Stochastic Scheduling with Predictions

Isaac Grosof

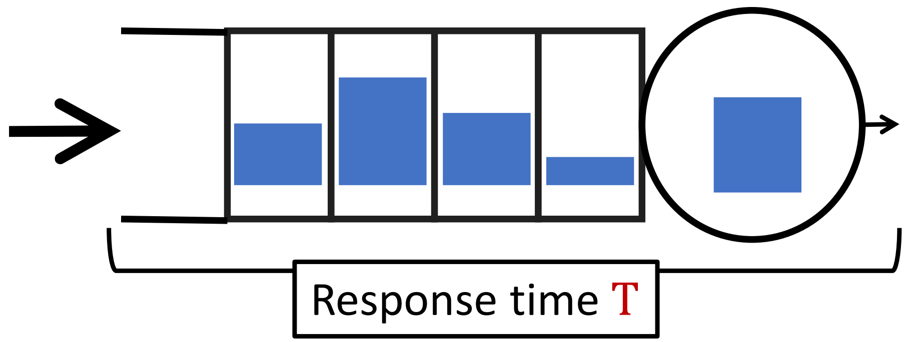
CMU Computer Science

Ziv Scully (CMU -> Cornell)

Michael Mitzenmacher (Harvard)

Scheduling with Predictions

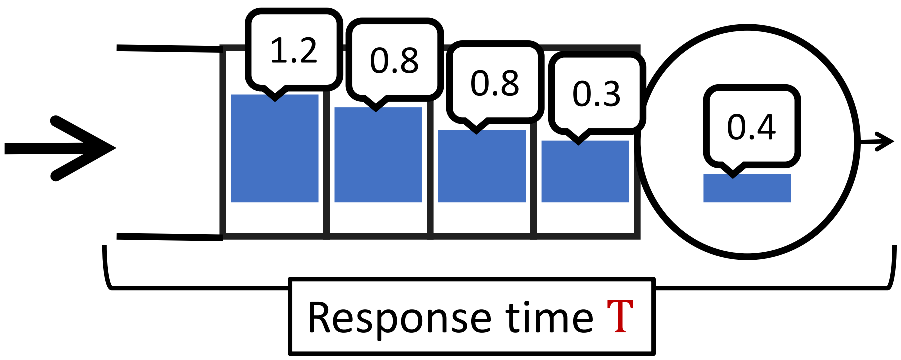
"Uniform Bounds for Scheduling with Job Size Estimates" ITCS 2022



Goal: Minimize mean response time (E[T])

Scheduling with Predictions

“Uniform Bounds for Scheduling with Job Size Estimates” ITCS 2022



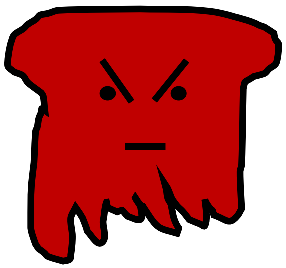
Goal: Minimize mean response time (E[T])

Known sizes: Shortest Remaining Processing Time (SRPT) is optimal

Predicted sizes: ?

Stochastic Scheduling with Predictions

Two ways to study:



Worst Case Analysis

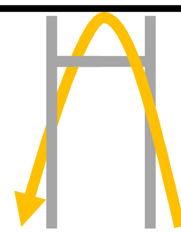
Stochastic Analysis

This talk

Why stochastic analysis for scheduling?

