A new $O\Big(1/(1-\rho)\Big)$-scaling bound for multiserver queues

via a leave-one-out technique

Yige Hong

Carnegie Mellon University

APS Conference, July 1, 2025

much smaller constant

A new $O\big(1/(1-\rho)\big)$-scaling bound for multiserver queues

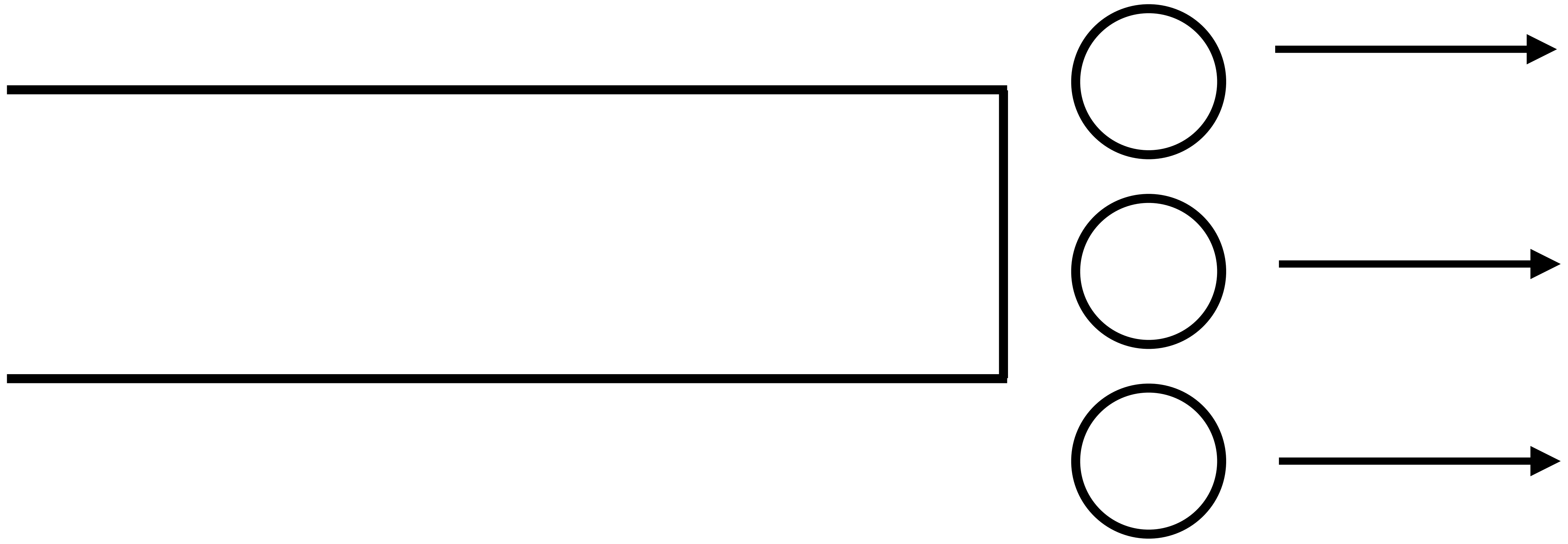
via a leave-one-out technique

Yige Hong

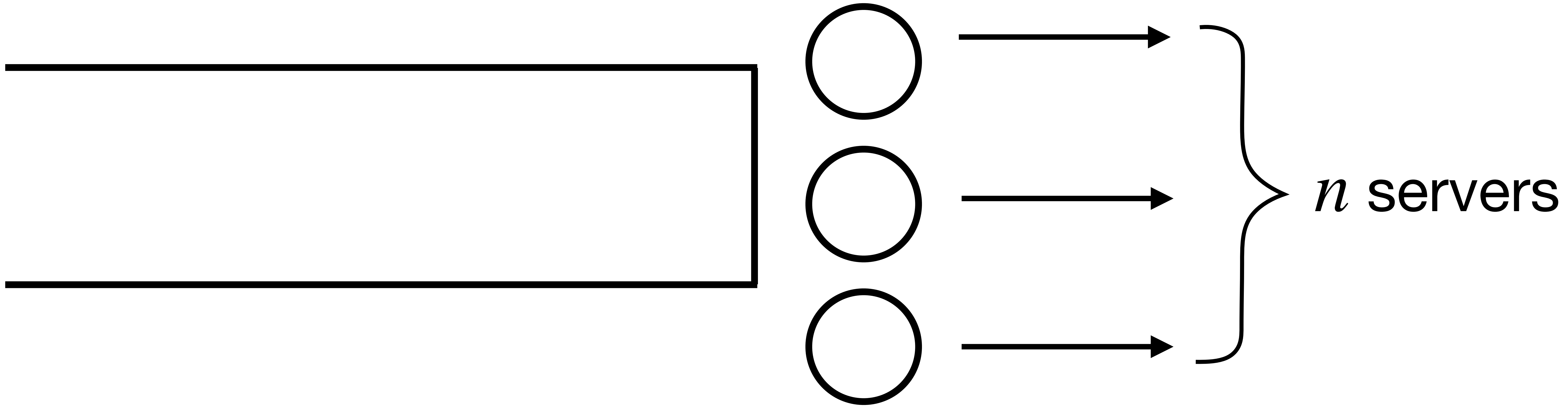
Carnegie Mellon University

APS Conference, July 1, 2025

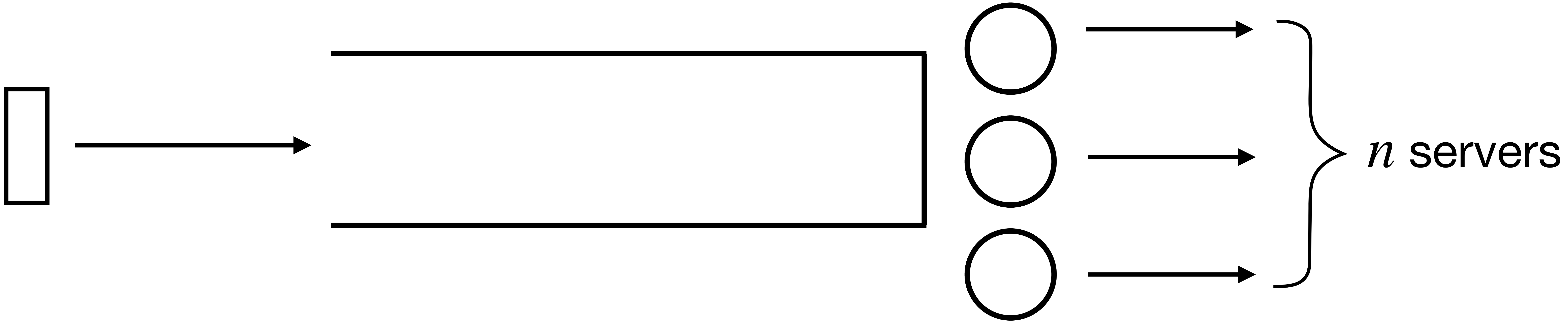
Model: GI/GI/n queue under FCFS



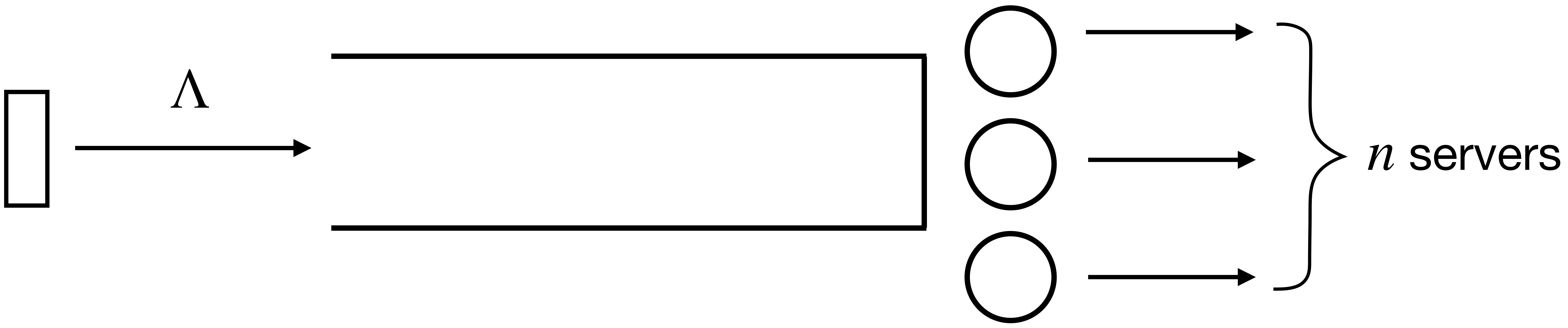
Model: GI/GI/n queue under FCFS



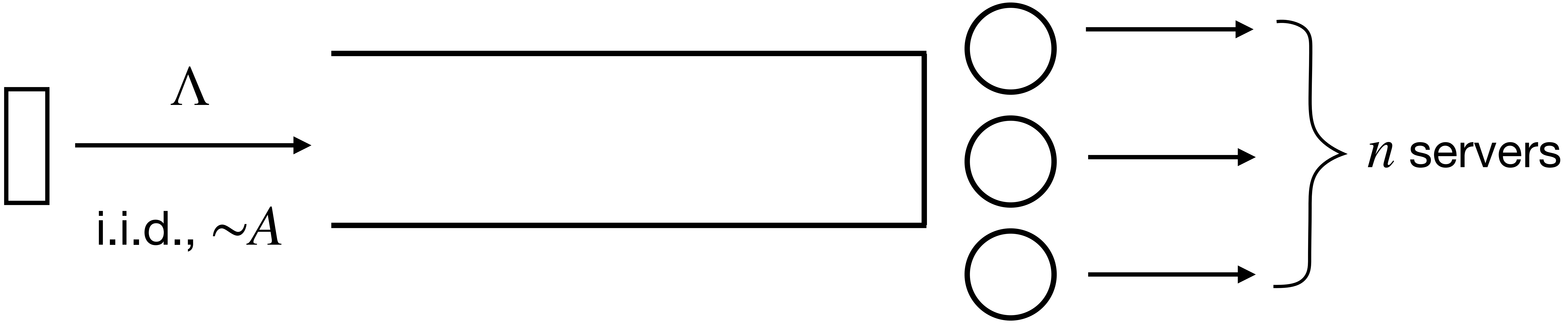
Model: GI/GI/n queue under FCFS



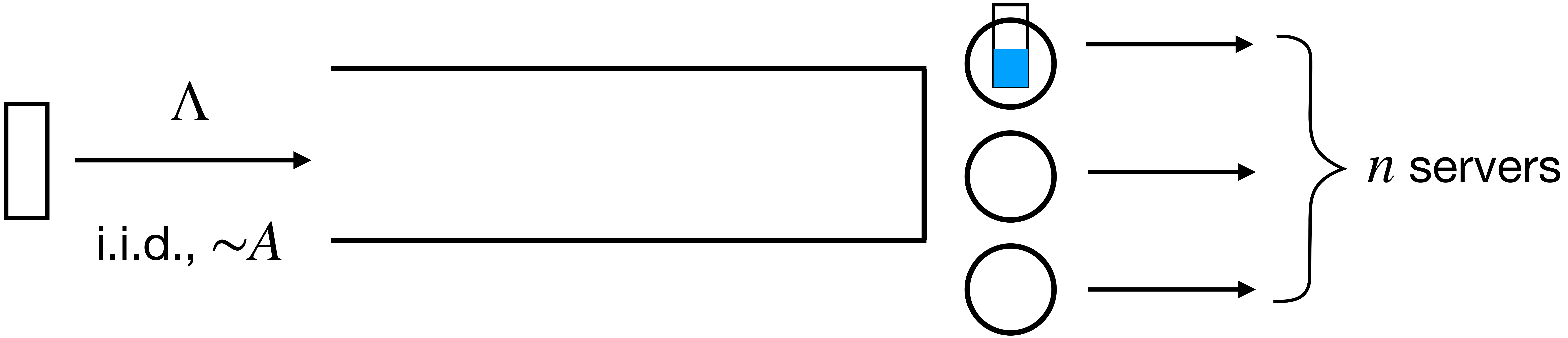
