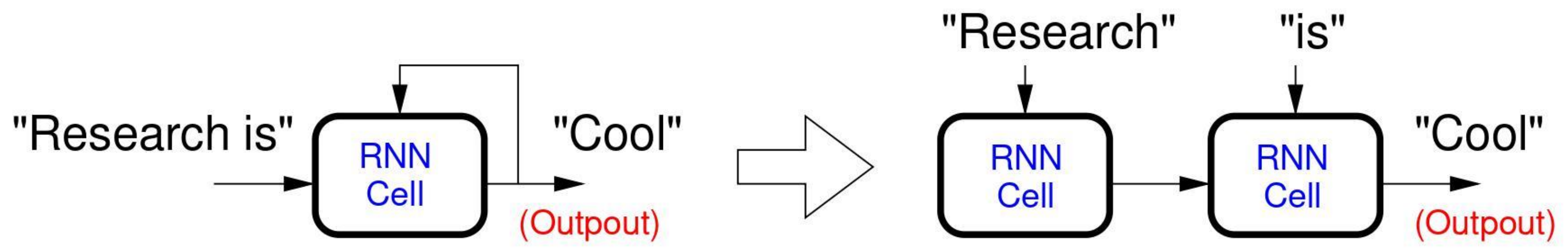
Author: John Wahlig
Advisor: Jia (Kevin) Liu

Research Motivation and Objectives

- Massive parallelism allows for expedition of training and inferences phases of deep learning systems
- Design efficient scheduling algorithms for RNN inference jobs to minimize average response time

Question: **Does Shortest Remaining Processing Time (SRPT) minimize overall job response time for Recursive Neural Network (RNN)?**

What is an RNN?



- Output of a cell is used as input for proceeding cell
- Used commonly for sequential data
- Weights are shared among cells