Reflects real world

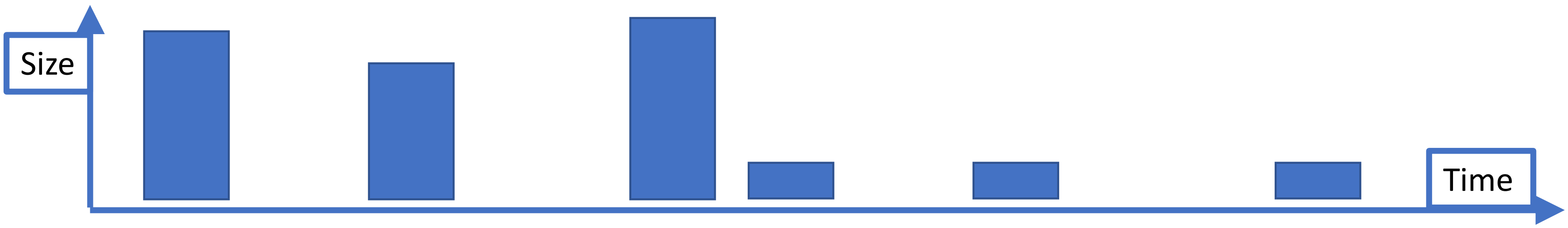
Overcomes worst-case barriers



Challenging, subtle problems

Why Stochastic Arrivals?

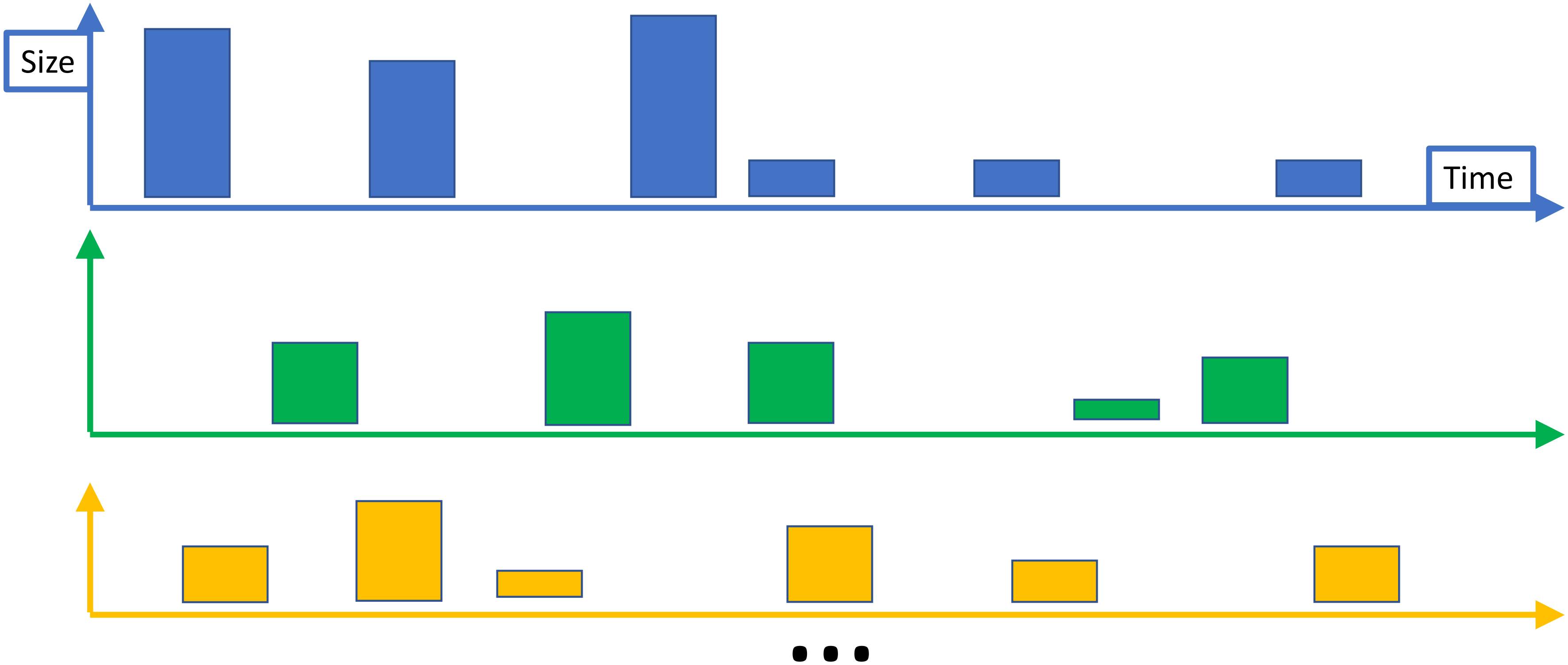
Worst case setting: General arrival sequence



Why Stochastic Arrivals?