Model: GI/GI/n queue under FCFS



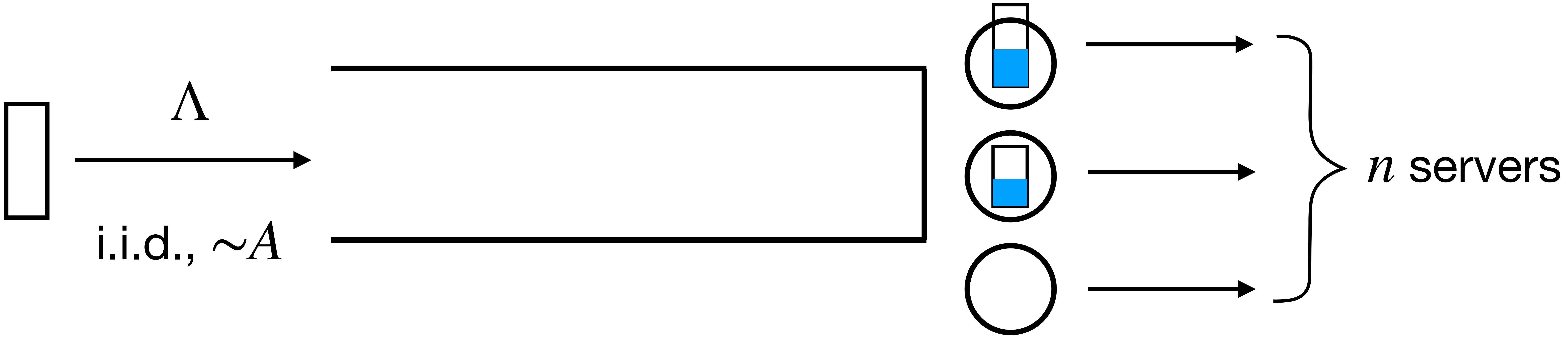
Model: GI/GI/n queue under FCFS



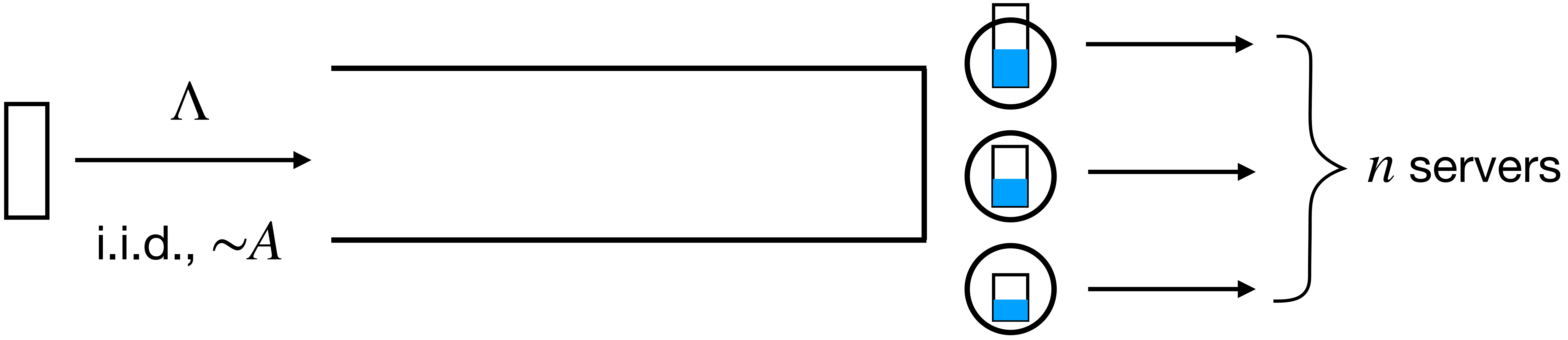
Model: GI/GI/n queue under FCFS



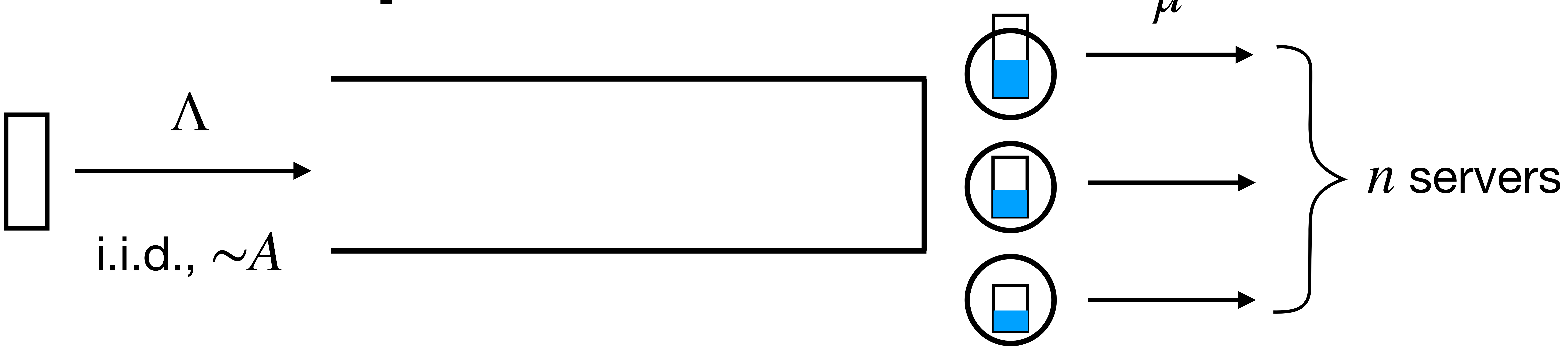
Model: GI/GI/n queue under FCFS



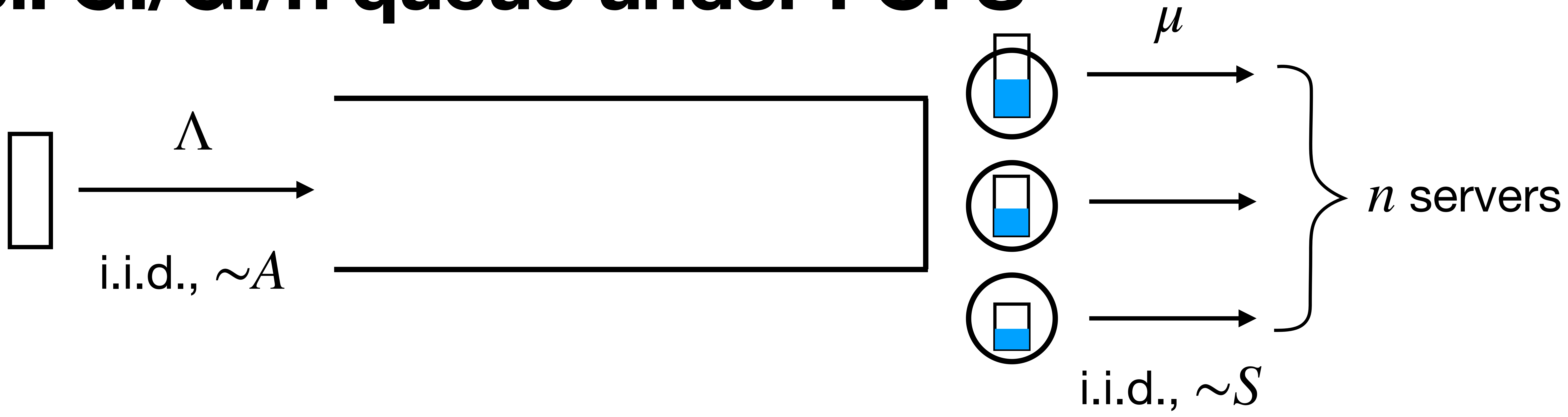
Model: GI/GI/n queue under FCFS



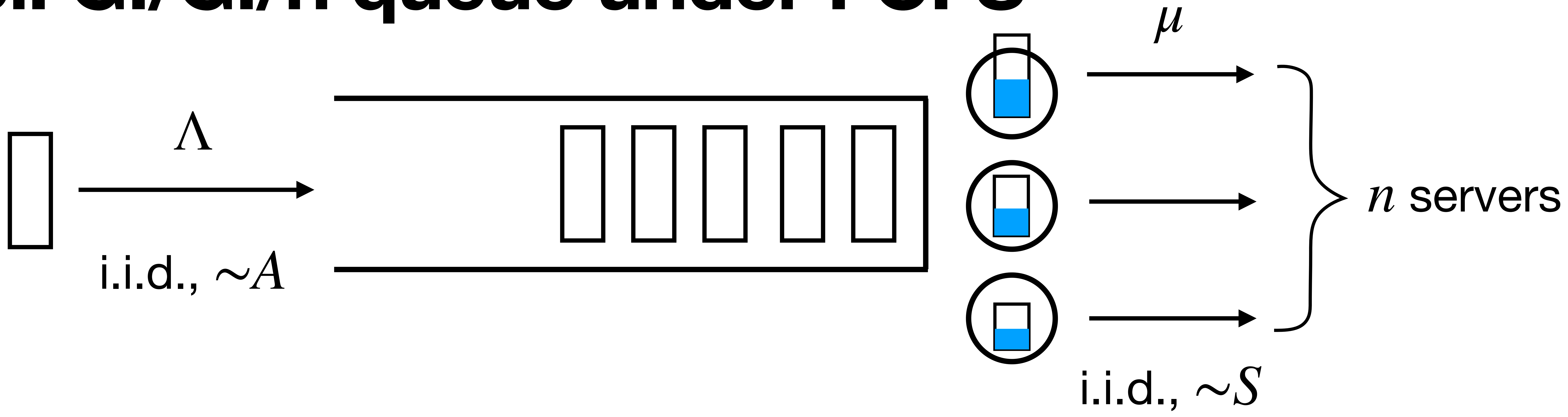
Model: GI/GI/n queue under FCFS



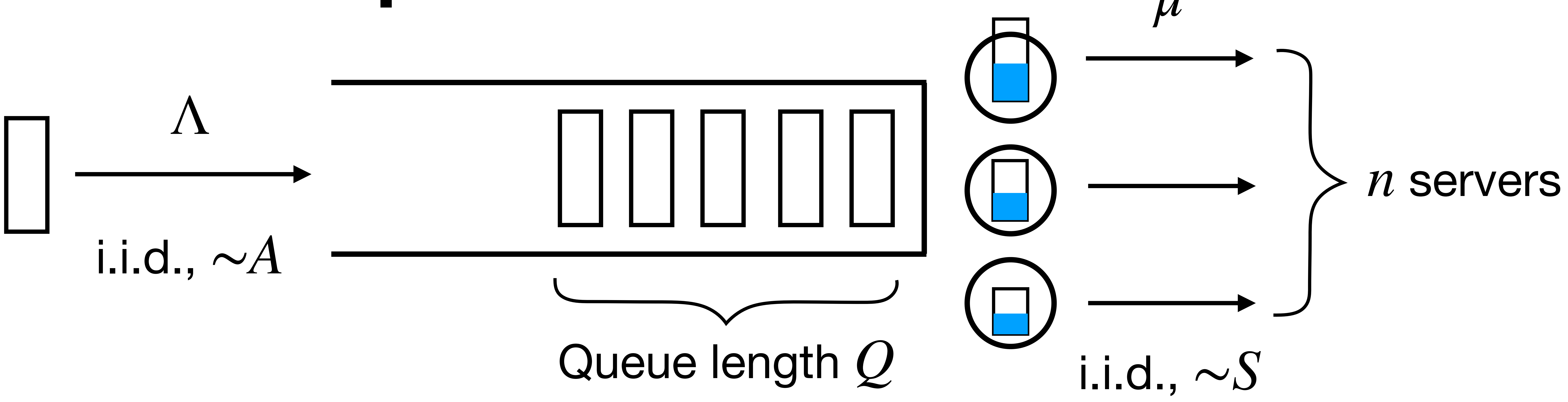
Model: GI/GI/n queue under FCFS



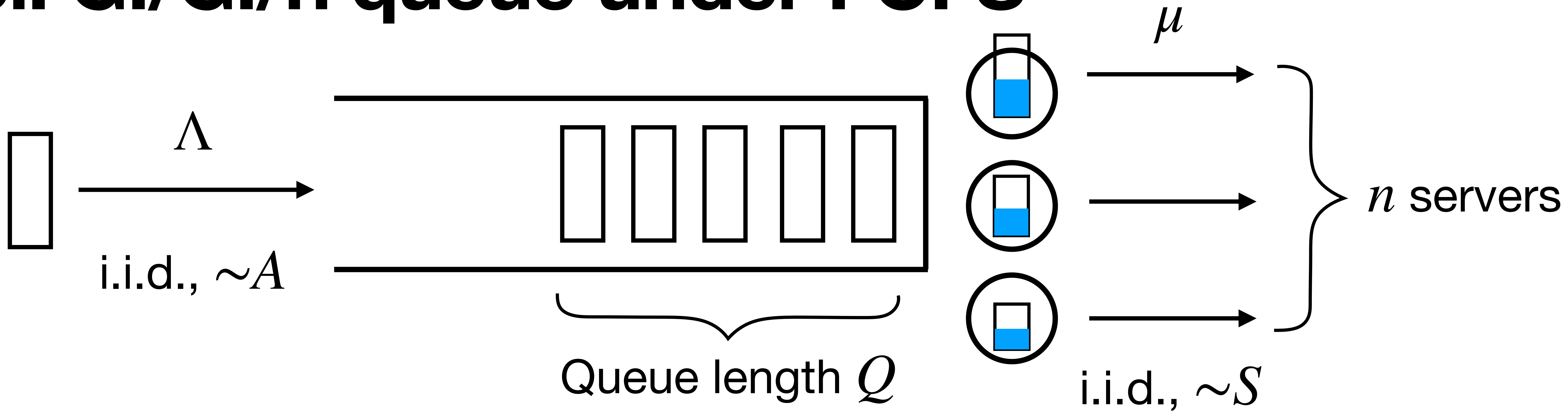
Model: GI/GI/n queue under FCFS



Model: GI/GI/n queue under FCFS

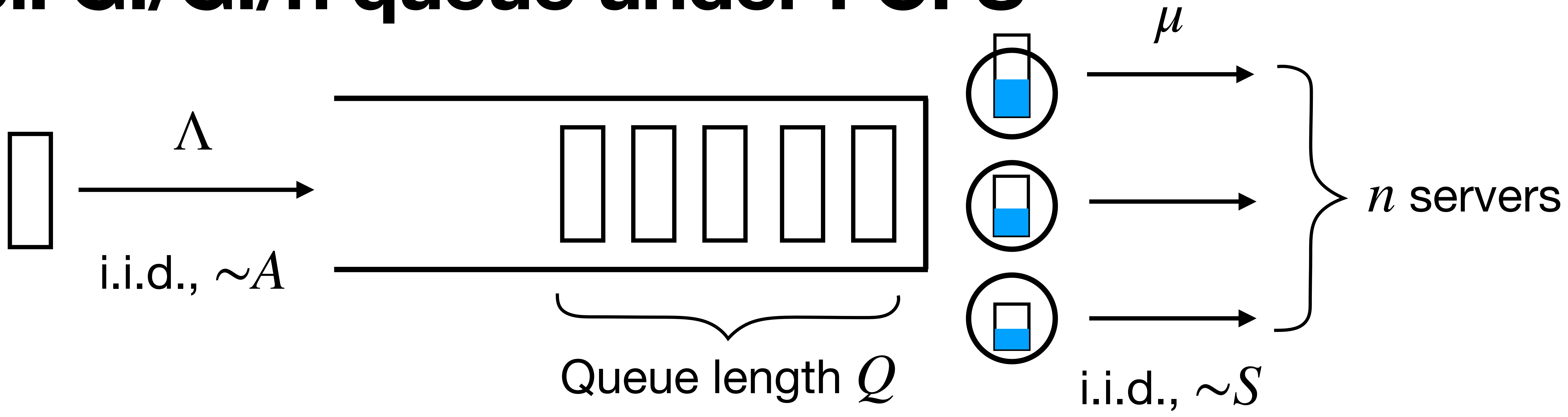


Model: GI/GI/n queue under FCFS

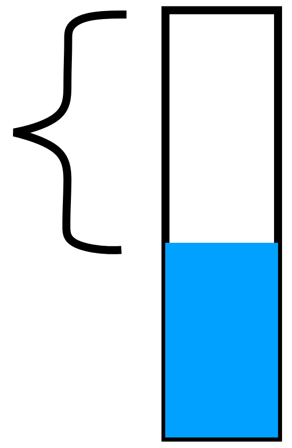


Notation:

