Optimal Multiserver Scheduling

with Unknown Job Sizes in Heavy Traffic

Ziv Scully
Isaac Grosof
Mor Harchol-Balter

Carnegie Mellon University
Optimal Multiserver Scheduling
with Unknown Job Sizes in Heavy Traffic

Ziv Scully
Isaac Grosof
Mor Harchol-Balter

Carnegie Mellon University
M/G/1 Queue
M/G/1 Queue
M/G/1 Queue
M/G/1 Queue

queue

server

job

size
M/G/1 Queue
M/G/1 Queue

queue

server

job

size
M/G/1 Queue

queue

server

job

size
M/G/1 Queue

queue

server

job

size

remaining size

age
M/G/1 Queue

random arrivals

size

remaining size

age

queue

server

job
M/G/1 Queue

random arrivals

queue

size

remaining size

age

server

job
M/G/1 Queue

\[ X = \text{size distribution} \]
\[ \lambda = \text{arrival rate} \]
\[ \rho = \lambda E[X] < 1 \]

random arrivals

queue

server

job

size

remaining size

age
M/G/1 Queue

\[ X = \text{size distribution} \]
\[ \lambda = \text{arrival rate} \]
\[ \rho = \lambda E[X] < 1 \]

Scheduling policy: picks which job to serve
M/G/1 Queue

$X = \text{size distribution}$
$\lambda = \text{arrival rate}$
$\rho = \lambda E[X] < 1$

random arrivals

Scheduling policy: picks which job to serve

queue

server

job

remaining size

size

age
M/G/1 Queue

\[ X = \text{size distribution} \]
\[ \lambda = \text{arrival rate} \]
\[ \rho = \lambda \mathbb{E}[X] < 1 \]

**Random arrivals**

**Queue**

- **Remaining size**
- **Age**

**Server**

**Scheduling policy:** picks which job to serve
M/G/1 Queue

- \( X = \) size distribution
- \( \lambda = \) arrival rate
- \( \rho = \lambda \mathbb{E}[X] < 1 \)

Scheduling policy:
- picks which job to serve

Random arrivals

Queue

Server

Job

Remaining size

Size

Age

\[ X = \text{size distribution} \]
\[ \lambda = \text{arrival rate} \]
\[ \rho = \lambda \mathbb{E}[X] < 1 \]
M/G/1 Queue

$X = \text{size distribution}$

$\lambda = \text{arrival rate}$

$\rho = \lambda E[X] < 1$

Scheduling policy:

picks which job to serve

random arrivals

queue

server

job

remaining size

size

age

Scheduling policy:
picks which job to serve
M/G/k Queue

random arrivals
M/G/k Queue

random arrivals
M/G/k Queue

$k$ servers, each speed $1/k$
M/G/k Queue

$k$ servers, each speed $1/k$

Scheduling policy: picks which $k$ jobs to serve
Response Time
Response Time
Response Time

$T = \text{response time}$
Response Time

\[ T = \text{response time} \]
Response Time

= $T = \text{response time}$

Goal: schedule to minimize

$\text{mean response time } E[T]$
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
</table>

5
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td></td>
</tr>
</tbody>
</table>
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td>SRPT</td>
</tr>
</tbody>
</table>
Minimizing $E[T]$

<table>
<thead>
<tr>
<th></th>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M/G/1</strong></td>
<td><strong>SRPT</strong></td>
<td><strong>Gittins</strong></td>
</tr>
</tbody>
</table>
What is Gittins?
What is Gittins?
What is Gittins?
What is **Gittins**?