Worst case setting: General arrival sequence

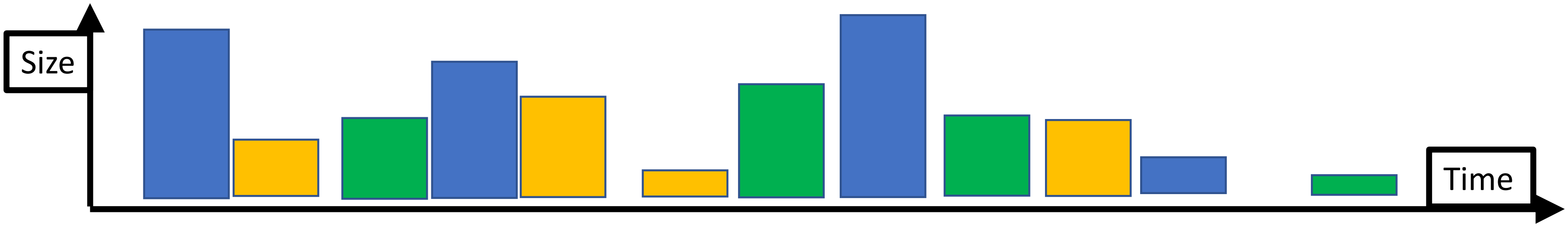
Real world: lots of independent arrival sequences



Why Stochastic Arrivals?

Worst case setting: General arrival sequence

Real world: lots of independent arrival sequences

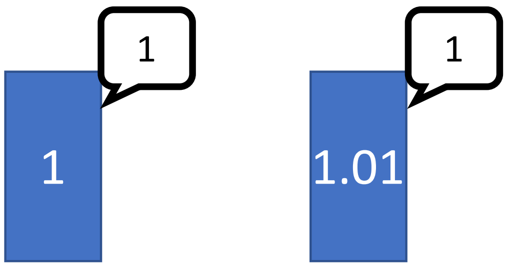


Approximation for large systems:

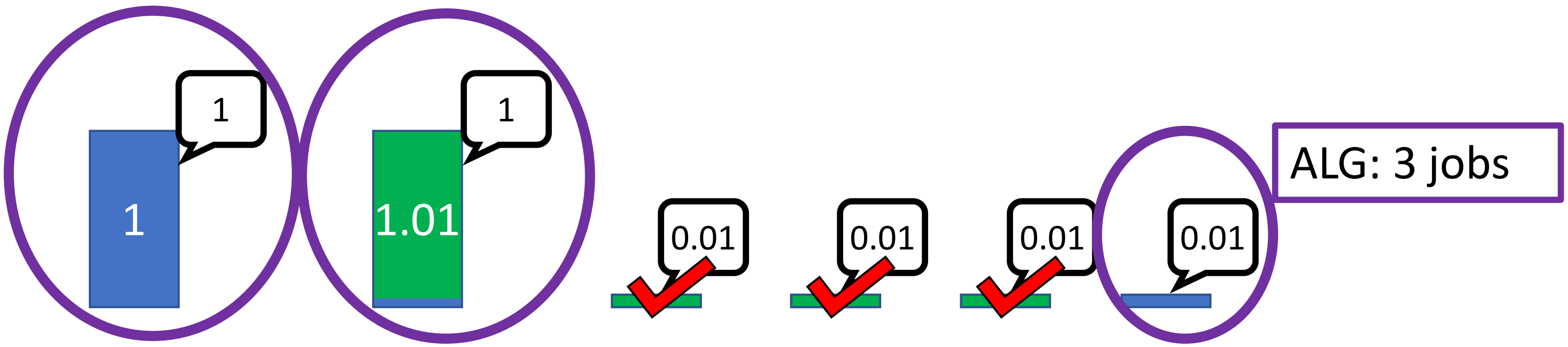
Exponential interarrival times (Poisson)