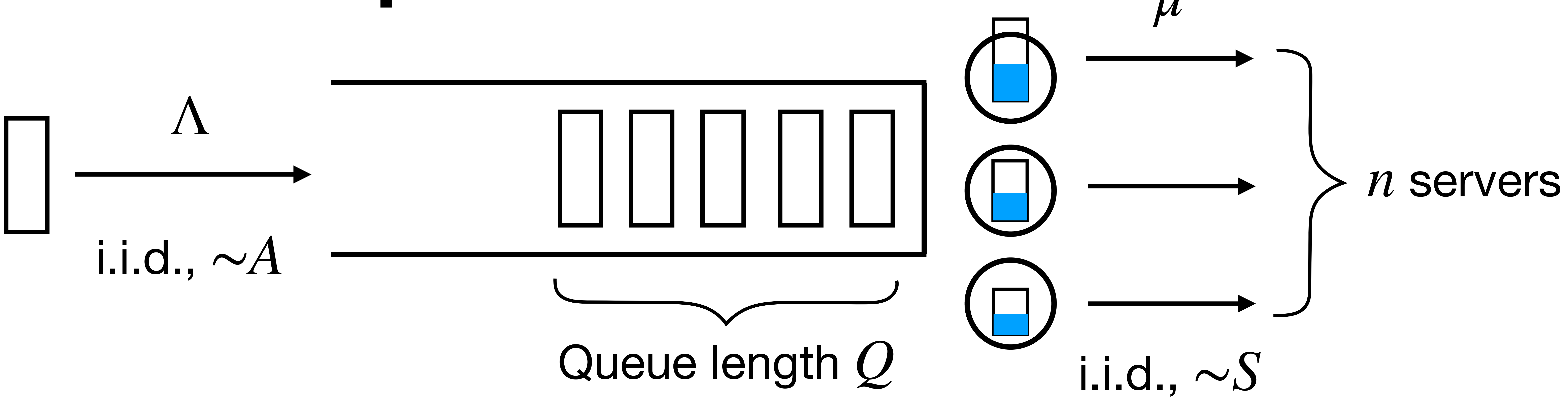
Model: GI/GI/n queue under FCFS



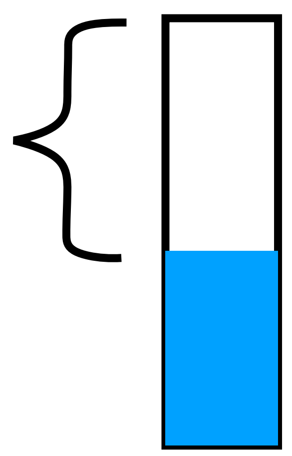
Notation:

- Residual arrival time R_a ; residual service time $R_{s,j}$ 

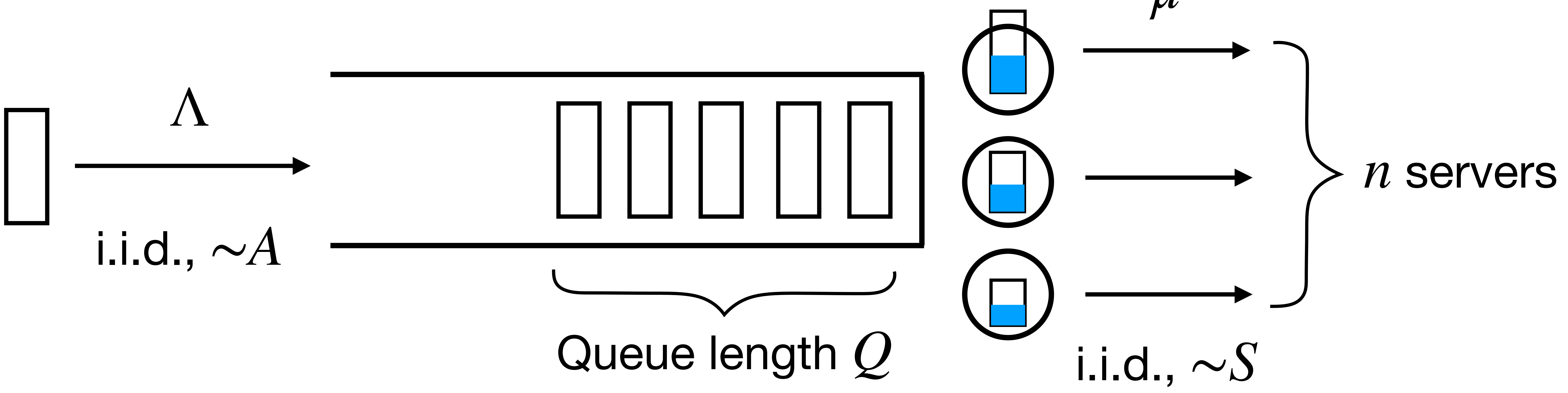
Model: GI/GI/n queue under FCFS



Notation:

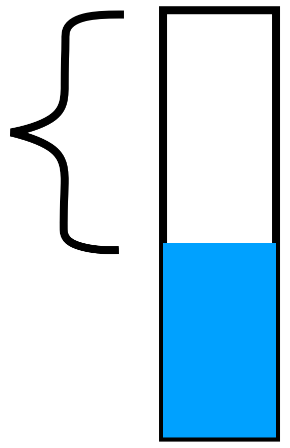
- Residual arrival time R_a ; residual service time $R_{s,j}$ 
- Load $\rho = \dfrac{\Lambda}{n\mu}$

Model: GI/GI/n queue under FCFS



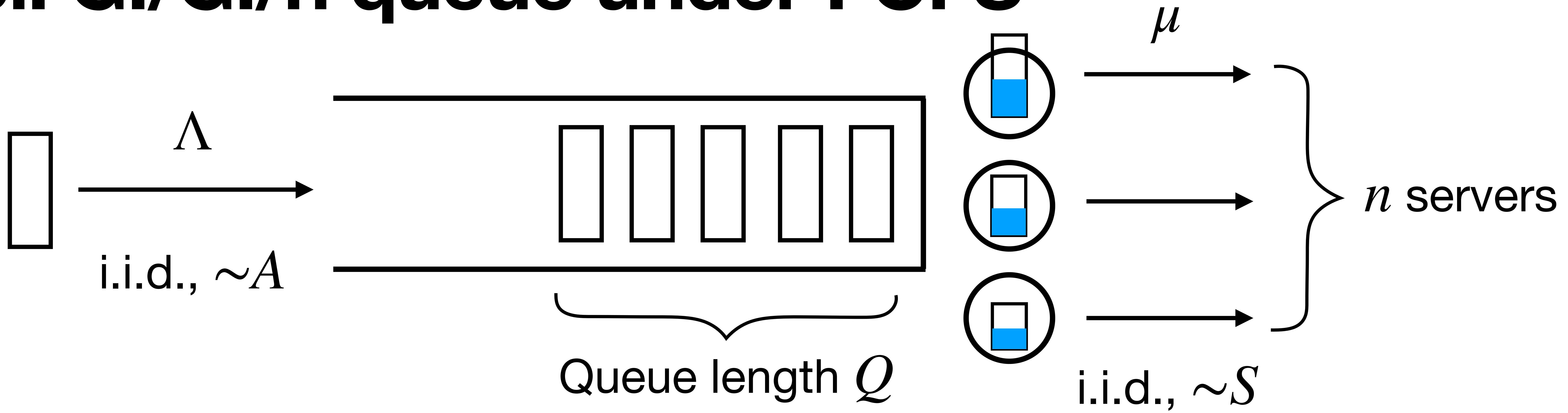
Notation:

- Residual arrival time R_a ; residual service time $R_{s,j}$
- Load $\rho = \dfrac{\Lambda}{n\mu}$



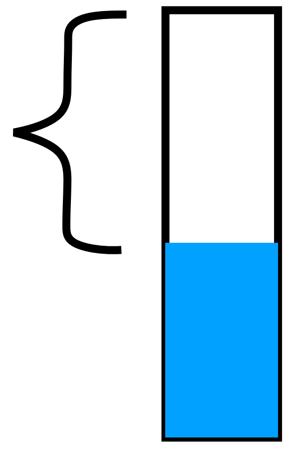
How does the steady-state $\mathbb{E}[Q]$ scale when n is large?

Model: GI/GI/n queue under FCFS



Notation:

- Residual arrival time R_a ; residual service time $R_{s,j}$
- Load $\rho = \dfrac{\Lambda}{n\mu}$



How does the steady-state $\mathbb{E}[Q]$ scale when n is large?

Goal: show $\mathbb{E}[Q] \leq O\left(1/(1-\rho)\right)$ for arbitrary $\rho < 1$ and n

Prior work

Prior work

Most prior work: focus on representative or useful scaling regimes

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2})$, $n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2})$, $n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1})$, $n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \left(2.1 \times 10^{21} \times \mathbb{E}[(\mu S)^2] \times \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2] \right) \times \frac{1}{1-\rho}$$

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65] [Iglehart and Whitt 70] [Kollerstrom 74, 79] [Loulou 73] [Scheller-Wolf 03] [Hokstad 85] [Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09] [Gamarnik and Momcilovic 08] [Gamarnik and Stolyar 12] [Braverman and Dai 15, 16] [Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11] [Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \left(2.1 \times 10^{21} \times \mathbb{E}[(\mu S)^2] \times \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2] \right) \boxed{\times \frac{1}{1-\rho}}$$

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2})$, $n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1})$, $n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \le \left(\boxed{2.1 \times 10^{21}} \times \mathbb{E}[(\mu S)^2] \times \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}]\right) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2]\right) \times \boxed{\frac{1}{1-\rho}}$$

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \le \Bigg(\boxed{2.1 \times 10^{21}} \times \mathbb{E}[\quad \cdots \quad + \mathbb{E}[(\mu S)^{2+\epsilon}] \Big) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2] \Bigg) \times \boxed{\frac{1}{1-\rho}}$$

Fun fact:
age of the universe
$= 4.35 \times 10^{17}$ **sec**

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:	Halfin-Whitt:	Non-Degenerate Slowdown
$\rho \uparrow 1$, n fixed, $\mathbb{E}[Q] = O(1/(1-\rho))$	$\rho = 1 - \Theta(n^{-1/2})$, $n \to \infty$ $\mathbb{E}[Q] = O(n^{1/2})$	$\rho = 1 - \Theta(n^{-1})$, $n \to \infty$ $\mathbb{E}[Q] = O(n)$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \left(\boxed{2.1 \times 10^{21}} \times \mathbb{E}[(\mu S)^2] \times \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2] \right) \boxed{\times \frac{1}{1-\rho}}$$

Prior work

Most prior work: focus on representative or useful scaling regimes

Classic heavy-traffic:
$\rho \uparrow 1$, n fixed,
$\mathbb{E}[Q] = O(1/(1-\rho))$

[Kingman 62, 70] [Borokov 65]
[Iglehart and Whitt 70]
[Kollerstrom 74, 79] [Loulou 73]
[Scheller-Wolf 03] [Hokstad 85]
[Grosof et al. 22]…

Halfin-Whitt:
$\rho = 1 - \Theta(n^{-1/2}), n \to \infty$
$\mathbb{E}[Q] = O(n^{1/2})$

[Hafin and Whitt 81] [Reed 09]
[Gamarnik and Momcilovic 08]
[Gamarnik and Stolyar 12]
[Braverman and Dai 15, 16]
[Aghajani and Ramanan 20]…

Non-Degenerate Slowdown
$\rho = 1 - \Theta(n^{-1}), n \to \infty$
$\mathbb{E}[Q] = O(n)$

[Whitt 03] [Atar and Solomon 11]
[Atar 12]…

$\mathbb{E}[Q] = O\big(1/(1-\rho)\big)$
for each scaling regime

Another type of results: universal across scaling regimes

Li and Goldberg 25 (Corollary 2, adapted):

M as shorthand

GI/GI/n. Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \left(\boxed{2.1 \times 10^{21}} \times \boxed{\mathbb{E}[(\mu S)^2] \times \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \times \left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}[(\Lambda A)^2]} \right) \times \boxed{\frac{1}{1-\rho}}$$

Known result for GI/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for GI/GI/n

Known result for GI/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \le \frac{2.1 \times 10^{21} \times M}{1-\rho}$$

Our result for GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \le \frac{2R_s^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

where

$$R_s^{\max} = \sup_{t \ge 0} \mathbb{E}[\mu S - t \,|\, \mu S \ge t]$$

Known result for GI/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2R_s^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

where

$$R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$$

Finite if "light-tailed"

Known result for GI/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2R_s^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

where

$$R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$$

Finite if "light-tailed"

(Omit some Harris ergodicity assumptions)

Known result for GI/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2R_s^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

where

$$R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$$

Finite if "light-tailed"

(Omit some Harris ergodicity assumptions)

⭐Generalizable to heterogeneous servers

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[S^{2+\epsilon}] < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \,|\, \mu S \geq t]$

(Omit some Harris ergodicity assumptions)

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$

Our result implies:

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \,|\, \mu S \geq t]$

Our result implies:

- $\mathbb{E}[Q] \leq \dfrac{\max(1/\alpha, 1)}{1 - \rho}$ if $S \sim$ Gamma with shape parameter α

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$

Our result implies:

- $\mathbb{E}[Q] \leq \dfrac{\max(1/\alpha, 1)}{1 - \rho}$ if $S \sim$ Gamma with shape parameter α
- $\mathbb{E}[Q] \leq \dfrac{\mu \max_{i \in [d]} \tau_i}{1 - \rho}$ if $S \sim$ Phase-type, where τ_i = expected completion time from phase i

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1-\rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$

Our result implies:

- $\mathbb{E}[Q] \leq \dfrac{\max(1/\alpha, 1)}{1-\rho}$ if $S \sim$ Gamma with shape parameter α
- $\mathbb{E}[Q] \leq \dfrac{\mu \max_{i \in [d]} \tau_i}{1-\rho}$ if $S \sim$ Phase-type, where τ_i = expected completion time from phase i
- $\mathbb{E}[Q] \leq \dfrac{1}{1-\rho}$ if $S \sim$ NBUE (e.g. Increasing Failure Rate, Erlang, Weibull with $k \geq 1$)

Known result for M/GI/n

(Li and Goldberg 25, Corollary 2, adapted)