- a.k.a. priority
- rank
- age
What is Gittins?

a.k.a. priority

lower is better

rank

age
What is **Gittins**?

$$r_{\text{Gittins}}(a) = \inf_{b > a} \frac{\mathbb{E}[\min\{X, b\} \mid X > a]}{\mathbb{P}[X \leq b \mid X > a]}$$

- a.k.a. priority
- lower is better
- rank

(age)
What is Gittins?

\[ r_{\text{Gittins}}(a) = \inf_{b > a} \frac{E[\min\{X, b\} | X > a]}{P[X \leq b | X > a]} \]

a.k.a. priority

rank

lower is better

age
What is Gittins?

$$r_{\text{Gittins}}(a) = \inf_{b > a} \frac{\mathbb{E}[\min\{X, b\} \mid X > a]}{\mathbb{P}[X \leq b \mid X > a]}$$

a.k.a. priority

lower is better

rank

age
What is **Gittins**?

\[
r_{\text{Gittins}}(a) = \inf_{b>a} \frac{\mathbb{E}[\min\{X, b\} \mid X > a]}{\mathbb{P}[X \leq b \mid X > a]}
\]

- **a.k.a. priority**
- **rank**: lower is better
- **age**
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>M/G/1</th>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRPT</td>
<td>Gittins</td>
</tr>
</tbody>
</table>
Minimizing $E[T]$

<table>
<thead>
<tr>
<th></th>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M/G/1</strong></td>
<td>SRPT</td>
<td>Gittins</td>
</tr>
<tr>
<td><strong>M/G/k</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in heavy traffic</td>
<td>$\rho \to 1$ limit</td>
<td></td>
</tr>
</tbody>
</table>
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td>SRPT</td>
</tr>
<tr>
<td>M/G/k in heavy traffic ($\rho \to 1$ limit)</td>
<td>SRPT-$k$</td>
</tr>
<tr>
<td></td>
<td>Gittins</td>
</tr>
</tbody>
</table>
## Minimizing $\text{E}[T]$:

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td>SRPT</td>
</tr>
<tr>
<td>M/G/k in heavy traffic ($\rho \to 1$ limit)</td>
<td>SRPT-$k$</td>
</tr>
</tbody>
</table>

- "finite variance" job size distributions
Minimizing $E[T]$

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td>SRPT</td>
</tr>
</tbody>
</table>
| M/G/k in heavy traffic ($\rho \to 1$ limit) | SRPT-$k$ | ???

“finite variance” job size distributions
Minimizing $\mathbb{E}[T]$  

<table>
<thead>
<tr>
<th>Known job sizes</th>
<th>Unknown job sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/G/1</td>
<td>SRPT</td>
</tr>
</tbody>
</table>

- M/G/k in heavy traffic ($\rho \to 1$ limit)
- SRPT-\(k\)
  - "finite variance" job size distributions

Does **Gittins-\(k\)** work?
Background:

SRPT-\(k\) optimality
Background: SRPT-$k$ optimality

(Grosof, Scully, & Harchol-Balter, 2018)
SRPT-$k$ Proof Sketch

SRPT-$k$
SRPT-\(k\) Proof Sketch
SRPT-\textit{k} Proof Sketch

SRPT-1

SRPT-\textit{k}
SRPT-\(k\) Proof Sketch

"close enough"
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]

\( T_1 = \text{M/G/1 response time} \)
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]

"queue delay"

\[ T_1 = \text{M/G/1 response time} \]
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]

\( T_1 \) = M/G/1 response time

\( Q_1 \) = delays due to jobs in *queue*
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]

- **\( T_1 \)** = M/G/1 response time
- **\( Q_1 \)** = delays due to jobs in *queue*
- **\( R_1 \)** = *run delay*
Response Time Decomposition

\[ T_1 = Q_1 + R_1 \]

- \( T_1 \) = M/G/1 response time
- \( Q_1 \) = delays due to jobs in queue
- \( R_1 \) = my size + delays while I’m running
Response Time Decomposition

\[ T_k = Q_k + R_k \]

- \( T_k \) = M/G/k response time
- \( Q_k \) = delays due to jobs in queue
- \( R_k \) = my size + delays while I’m running

“queue delay”

“run delay”
SRPT-\(k\) Proof Sketch

“close enough”