I.i.d. sizes

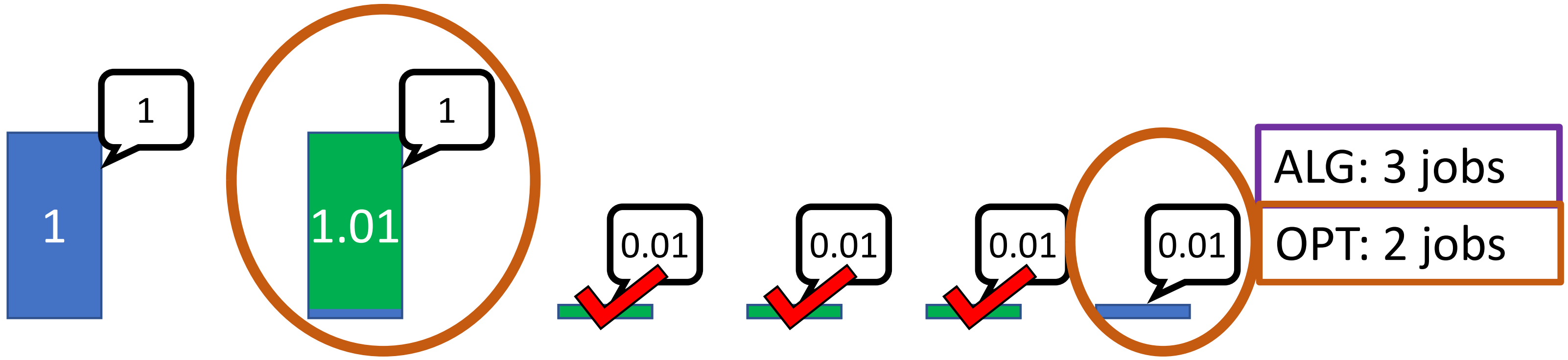
Example of Scheduling with Predictions



Example of Scheduling with Predictions



Example of Scheduling with Predictions

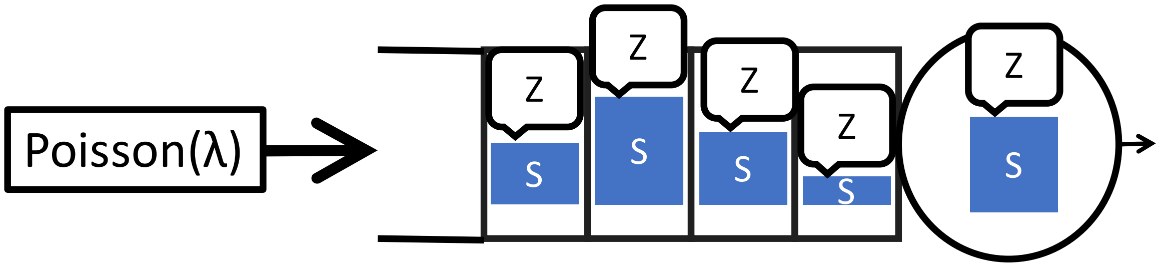


Tiny prediction error → 3/2 competitive ratio

Very unlikely in our stochastic model

Goal: Tiny prediction error → Near-perfect performance

Specific model



Arrival process: Poisson(λ)

Size and predicted size i.i.d. from (S, Z)