System Model

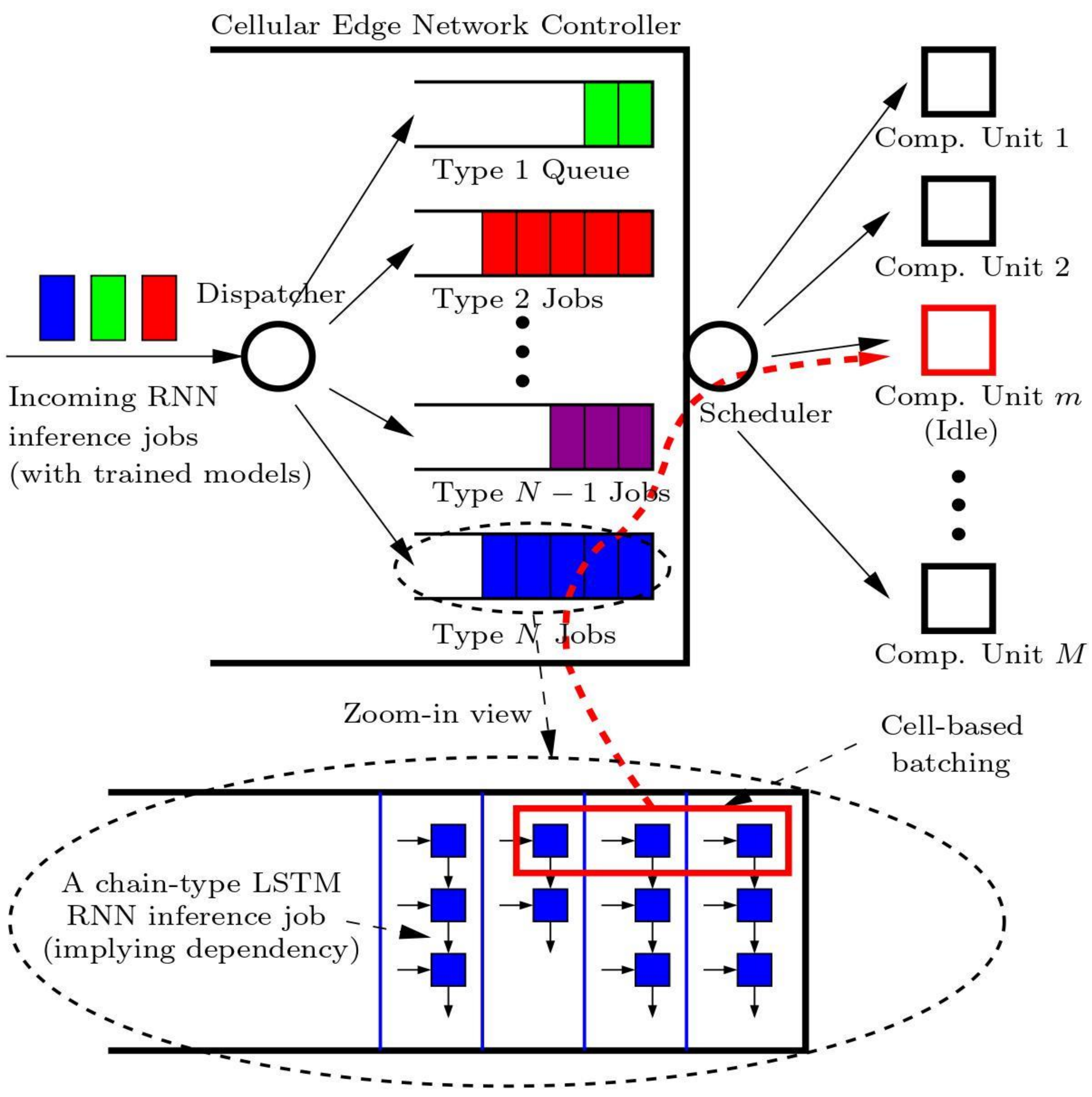


Figure 1. RNN Job Scheduler and Chain-like Job Structure