SRPT-1

speed 1

SRPT-\(k\)
SRPT-$k$ Proof Sketch

\[ T_1 = Q_1 + R_1 \]

speed 1

SRPT-1

“close enough”

SRPT-$k$

\[ \frac{1}{k} \]
SRPT-$k$ Proof Sketch

$T_1 = Q_1 + R_1$

"close enough"

$Q_k \leq_{st} Q_1 + O(k) \cdot R_1$
SRPT-\(k\) Proof Sketch

\[ T_1 = Q_1 + R_1 \]

"close enough"

\[ Q_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

\[ R_k \leq_{st} kR_1 \]
SRPT-\(k\) Proof Sketch

\[
T_1 = Q_1 + R_1
\]

**Step 1:**
\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

“close enough”

\[
Q_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

\[
R_k \leq_{st} kR_1
\]
SRPT-$k$ Proof Sketch

Step 1:

\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2:

\[ T_1 = Q_1 + R_1 \]

\[ \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0 \]
SRPT-k Proof Sketch

Step 1:
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2:
\[ \lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0 \]

SRPT in heavy-traffic M/G/1
(Lin, Wierman, & Zwart, 2011)
SRPT-\(k\) Proof Sketch

Step 1: \[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2: \[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]

Theorem: \[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_k]}{\mathbb{E}[T_1]} = 1
\]

SRPT in heavy-traffic M/G/1

(Lin, Wierman, & Zwart, 2011)

\[ T_1 = Q_1 + R_1 \]
Generalizing to **Gittins-\(k\)?**

**Step 1:**

\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

**Step 2:**

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]
Generalizing to **Gittins-**~\(k\)~?

- **Step 1:**
  \[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

- **Step 2:**
  \[ \lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0 \]
Generalizing to Gittins-\(k\)?

Step 1:
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2:
\[ \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0 \]

might not be “close enough”
Generalizing to Gittins-$k$?

Step 1:

Gittins-$1$

might not be “close enough”

Gittins-$k$

Step 2:

Heavy-traffic scaling of Gittins-$1$ unknown

Step 1:

$T_k \leq Q_1 + O(k) \cdot R_1$

Step 2:

$\lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0$
Generalizing to **Gittins-\( k \)?**

- **Gittins-1**
  - Step 1:
    \[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]
  - Step 2:
    \[ \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0 \]

- **Gittins-\( k \)**
  - Might not be "close enough"

Heavy-traffic scaling of **Gittins-1** unknown
New policy: M-Gittins
New policy: M-Gittins

Theorem:

$$\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-}k}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1$$

OPT
New policy: M-Gittins

Theorem:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-k}}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1
\]

Step 1:
\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

Step 2:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]
New policy: M-Gittins

Theorem:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{M-Gittins-k}]}{\mathbb{E}[T_{Gittins-1}]} = 1
\]

Step 1:
\[T_k \leq_{st} Q_1 + O(k) \cdot R_1\]

Step 2:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]
New policy: M-Gittins

Theorem:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{M \text{-Gittins} - k}]}{\mathbb{E}[T_{\text{Gittins} - 1}]} = 1
\]

Step 1:
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]
What is M-Gittins?
What is M-Gittins?

a.k.a. priority

rank

age
What is M-Gittins?

\[ r_{\text{M-Gittins}}(a) = \max_{0 \leq b \leq a} r_{\text{Gittins}}(b) \]

a.k.a. priority

rank

age
What is M-Gittins?

\[ r_{M-Gittins}(a) = \max_{0 \leq b \leq a} r_{\text{Gittins}}(b) \]

a.k.a. priority

rank

age

Gittins
What is M-Gittins?

\[ r_{M-Gittins}(a) = \max_{0 \leq b \leq a} r_{Gittins}(b) \]

a.k.a. priority

rank

age

M-Gittins

Gittins
What is M-Gittins?

\[ r_{M-Gittins}(a) = \max_{0 \leq b \leq a} r_{Gittins}(b) \]
Why Monotonicity?
Why Monotonicity?