$Z \in [\beta S, \alpha S]$

(S,Z) distribution chosen adversarially

Scheduler does not know S, Z, α, β

Metric: $\frac{E[T^{\pi}]}{E[T^{OPT}]} = \frac{E[T^{\pi}]}{E[T^{SRPT}]}$

Performance Goals

Goals for $\frac{E[T^{\pi}]}{E[T^{SRPT}]}$:

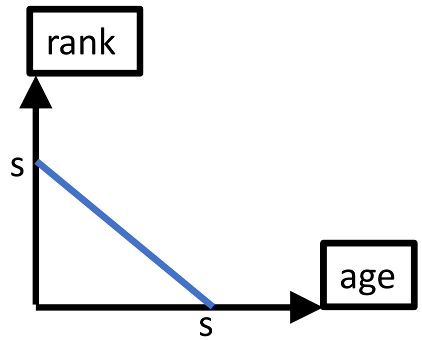
Consistency: $\lim_{\alpha,\beta\to 1} \frac{E[T^{\pi}]}{E[T^{SRPT}]} = 1$

Graceful Degradation: $\forall\alpha,\beta,\ \frac{E[T^{\pi}]}{E[T^{SRPT}]} \leq c\,\frac{\alpha}{\beta}$

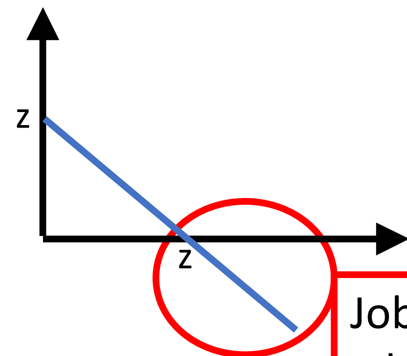
Robustness: $\forall\alpha,\beta,\ \frac{E[T^{\pi}]}{E[T^{SRPT}]} \leq d$

Naïve Scheduling Policy

SRPT:
Rank $= s - a$
Lower is better



SRPT-E: $z - a$



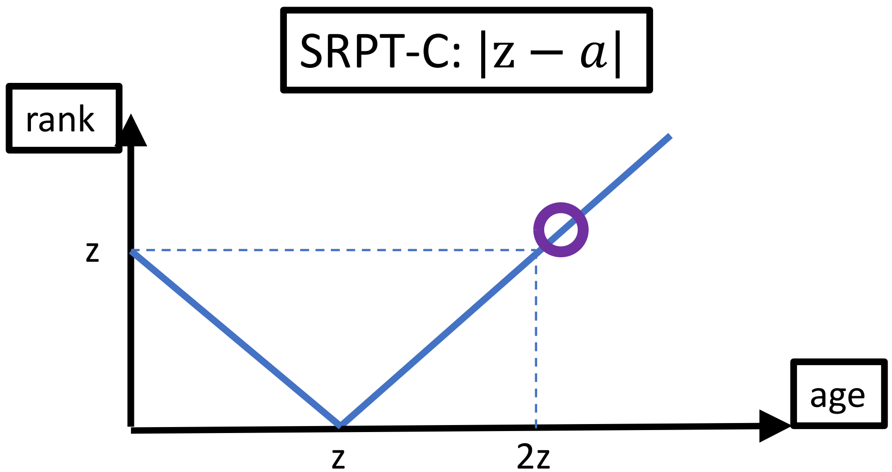
Jobs uninterrupted
when exceed prediction

Lower bound: $E[T^{SRPT-}\ \ ] \geq \frac{\lambda}{2}(1-\beta)^2 E[S^2]$

If $E[S^2] = \infty, E[T] = \infty$.