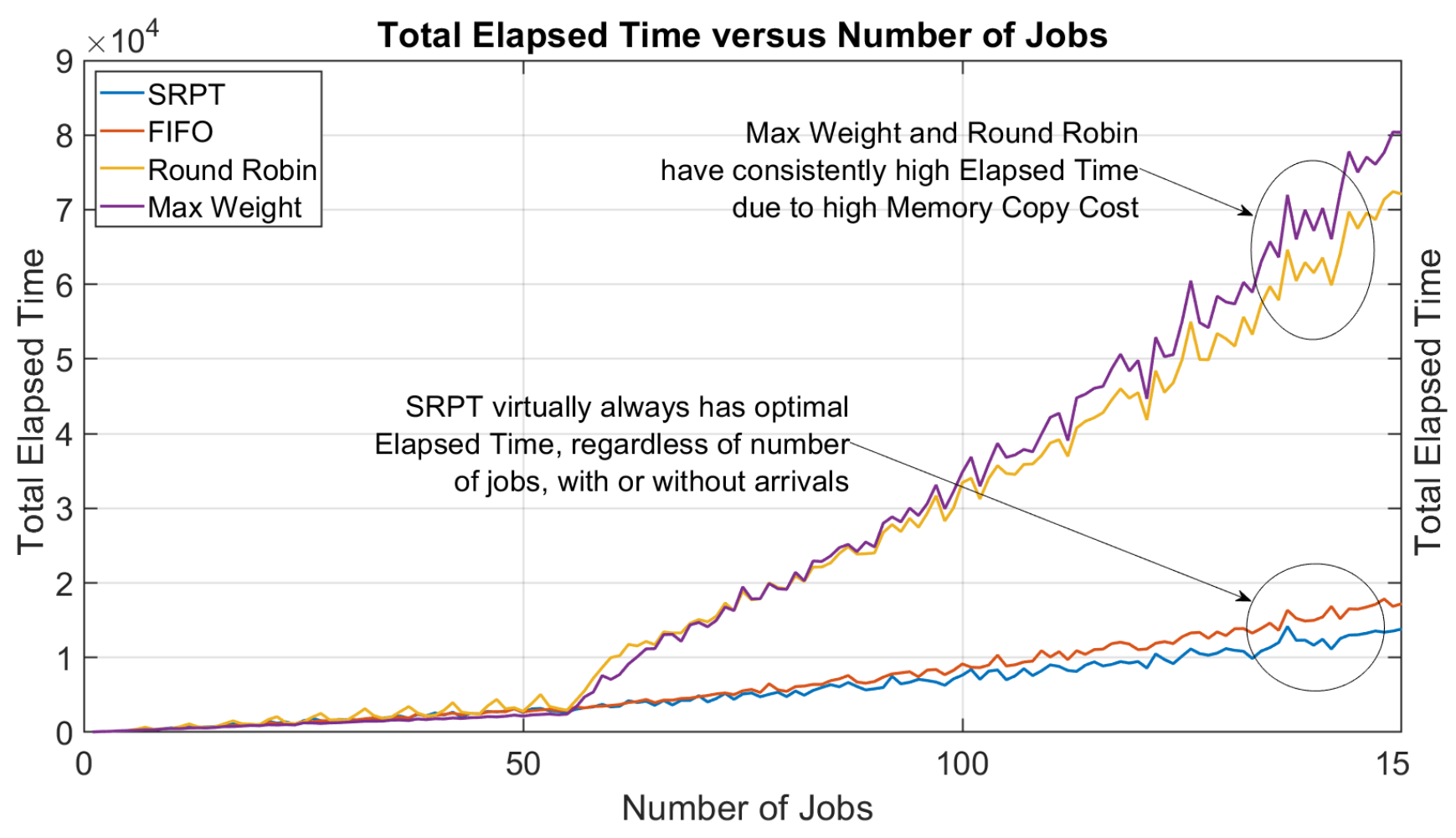
Methodology

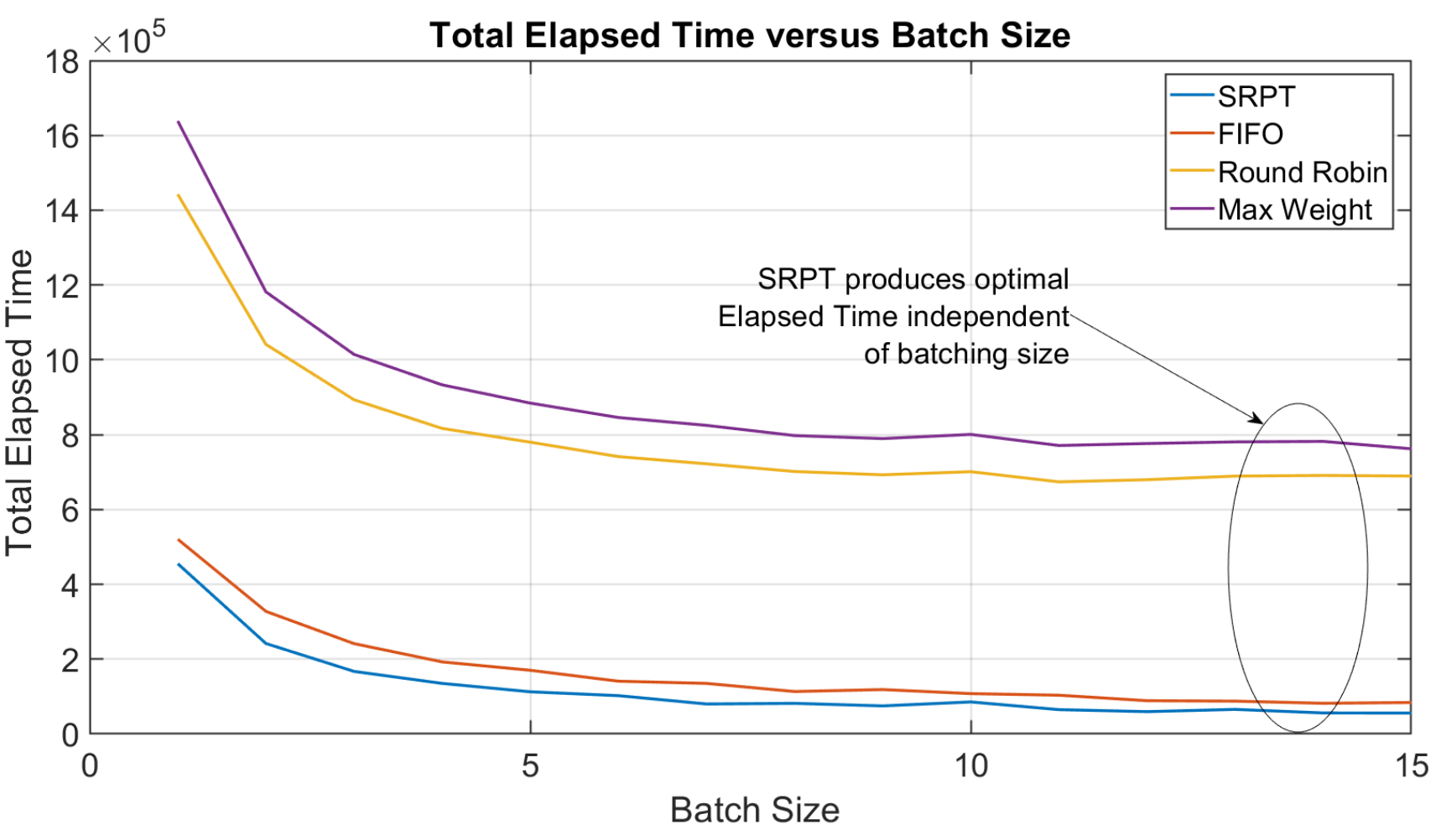
- Open problem even in single server system with batching
- Study multi-server SRPT with batching
- Conjure on M-SRPT-B's optimality