Conceptually easier: rank only gets worse with age
Why Monotonicity?

Conceptually easier: rank only gets worse with age

solves Step 1
Why Monotonicity?

Conceptually easier: rank only gets worse with age

Analytically easier: simple formulas for $E[Q_1]$ and $E[R_1]$
Why Monotonicity?

Conceptually easier: rank only gets worse with age

Analytically easier: simple formulas for $E[Q_1]$ and $E[R_1]$

solves Step 1

helps Step 2
Heavy-Traffic Optimality
Theorem 1: \( \text{M-Gittins-k} \) is heavy-traffic optimal in the \( M/G/k \), specifically

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[ T_{\text{M-Gittins-k}} ]}{\mathbb{E}[ T_{\text{Gittins-1}} ]} = 1,
\]
Heavy-Traffic Optimality

Theorem 1:

**M-Gittins-k** is *heavy-traffic optimal* in the M/G/k, specifically

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-k}}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1,
\]

if X is in *any* of the following classes:
Heavy-Traffic Optimality

Theorem 1:

**M-Gittins-k** is heavy-traffic optimal in the M/G/k, specifically if $X$ is in any of the following classes:

- bounded

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-k}}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1,
\]
Theorem 1:

\( \text{M-Gittins-k} \) is heavy-traffic optimal in the \( \text{M/G/}k \), specifically

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-k}}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1,
\]

if \( X \) is in any of the following classes:

- bounded
- “finite-variance heavy-tailed”
  \( (O\text{-regularly varying with Matuszewska indices less than } -2) \)
Theorem 1: \textbf{M-Gittins-}k \textit{is heavy-traffic optimal} in the M/G/k, specifically

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-}k}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1,
\]

if \( X \) is in \textit{any} of the following classes:

- bounded
- “finite-variance heavy-tailed” (\( O \)-regularly varying with Matuszewska indices less than \(-2\))
- \( \text{MDA}(\Lambda) \) with “quasi-decreasing hazard rate”, e.g. \( h(x) = \Theta(x^{-\gamma}) \)
Heavy-Traffic Optimality

Theorem 1:

M-Gittins-\(k\) is heavy-traffic optimal in the M/G/\(k\), specifically

\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-}k}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1,
\]

if \(X\) is in any of the following classes:

• bounded
• “finite-variance heavy-tailed”
  \(O\text{-regularly varying with Matuszewska indices less than } -2\)
• MDA(\(\Lambda\)) with “quasi-decreasing hazard rate”, e.g. \(h(x) = \Theta(x^{-\gamma})\)

exponential, log-normal, Weibull…
M-Gittins-k Proof Sketch

M-Gittins-k

speed 1

M-Gittins-1

M-Gittins-k

1/k
M-Gittins-\(k\) Proof Sketch

“close enough”

M-Gittins-1

M-Gittins-\(k\)
M-Gittins-k Proof Sketch

Step 1:

\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

(speed 1)

M-Gittins-1

“close enough”

M-Gittins-k

M-Gittins-k

M-Gittins-1
M-Gittins-k Proof Sketch

**Step 1:**

\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

**Step 2:**

\[ \lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0 \]
**M-Gittins-k Proof Sketch**

**Step 1:**
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

**Step 2:**
\[ \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0 \]

**Key fact:**
\[ Q_{M-Gittins} \leq_{st} Q_{Gittins} \]
M-Gittins-k Proof Sketch

Step 1:
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

Step 2:
\[ \lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0 \]

Key fact:
\[ Q_{M-Gittins} \leq_{st} Q_{Gittins} \]

Theorem:
\[ \lim_{\rho \to 1} \frac{\mathbb{E}[T_{M-Gittins-k}]}{\mathbb{E}[T_{Gittins-1}]} = 1 \]
**M-Gittins-k Proof Sketch**