In datacenter and webserver traces, $E[S^2]$ very large

SRPT with Checkmark



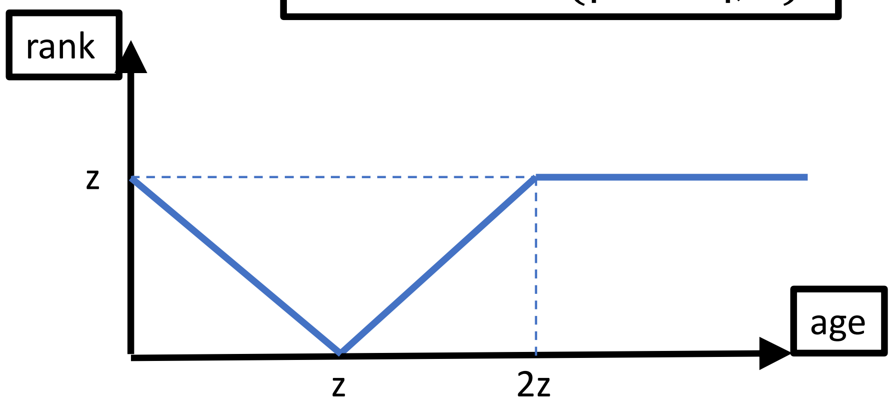
SRPT-C is consistent: $\lim_{\alpha,\beta \to 1} \frac{E[T^{SRPT-}]}{E[T^{SRPT}]} = 1$

First consistent policy!

Not gracefully degrading: if $\beta < \frac{1}{2}$, $\frac{E[T^{SRPT-}]}{E[T^{SRPT}]}$ unbounded.

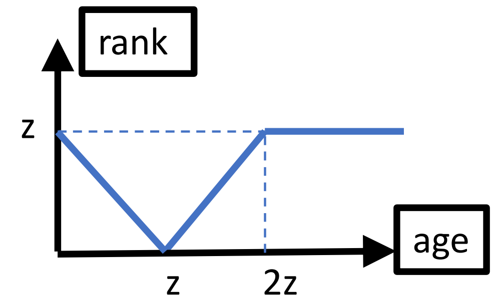
SRPT with Bounce

$\boxed{\text{SRPT-B: } \min(|z - a|, z)}$



First policy to be consistent and gracefully degrading!

Results



$$\frac{E[T^{SRPT}]}{E[T^{SRPT}]} \leq \frac{\alpha}{\beta} K(\alpha, \beta)$$

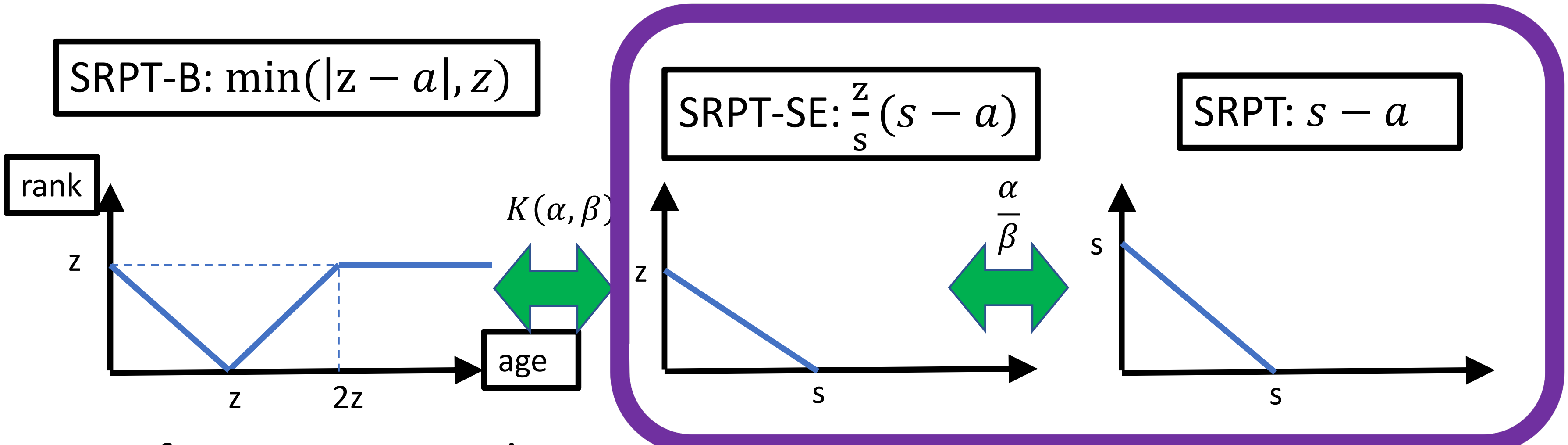
where $K(\alpha,\beta) = 1 + \left(\frac{3}{2}\alpha 1\{\beta < 1\} + 1\right) \min\left\{1, \max\left\{1 - \frac{1}{\alpha}, \frac{1}{\beta} - 1\right\}\right\}$