Assume $\mathbb{E}[A^2] < \infty$, $\mathbb{E}[S^{2+\epsilon}] < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2.1 \times 10^{21} \times M}{1 - \rho}$$

Our result for M/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

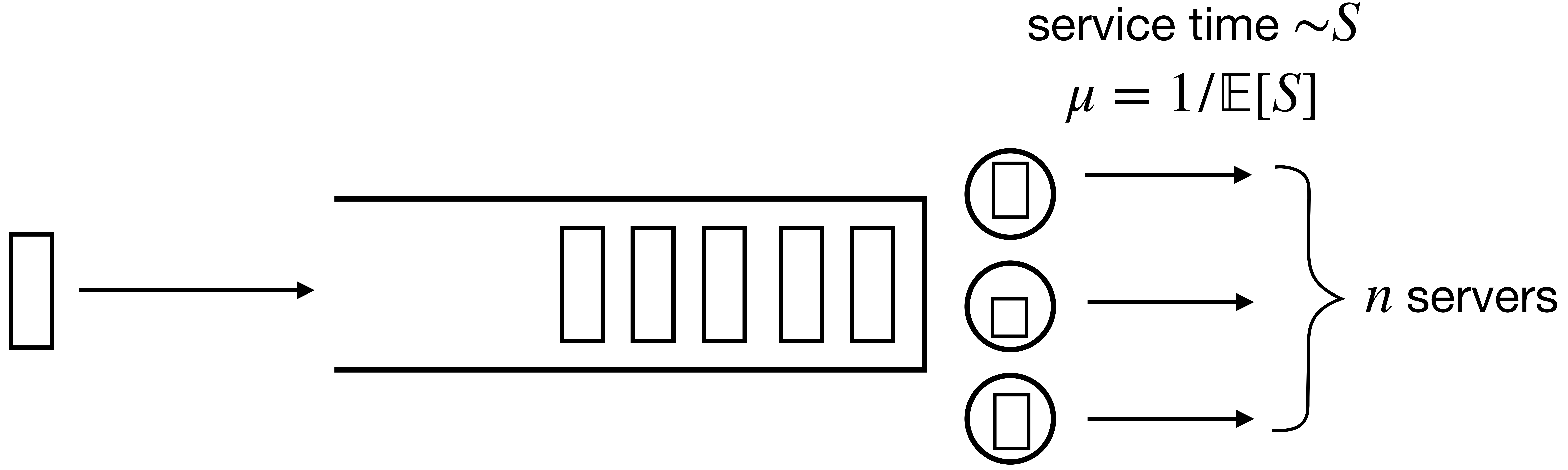
$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1 - \rho}$$

where $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$

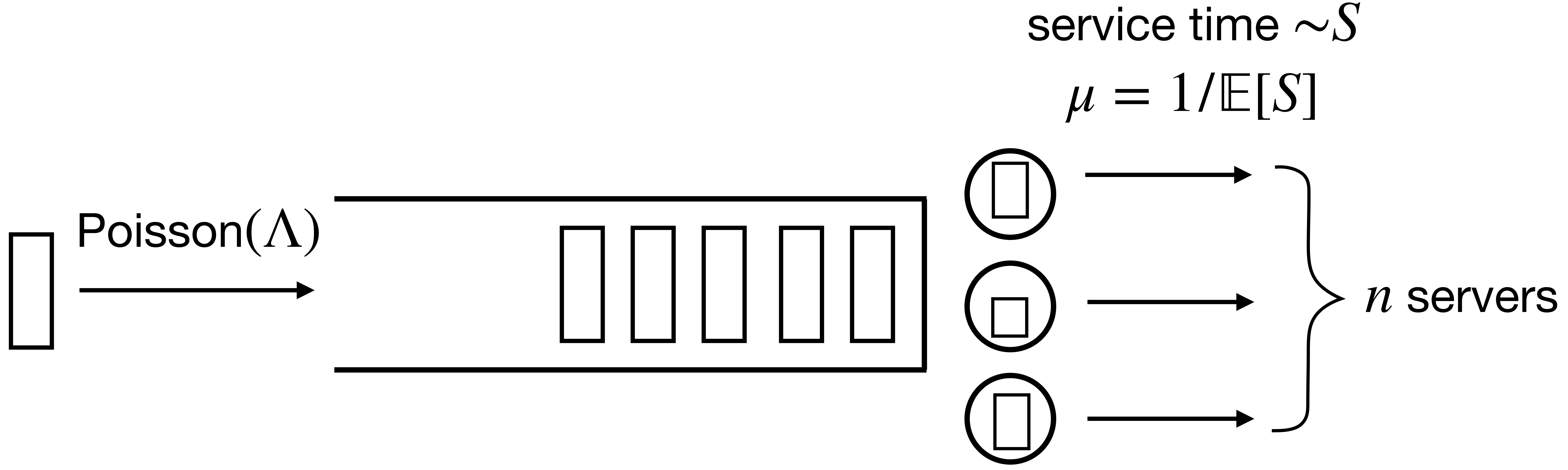
Our result implies:

- $\mathbb{E}[Q] \leq \dfrac{\max(1/\alpha, 1)}{1 - \rho}$ if $S \sim$ Gamma with shape parameter α
- $\mathbb{E}[Q] \leq \dfrac{\mu \max_{i \in [d]} \tau_i}{1 - \rho}$ if $S \sim$ Phase-type, where τ_i = expected completion time from phase i
- $\mathbb{E}[Q] \leq \dfrac{1}{1 - \rho}$ if $S \sim$ NBUE (e.g. Increasing Failure Rate, Erlang, Weibull with $k \geq 1$)
- $\mathbb{E}[Q] \leq \dfrac{\mu S_{\max}}{1 - \rho}$ if S is bounded by $S_{\max}$