**Step 1:**
\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

**Step 2:**
\[ \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0 \]

**Theorem:**
\[ \lim_{\rho \to 1} \frac{E[T_{M-Gittins-k}]}{E[T_{Gittins-1}]} = 1 \]
M/G/1 Heavy-Traffic Scaling
M/G/1 Heavy-Traffic Scaling

Theorem 2:

“infinite variance”

“finite variance”
M/G/1 Heavy-Traffic Scaling

Theorem 2:
For each of Gittins and M-Gittins, if $X \in \text{OR}(-2,-1)$, then

$$E[Q_1], E[R_1] = \Theta\left( \log \frac{1}{1-\rho} \right),$$

"infinite variance"

"finite variance"
M/G/1 Heavy-Traffic Scaling

Theorem 2:

For each of Gittins and M-Gittins, if $X \in \text{OR}(-2, -1)$, then

$$\mathbb{E}[Q_1], \mathbb{E}[R_1] = \Theta\left(\log \frac{1}{1 - \rho}\right),$$

and if $X \in \text{OR}(-\infty, -2) \cup \text{MDA}(\Lambda) \cup \text{ENBUE}$, then

$$\mathbb{E}[Q_1] = \Theta\left(\frac{1}{1 - \rho} \max_{0 \leq b \leq F_e^{-1}(1 - \rho)} \mathbb{E}[X - b \mid X > b]\right),$$

$$\mathbb{E}[R_1] = o(\mathbb{E}[Q_1]).$$
M/G/1 Heavy-Traffic Scaling

Theorem 2:

For each of Gittins and M-Gittins, if $X \in \text{OR}(-2, -1)$, then

$$E[Q_1], E[R_1] = \Theta\left(\log \frac{1}{1 - \rho}\right),$$

and if $X \in \text{OR}(-\infty, -2) \cup \text{MDA}(\Lambda) \cup \text{ENBUE}$, then

$$E[Q_1] = \Theta\left(\frac{1}{1 - \rho} \max_{0 \leq b \leq F_e^{-1}(1 - \rho)} E[X - b | X > b]\right)$$

$$E[R_1] = o(E[Q_1]).$$

Step 2
Key Building Blocks
Key Building Blocks

**SRPT-**$^k$ optimality
(Grosof, Scully, & Harchol-Balter, 2018)
Key Building Blocks

**SRPT-\(k\) optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: \(E[Q_1]\) and \(E[R_1]\)
(Scully, Harchol-Balter & Scheller-Wolf, 2018)
Key Building Blocks

**SRPT-k optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: $E[Q_1]$ and $E[R_1]$
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]
Key Building Blocks

**SRPT-**\( k \) optimality
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: \( E[Q_1] \) and \( E[R_1] \)
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

\( \text{needs monotonicity} \)
**Key Building Blocks**

**SRPT-$k$ optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: $E[Q_1]$ and $E[R_1]$
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

Heavy-traffic analysis of FB in M/G/1
(Kamphorst & Zwart, 2020)

\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

⚠️ needs **monotonicity**

Step 1:
Key Building Blocks

SRPT-\(k\) optimality
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: \(E[Q_1] \) and \(E[R_1]\)
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

Heavy-traffic analysis of FB in M/G/1
(Kamphorst & Zwart, 2020)

Understanding Gittins rank
(Aalto, Ayesta, & Righter, 2009, 2011)

\[
\text{Step 1: } T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

⚠️ needs \textbf{monotonicity}
Key Building Blocks

**SRPT-$$k$$ optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: $$E[Q_1]$$ and $$E[R_1]$$
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

Heavy-traffic analysis of FB in M/G/1
(Kamphorst & Zwart, 2020)

Understanding **Gittins rank**
(Aalto, Ayesta, & Righter, 2009, 2011)

Simple policy with near-optimal $$E[T_1]$$
(Scully, Harchol-Balter & Scheller-Wolf, 2020)