$\lim_{\alpha,\beta \to 1} K(\alpha,\beta) = 1$ (Consistency)

$\forall \alpha, \beta,\ K(\alpha,\beta) \leq 3.5$ (Graceful degradation)

Robustness impossible, Gittins policy lower bound

Proof Methods



Proof uses cutting-edge queueing theory:

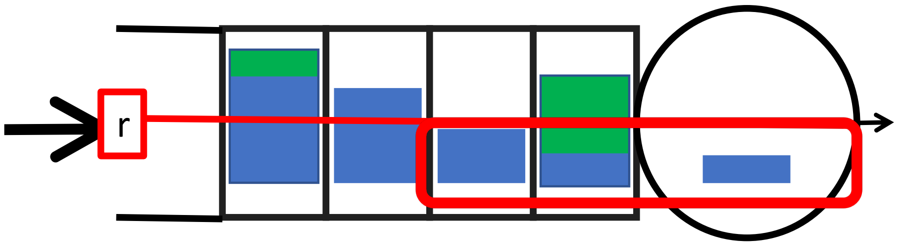
SOAP

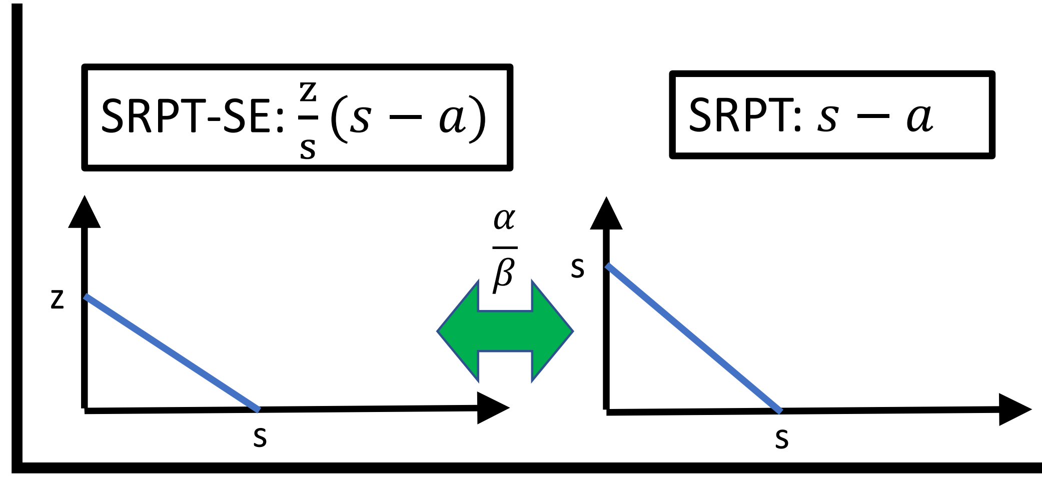
Analysis of E[T] for
rank-function policies