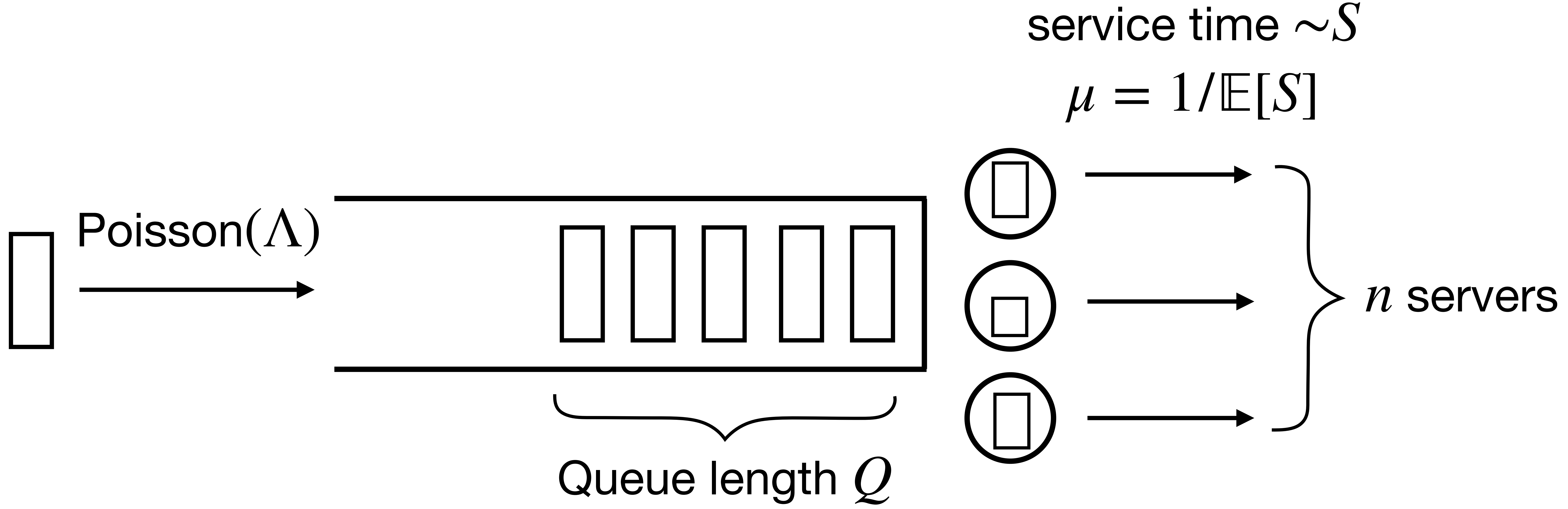
Proof sketch: focus on M/GI/n



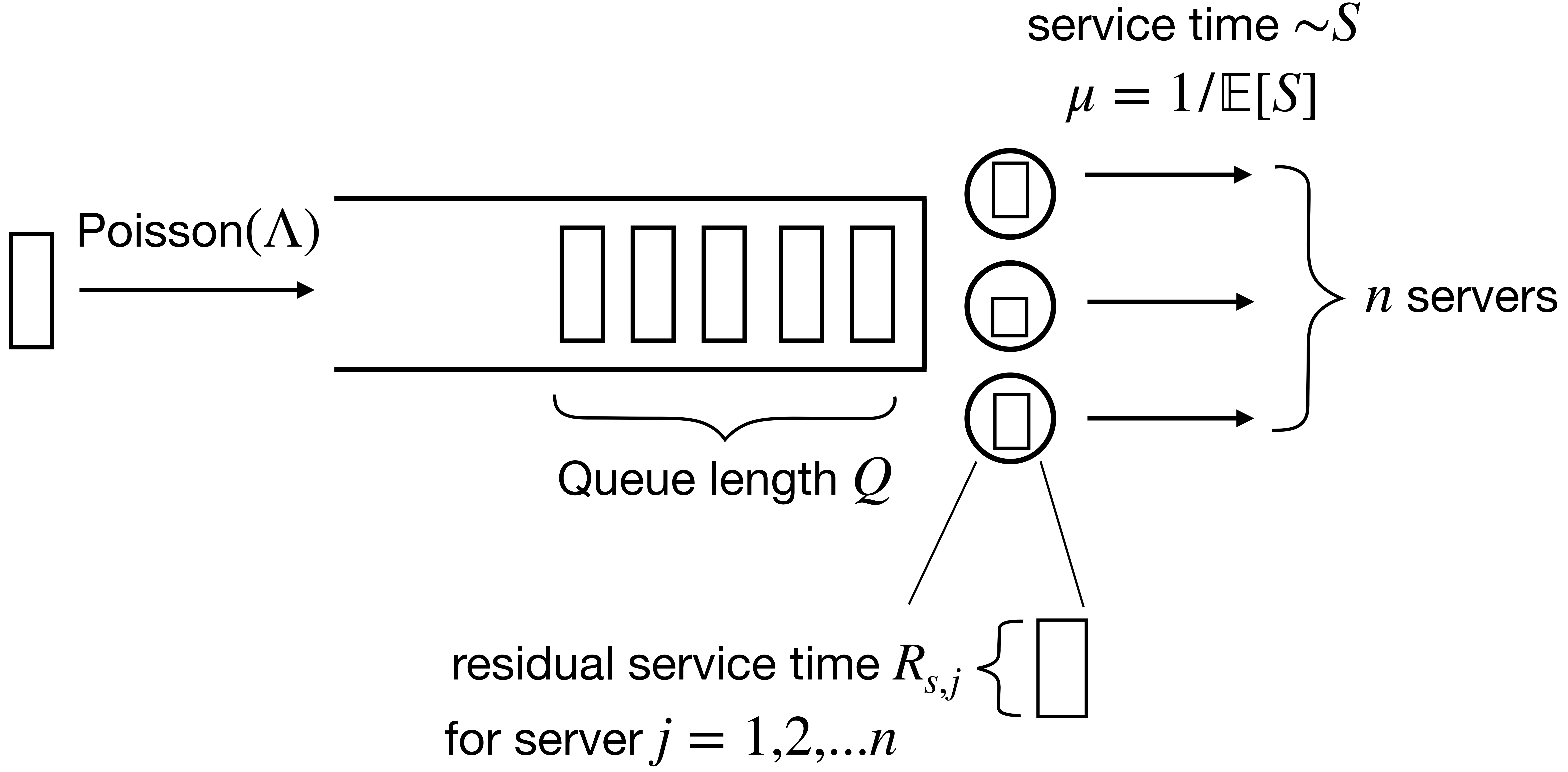
Proof sketch: focus on M/GI/n



Proof sketch: focus on M/GI/n

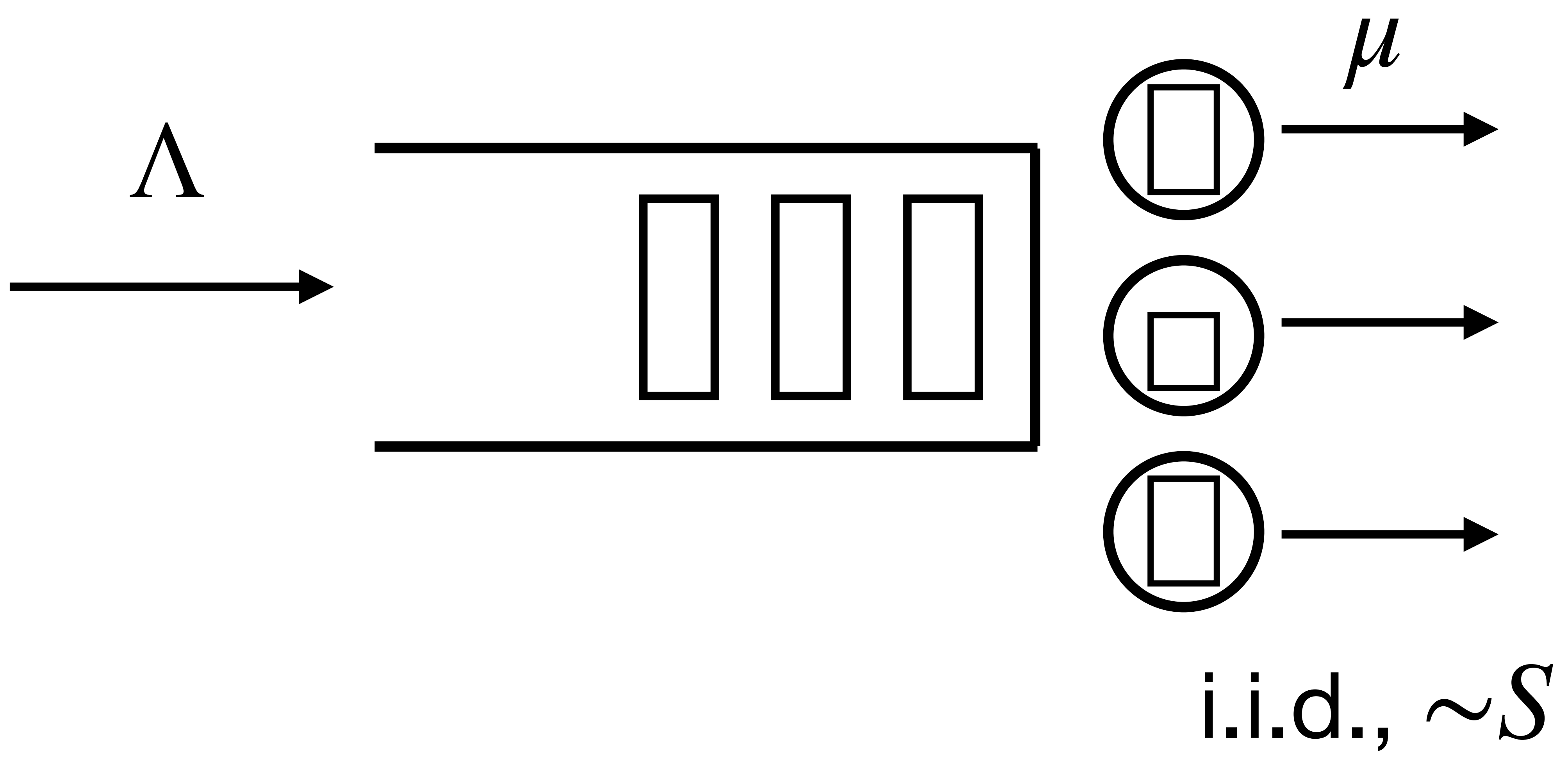


Proof sketch: focus on M/GI/n

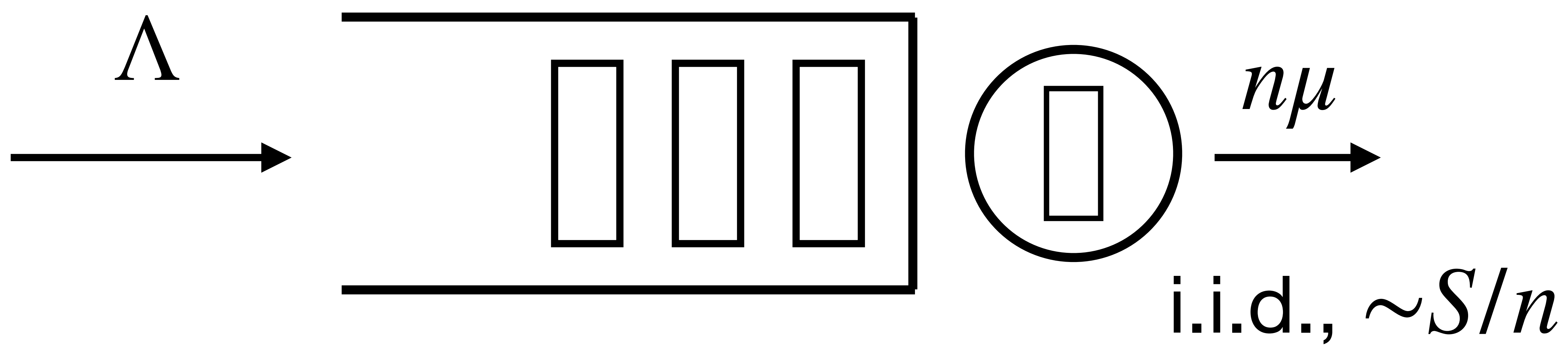


Comparing with M/GI/1

M/GI/n

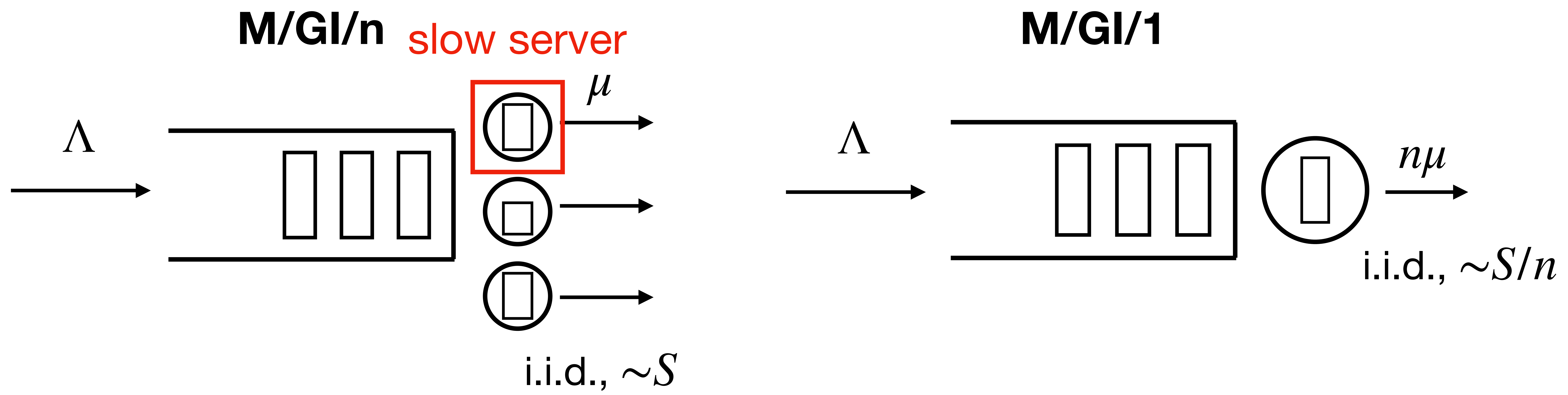


M/GI/1



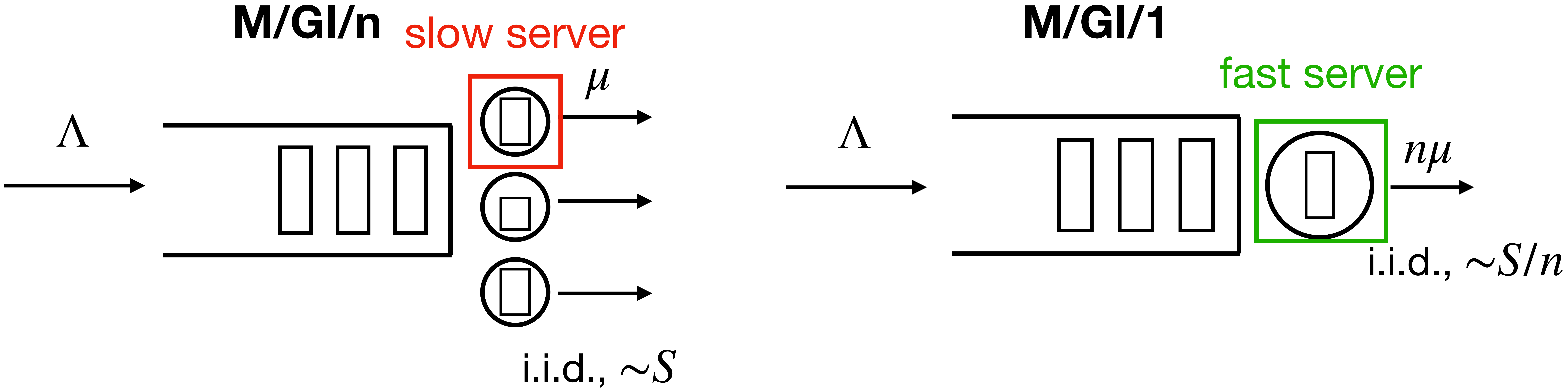
$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{M/GI/1}} + \; ?$$

Comparing with M/GI/1



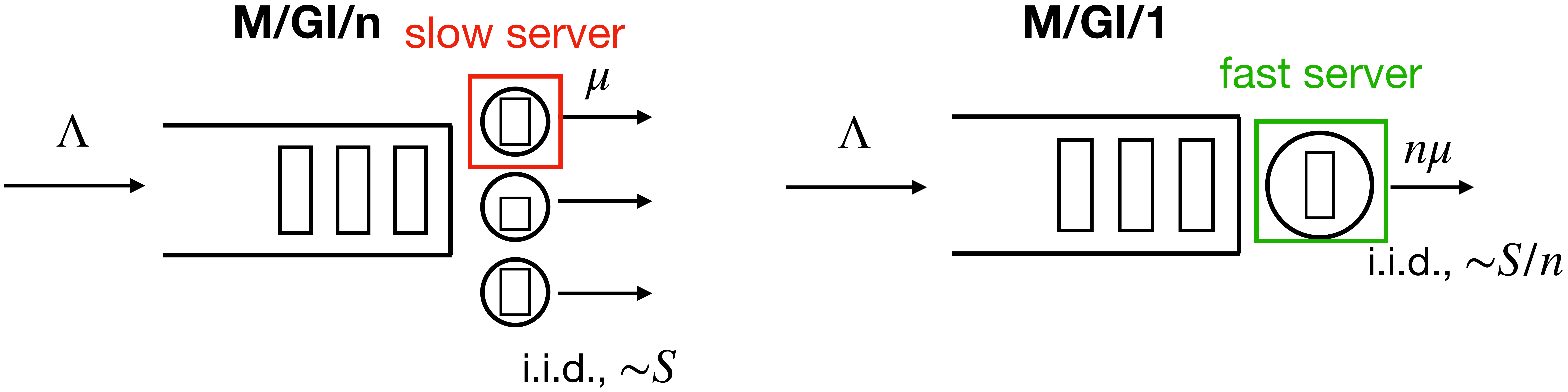
$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{M/GI/1}} + \ ?$$

Comparing with M/GI/1



$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{M/GI/1}} + ?$$

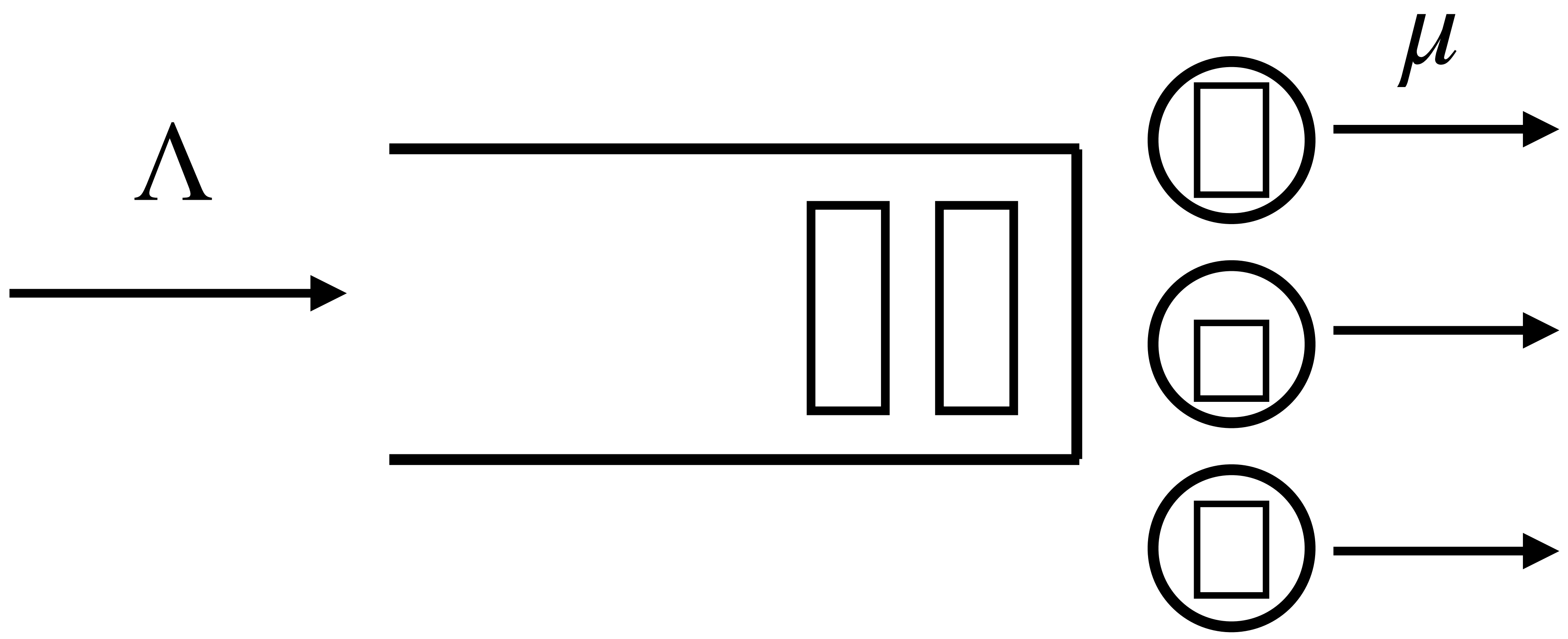
Comparing with M/GI/1



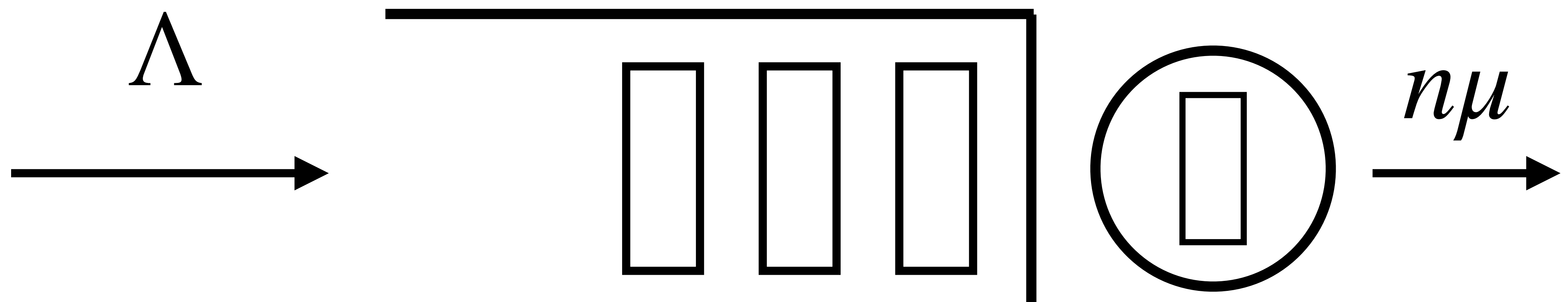
$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{M/GI/1}} + ?$$