⚠️ **Step 1:**
$$T_k \leq_{st} Q_1 + O(k) \cdot R_1$$

⚠️ needs **monotonicity**

$$T_k \leq_{st} Q_1 + O(k) \cdot R_1$$
Key Building Blocks

**SRPT-\(k\) optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: \(E[Q_1]\) and \(E[R_1]\)
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

Heavy-traffic analysis of FB in M/G/1
(Kamphorst & Zwart, 2020)

Understanding **Gittins rank**
(Aalto, Ayesta, & Righter, 2009, 2011)

Simple policy with near-optimal \(E[T_1]\)
(Scully, Harchol-Balter & Scheller-Wolf, 2020)

\begin{align*}
\textbf{Step 1:} & \quad T_k \leq_{st} Q_1 + O(k) \cdot R_1 \\
\textbf{Step 2:} & \quad \lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0
\end{align*}

⚠️ needs **monotonicity**
Key Building Blocks

**SRPT-\( k \) optimality**
(Grosof, Scully, & Harchol-Balter, 2018)

“SOAP” analysis: \( \mathbb{E}[Q_1] \) and \( \mathbb{E}[R_1] \)
(Scully, Harchol-Balter & Scheller-Wolf, 2018)

Heavy-traffic analysis of FB in M/G/1
(Kamphorst & Zwart, 2020)

Understanding **Gittins rank**
(Aalto, Ayesta, & Righter, 2009, 2011)

Simple policy with near-optimal \( \mathbb{E}[T_1] \)
(Scully, Harchol-Balter & Scheller-Wolf, 2020)

\[ T_k \leq_{st} Q_1 + O(k) \cdot R_1 \]

\[ \lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0 \]
Summary

**Theorem:**

$$\lim_{\rho \to 1} \frac{\mathbb{E}[T_{M-Gittins-k}]}{\mathbb{E}[T_{Gittins-1}]} = 1$$
Summary

Theorem:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{\text{M-Gittins-k}}]}{\mathbb{E}[T_{\text{Gittins-1}}]} = 1
\]

Step 1:
\[
T_k \leq_{st} Q_1 + O(k) \cdot R_1
\]

Step 2:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]
Summary

Theorem:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[T_{M-Gittins-k}]}{\mathbb{E}[T_{Gittins-1}]} = 1
\]

Step 1:
\[T_k \leq_{st} Q_1 + O(k) \cdot R_1\]

Step 2:
\[
\lim_{\rho \to 1} \frac{\mathbb{E}[R_1]}{\mathbb{E}[Q_1]} = 0
\]

Characterized M/G/1 heavy-traffic scaling of Gittins and M-Gittins
Summary

Theorem:
\[
\lim_{\rho \to 1} \frac{E[T_{\text{M-Gittins-k}}]}{E[T_{\text{Gittins-1}}]} = 1
\]

Step 1:
\[T_k \leq Q_1 + O(k) \cdot R_1\]

Step 2:
\[
\lim_{\rho \to 1} \frac{E[R_1]}{E[Q_1]} = 0
\]

Characterized M/G/1 heavy-traffic scaling of Gittins and M-Gittins

Get in touch: zscully@cs.cmu.edu
Bonus Slides
Comparison with FB
Comparison with FB

“finite variance”
Comparison with FB

“finite variance”

serves job of least age
Comparison with FB

Foreground-Background (FB)

$$E[Q_1] = \Theta\left(\frac{1}{1-\rho} \bigg/ E[X - \bar{F}_e^{-1}(1-\rho) \mid X > \bar{F}_e^{-1}(1-\rho)]\right)$$
$$E[R_1] = o(E[Q_1])$$

(Kamphorst & Zwart, 2020)

Gittins and M-Gittins

$$E[Q_1] = \Theta\left(\frac{1}{1-\rho} \bigg/ \max_{0 \leq b \leq \bar{F}_e^{-1}(1-\rho)} E[X - b \mid X > b] \right)$$
$$E[R_1] = o(E[Q_1])$$