WINE

$$\lambda E[T] = E[N] = \int_0^\infty \frac{E[W(r)]}{r^2} dr$$

SRPT-SE vs. SRPT





r-relevant work: $W(r) = \sum (s_i - a) 1\{s_i - a \le r\}$

Thm: SRPT minimizes $W(r)$: $\mathrm{E}\left[\mathrm{W}^{\mathrm{SRPT}}(\mathrm{r})\right] \le \mathrm{E}[\mathrm{W}^{\pi}(\mathrm{r})]\ \forall \pi, r$

Thm: SRPT-SE nearly min. $W(r)$: $\mathrm{E}\left[\mathrm{W}^{\mathrm{SRPT-SE}}(\mathrm{r})\right] \le \mathrm{E}\left[\mathrm{W}^{\pi}\left(\frac{\alpha}{\beta}\mathrm{r}\right)\right]\ \forall \pi, r$

$\mathrm{E}\left[\mathrm{W}^{\mathrm{SRPT-SE}}(\mathrm{r})\right] \le \mathrm{E}\left[\mathrm{W}^{SRPT}\left(\frac{\alpha}{\beta}\mathrm{r}\right)\right]$

$E[T^{SRPT-S}] \le \frac{\alpha}{\beta} E[T^{SRPT}]$

$\boxed{\text{WINE: } E[T] = \frac{1}{\lambda}\int_0^{\infty} \frac{E[W(r)]}{r^2} dr}$

Zooming Out on Scheduling with Predictions

Today: Unknown prediction distribution (S, Z)

Known prediction distribution:

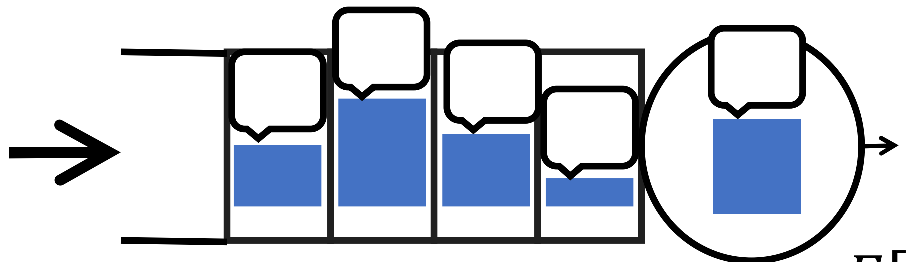
Single server solved by Gittins policy

Multiserver: “The Gittins Policy is Nearly Optimal in the M/G/k under Extremely General Conditions”, [G., Scully, Harchol-Balter], SIGMETRICS 2021

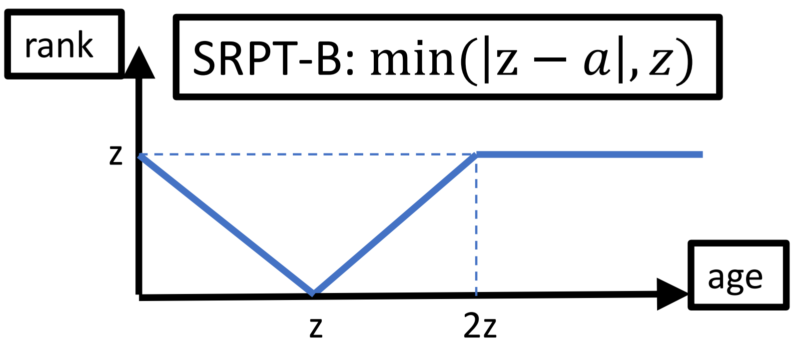
Strategic predictions:

“Incentive Compatible Queues Without Money”, [G., Mitzenmacher], arXiv

Stochastic Scheduling with Predictions



igrosof@cmu.edu
isaacg1.github.io



$$\lim_{\alpha,\beta\to 1} \frac{E[T^{SRPT}]}{E[T^{SRPT}]} = 1$$

$$\frac{E[T^{SRPT-}]}{E[T^{SRPT}]} \leq 3.5\frac{\alpha}{\beta}$$