How to quantify the difference?

modified M/GI/n

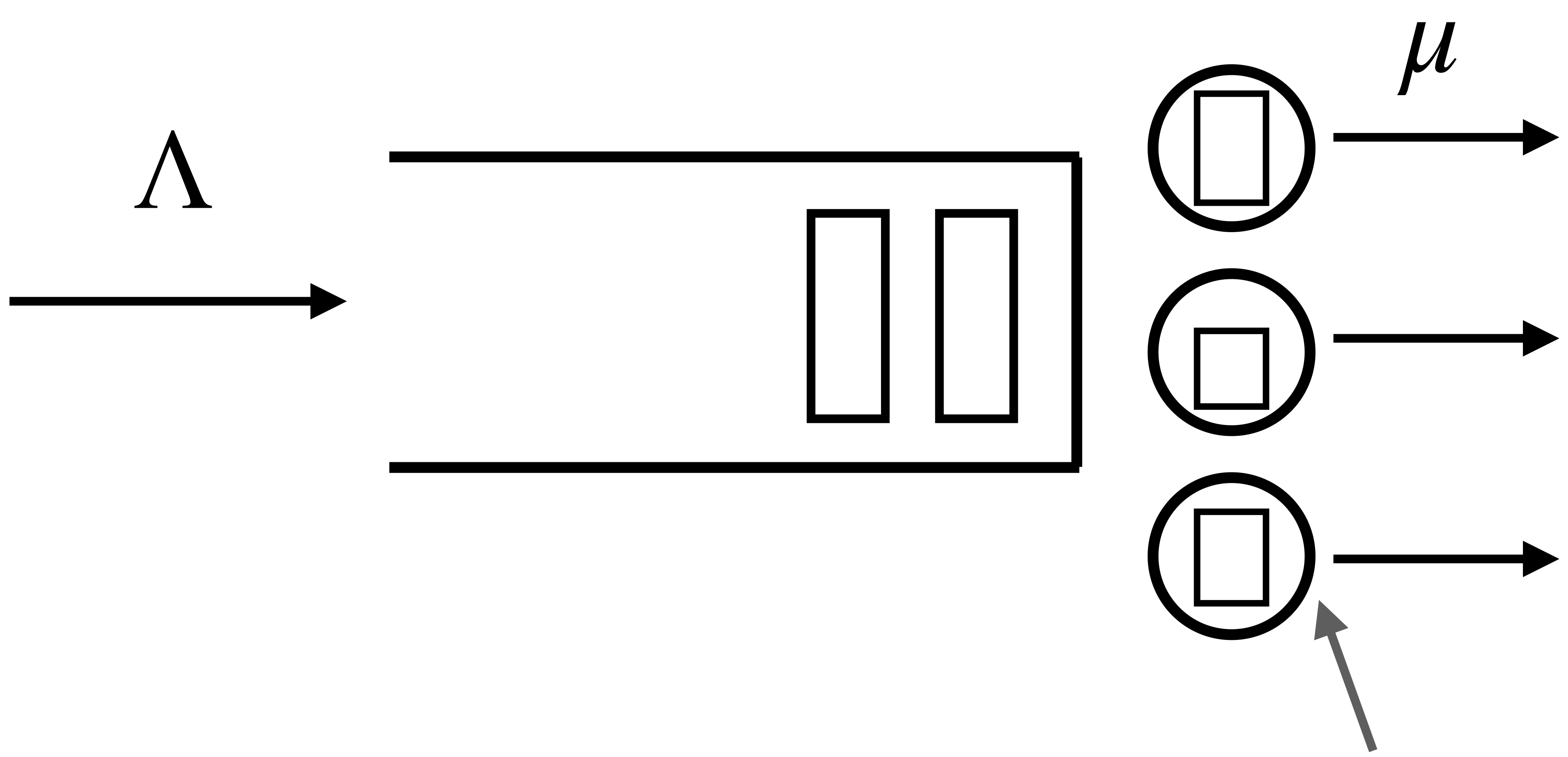


M/GI/1

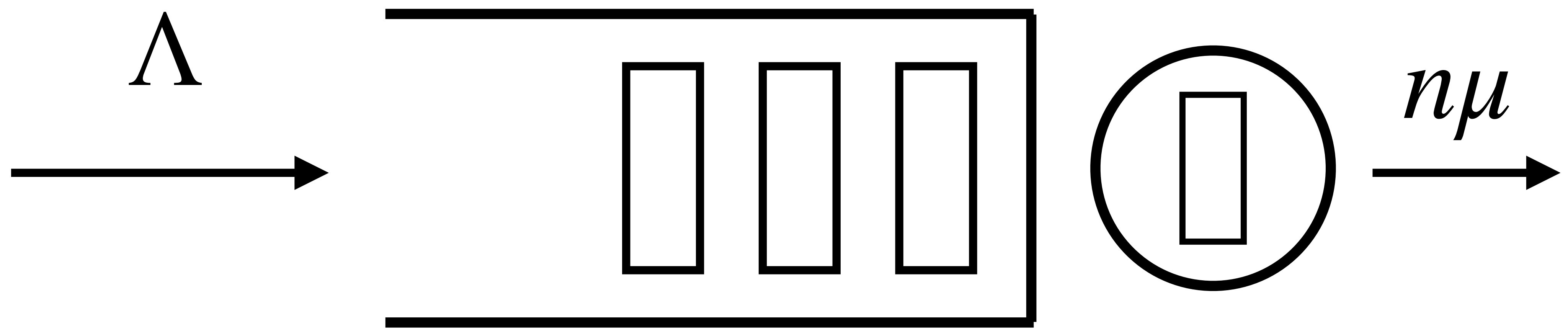


$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}}$$

modified M/GI/n

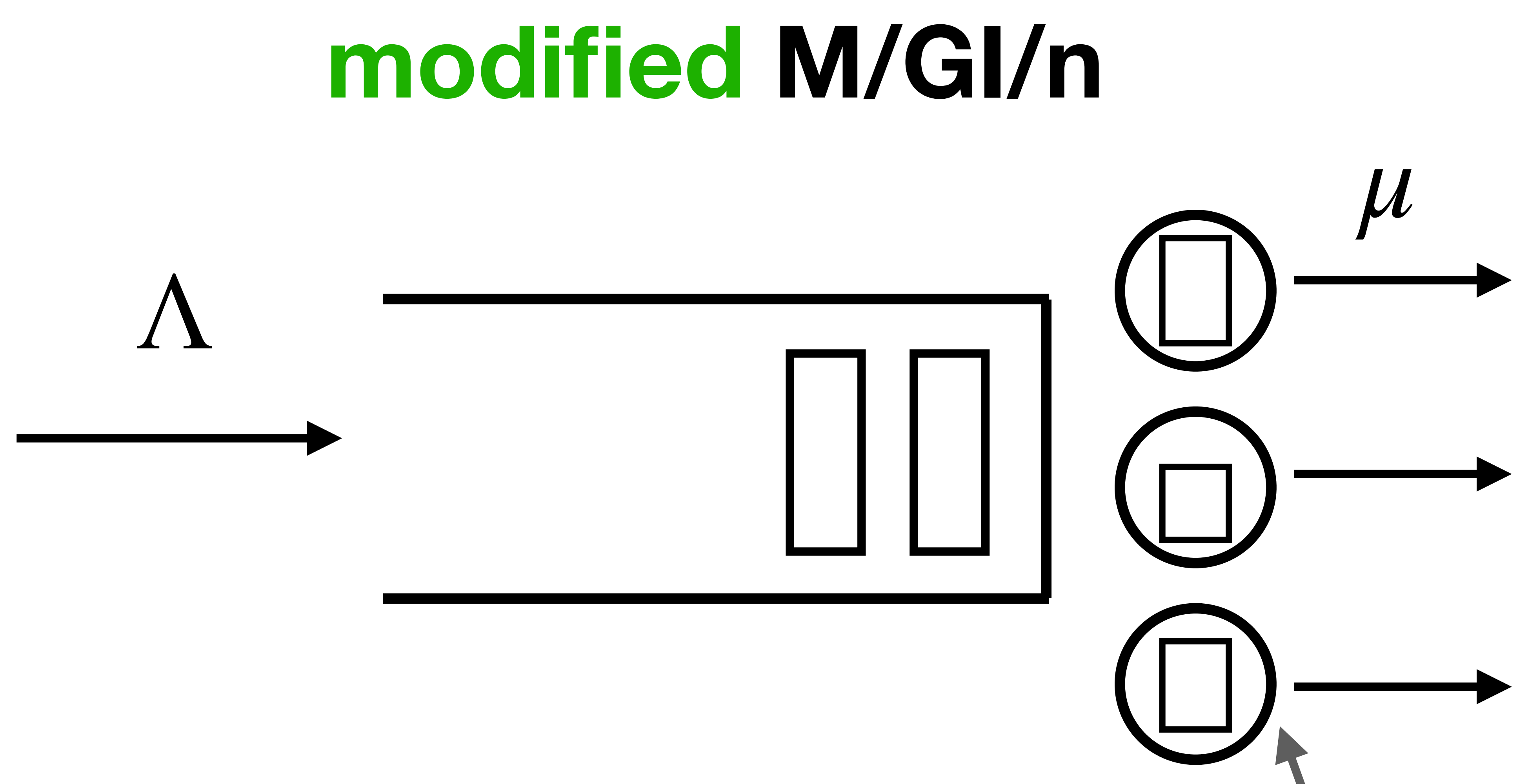


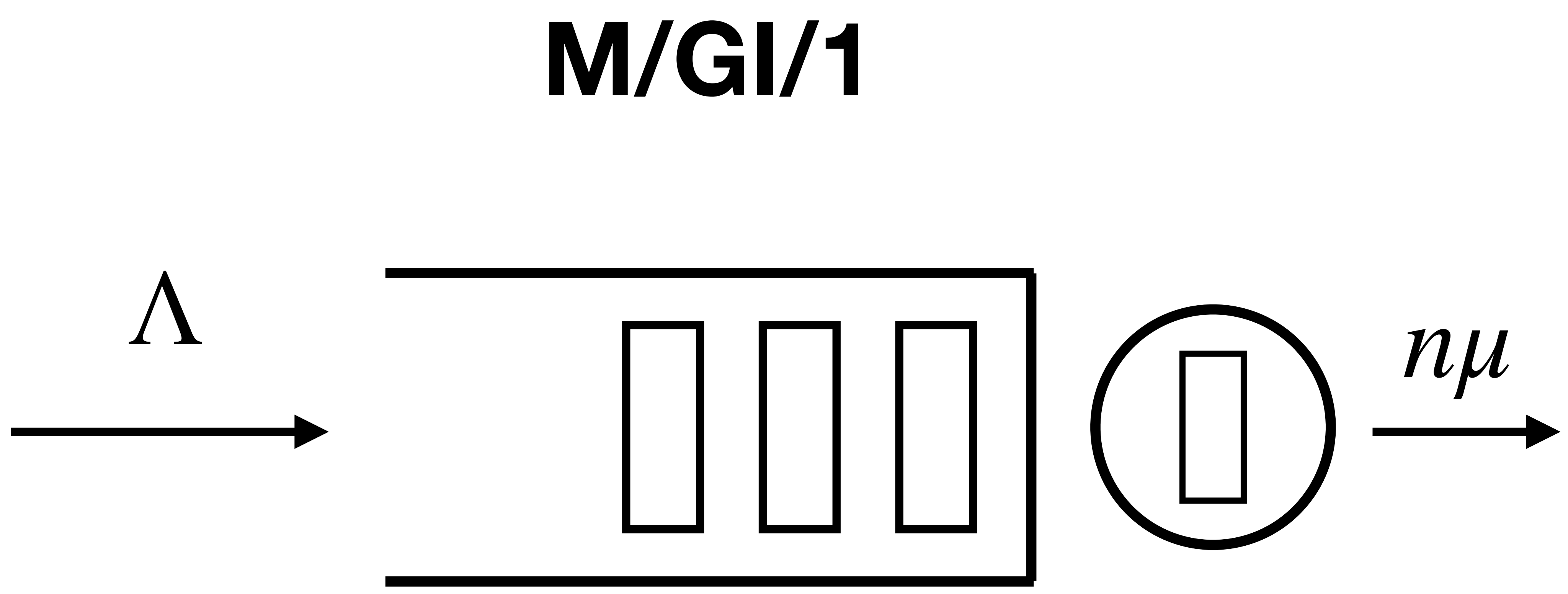
M/GI/1



Modification: when server gets idle and $Q = 0$, add a virtual job to keep it busy