Simulation Setting

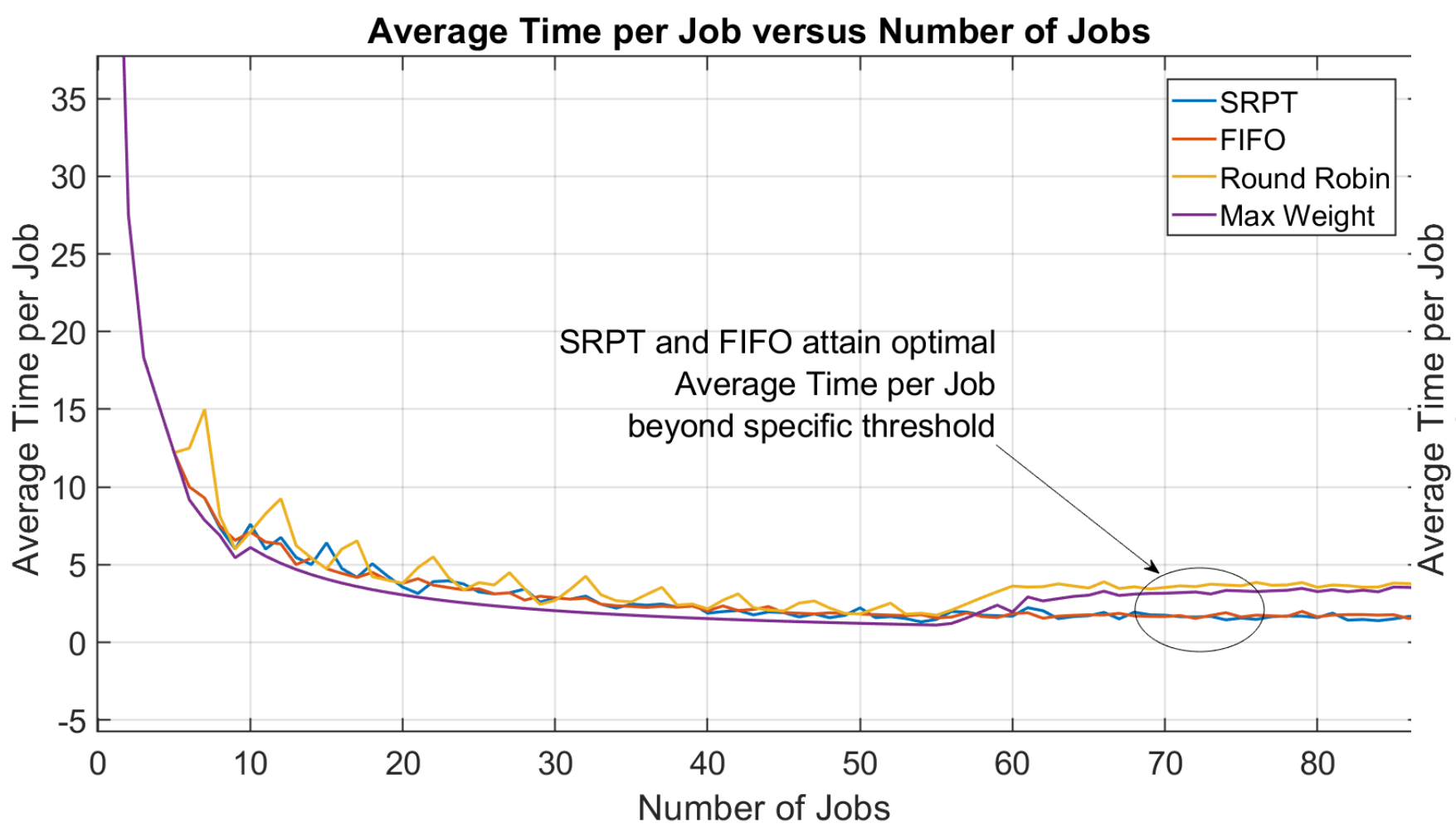
- Baseline policies:
  - SRPT, FIFO, Round Robin, Max Weight
- Arrival patterns:
  - Poisson Arrival, Bernoulli Arrival
- Multiple servers with various processing capability
- Chain-like jobs (single type/queue)
- Memory copy considered when switch among servers
- Heavy traffic vs light traffic