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}}$$



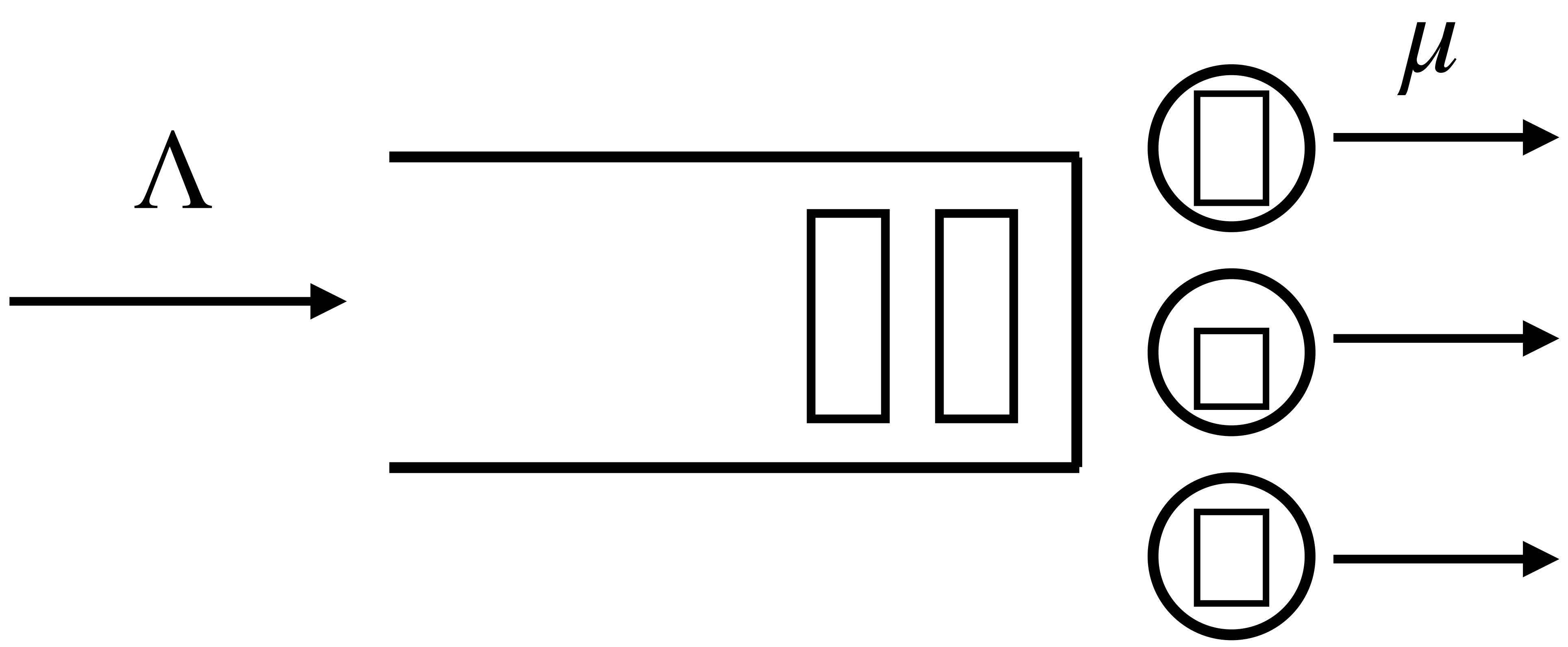


Modification: when server gets idle and $Q = 0$, add a virtual job to keep it busy

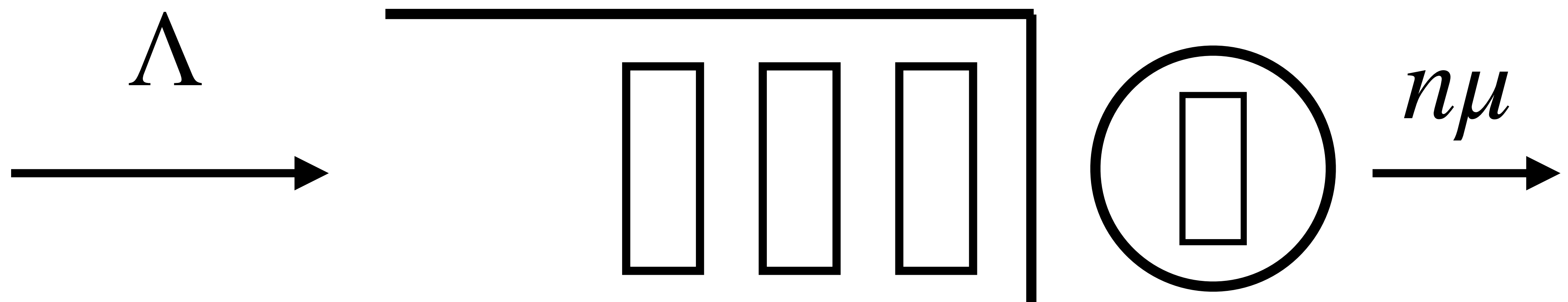
$\Longrightarrow$ completion processes: i.i.d. renewal processes