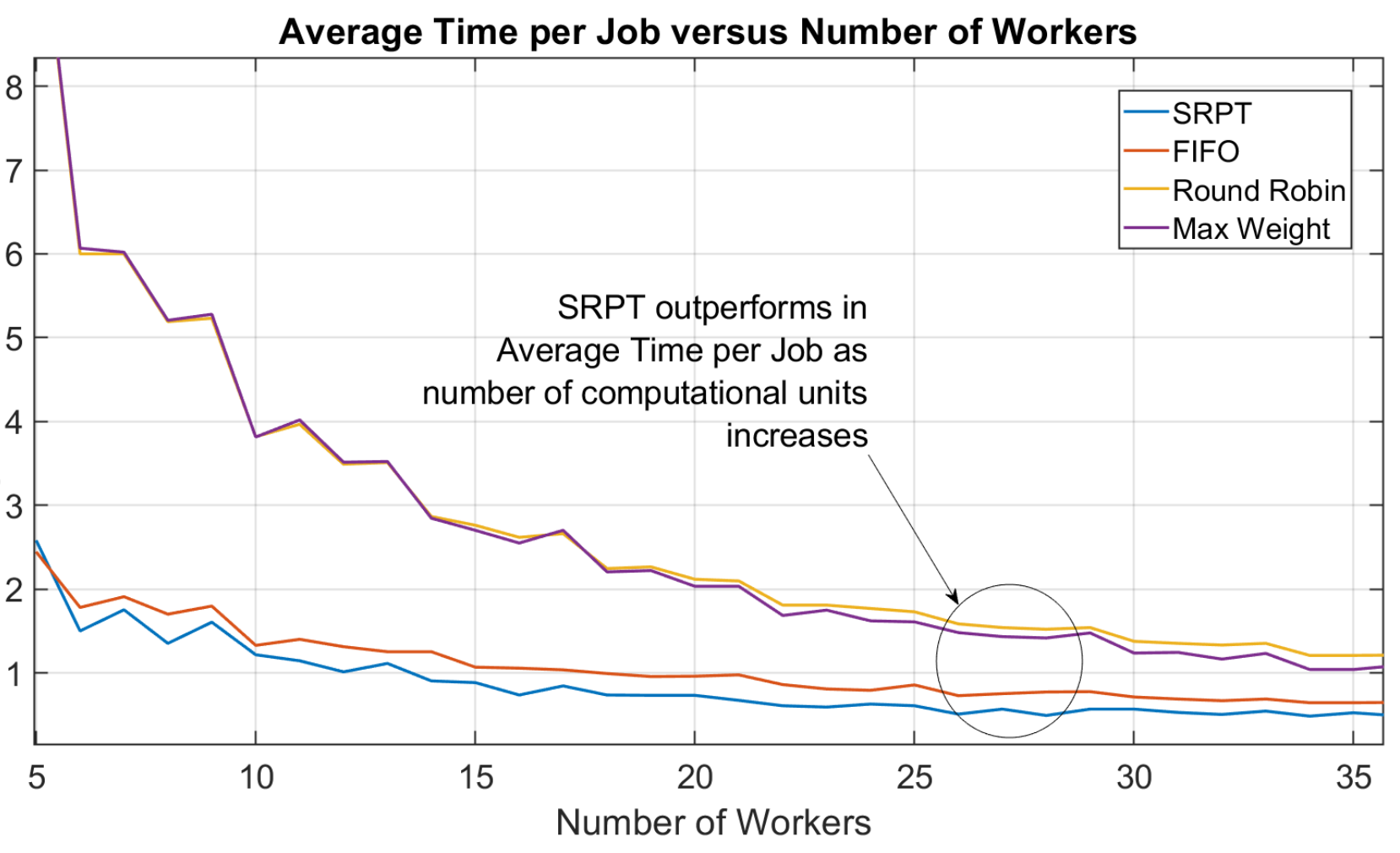
Simulation Results





- SRPT outperforms other policies in Total Elapsed Time





- Under heavy traffic, SRPT outperforms other policies in Average Time per Job
- Largest difference between Round Robin and Max Weight versus SRPT and FIFO is Memory Copy Cost Time

Conclusions

- SRPT and FIFO perform comparably, but SRPT remains almost entirely optimal in Total Elapsed Time
- SRPT ensures minimal Memory Copy Cost Time
- SRPT only produces sub-optimal results under specific conditions:
  - Exceptionally large number of Computational Units
  - Specifically sized arriving jobs

Future Work

- Implement scheduling algorithms into actual RNN framework
- Extend to tree-like job structure
- Multiple queues to support multiple cell types
- Prove optimality of SRPT theoretically under no arrivals

[1]. Pin Gao, Lingfan Yu, Yongwei Wu, and Jinyang Li. Low latency RNN inference with cellular batching. In Proceedings of the Thirteenth EuroSys Conference, EuroSys 2018, Porto, Portugal, April 23-26, 2018, pages 31:1--31:15, 2018.
[2]. Weina Wang, Mor Harchol-Balter, Haotian Jiang, Alan Scheller-Wolf, R. Srikant. ``Delay Asymptotics and Bounds for Multi-Task Parallel Jobs." Queueing Systems , vol. 91, no. 3-4, March 2019, pp. 207—239.
[3]. Grosof, I., Z. Scully, and M. Harchol-Balter (2018). Srpt for multiserver systems. arXiv preprint arXiv:1805.07686.