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}}$$

modified M/GI/n

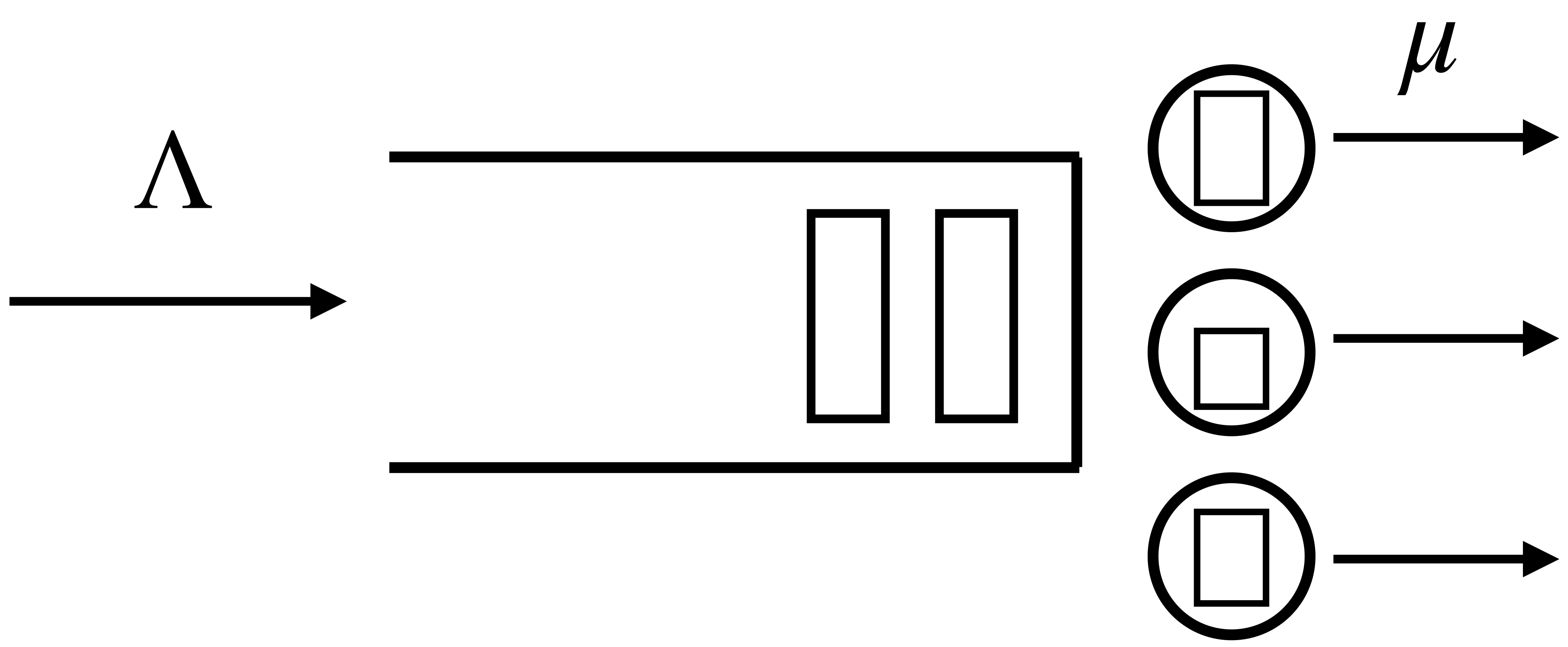


M/GI/1

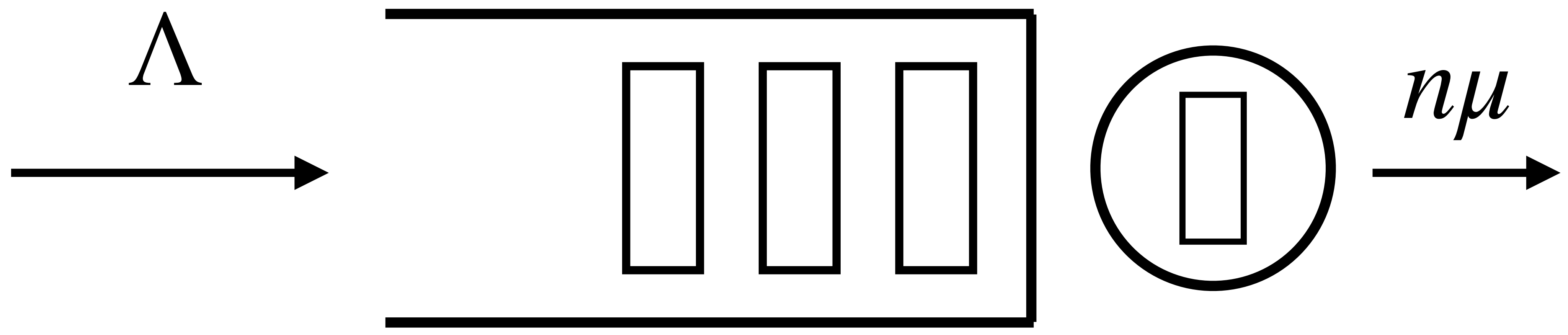


$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}}$$

modified M/GI/n

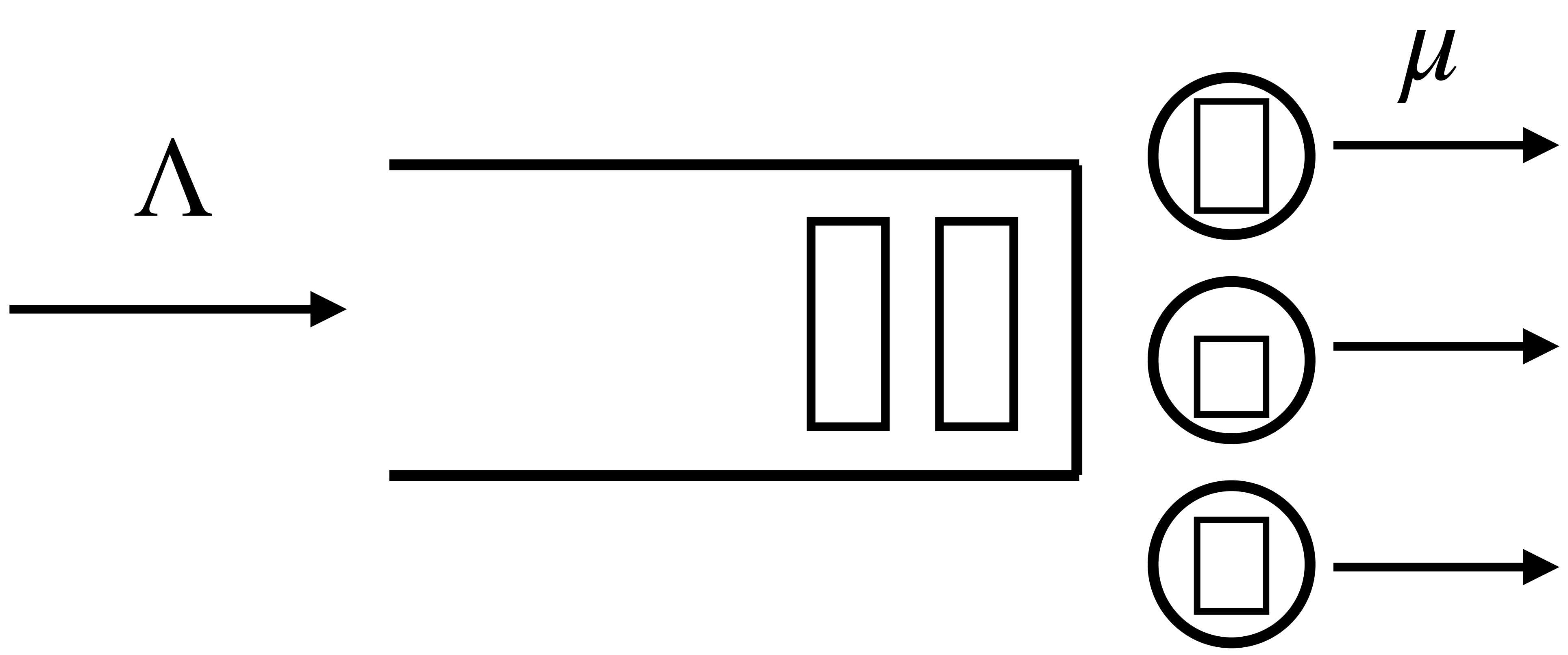


M/GI/1

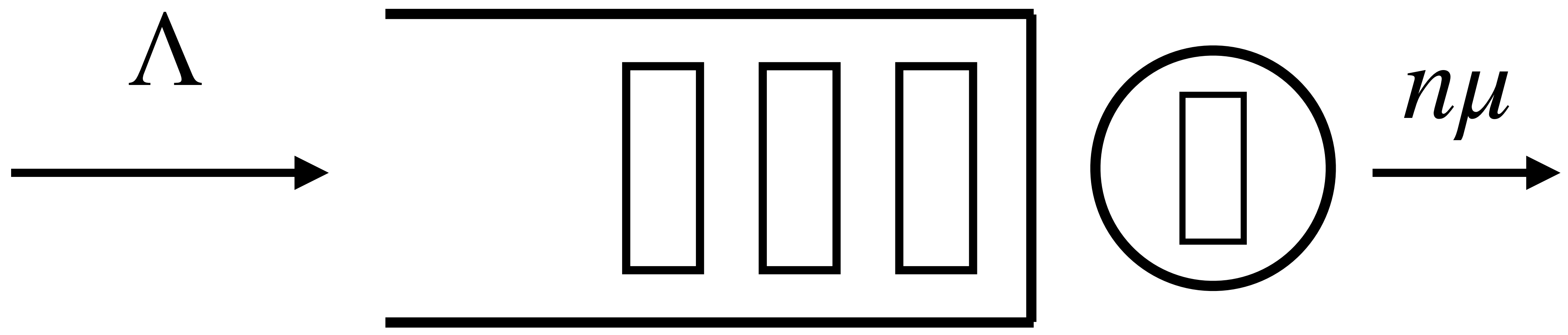


$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n}\Gamma_{s,j}$$

modified M/GI/n



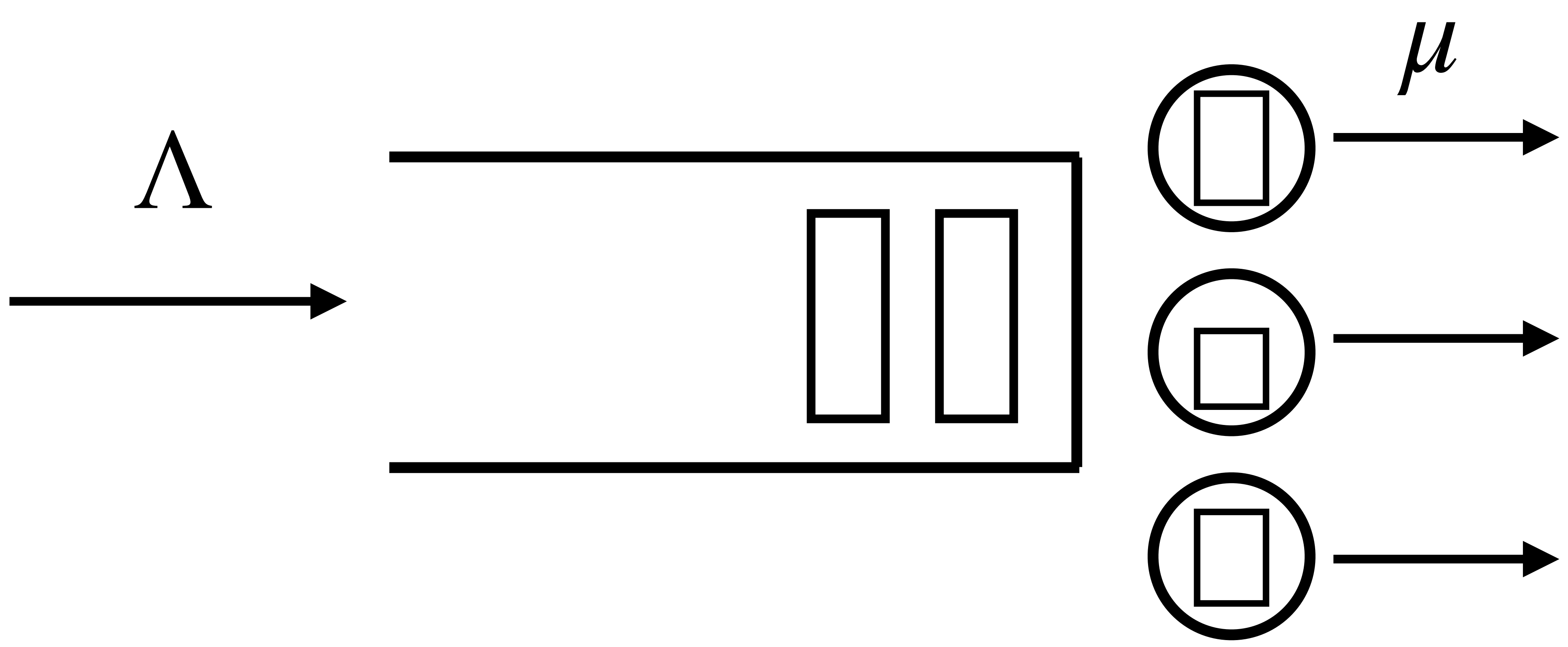
M/GI/1



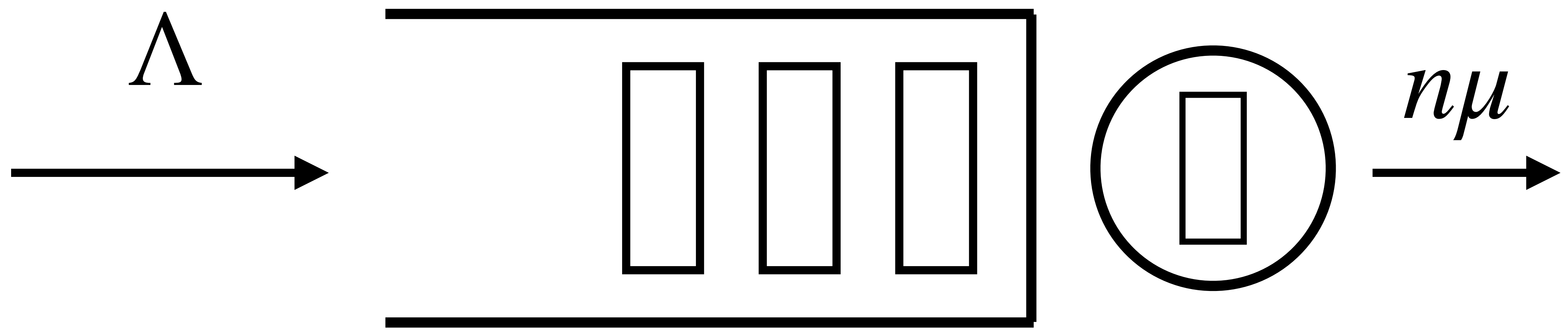
$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

modified M/GI/n



M/GI/1

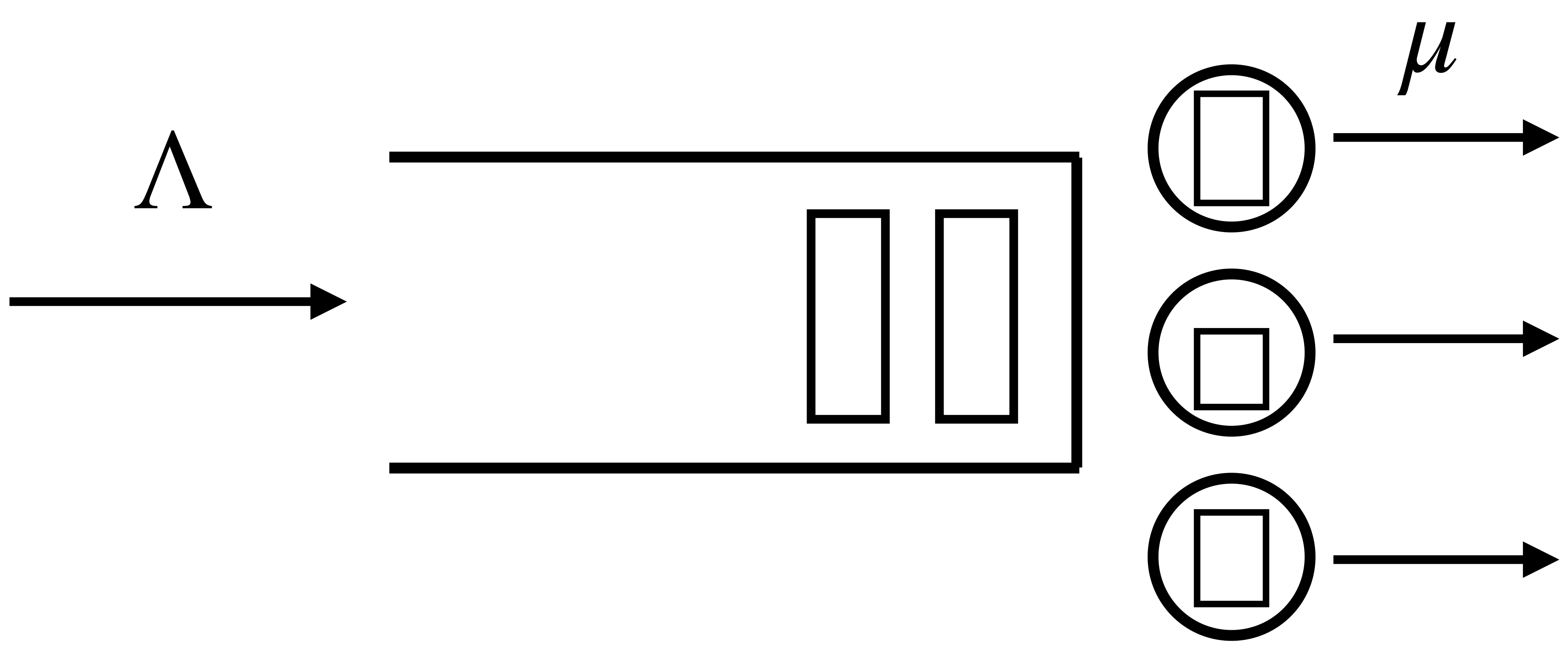


$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n}\Gamma_{s,j}$$

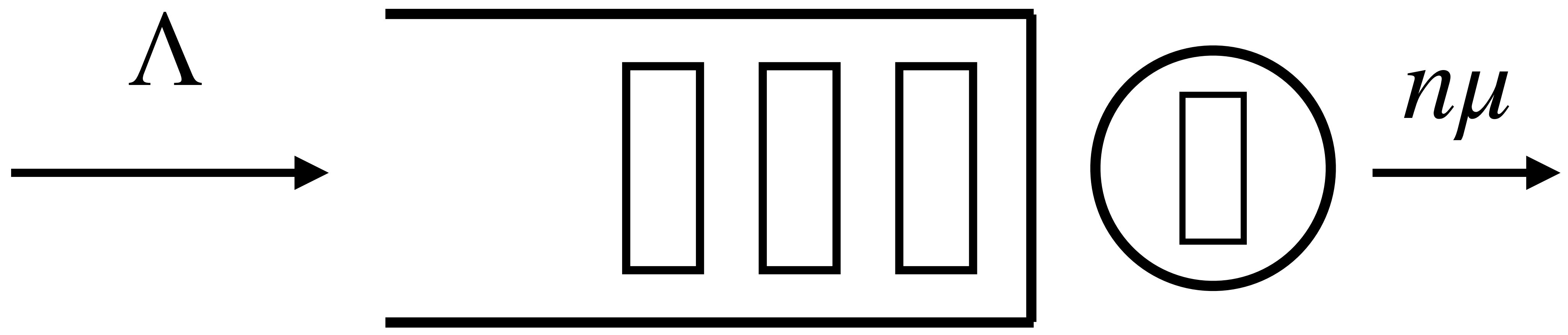
$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

$$\rho = \frac{\Lambda}{n\mu}$$

modified M/GI/n



M/GI/1

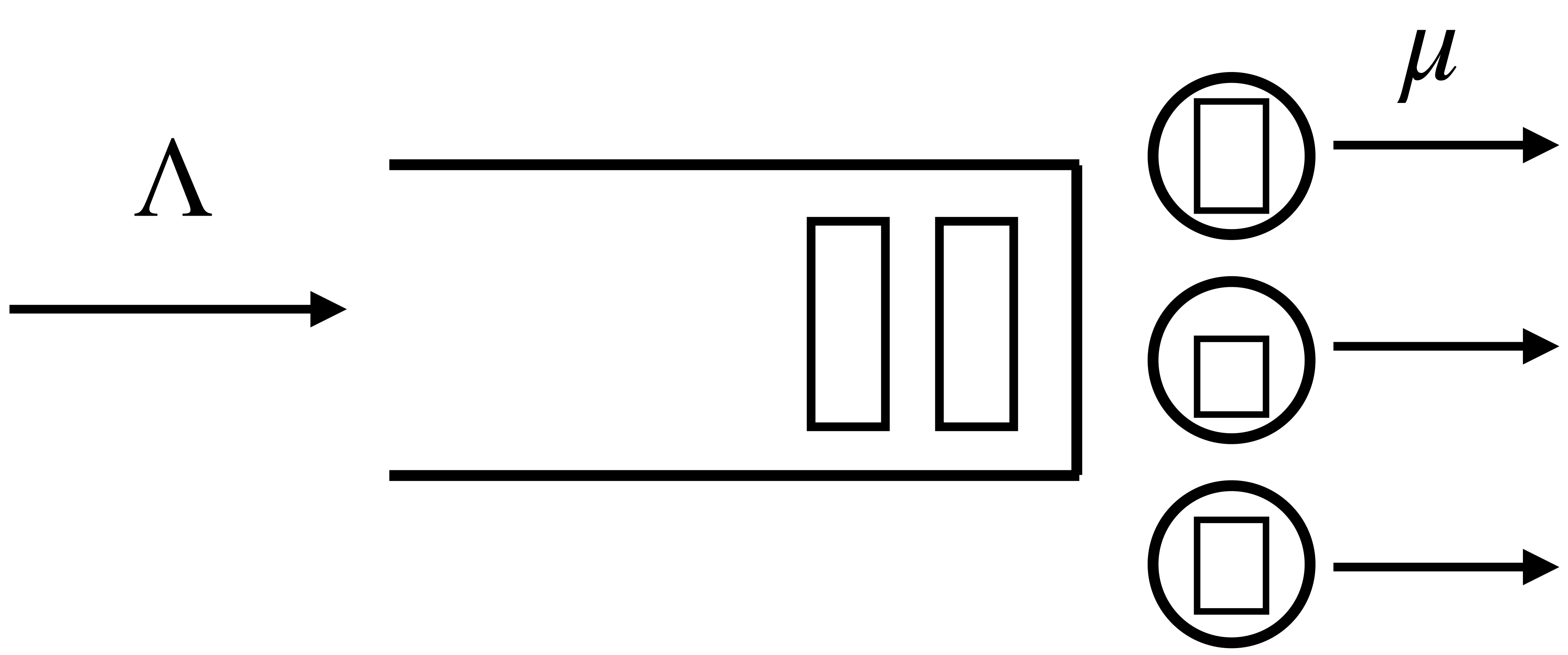


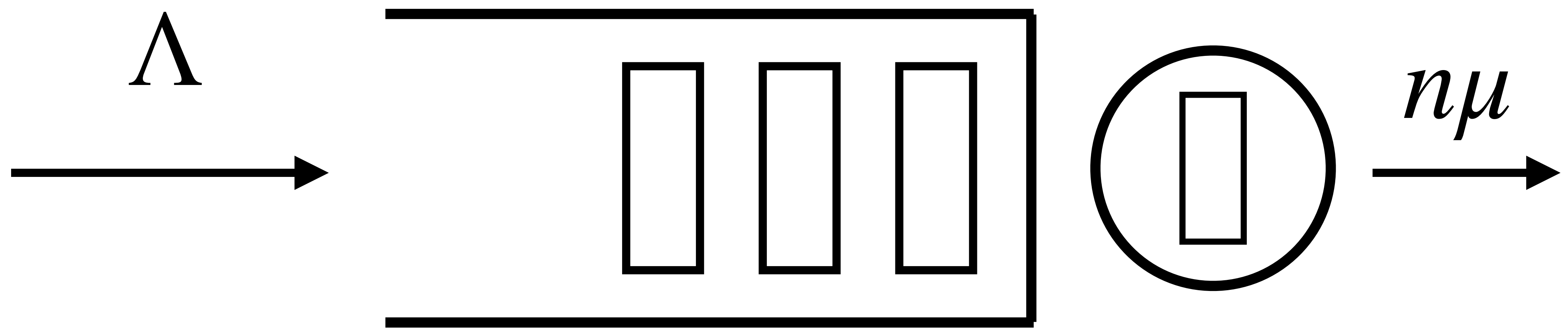
$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

modified M/GI/n **M/GI/1**





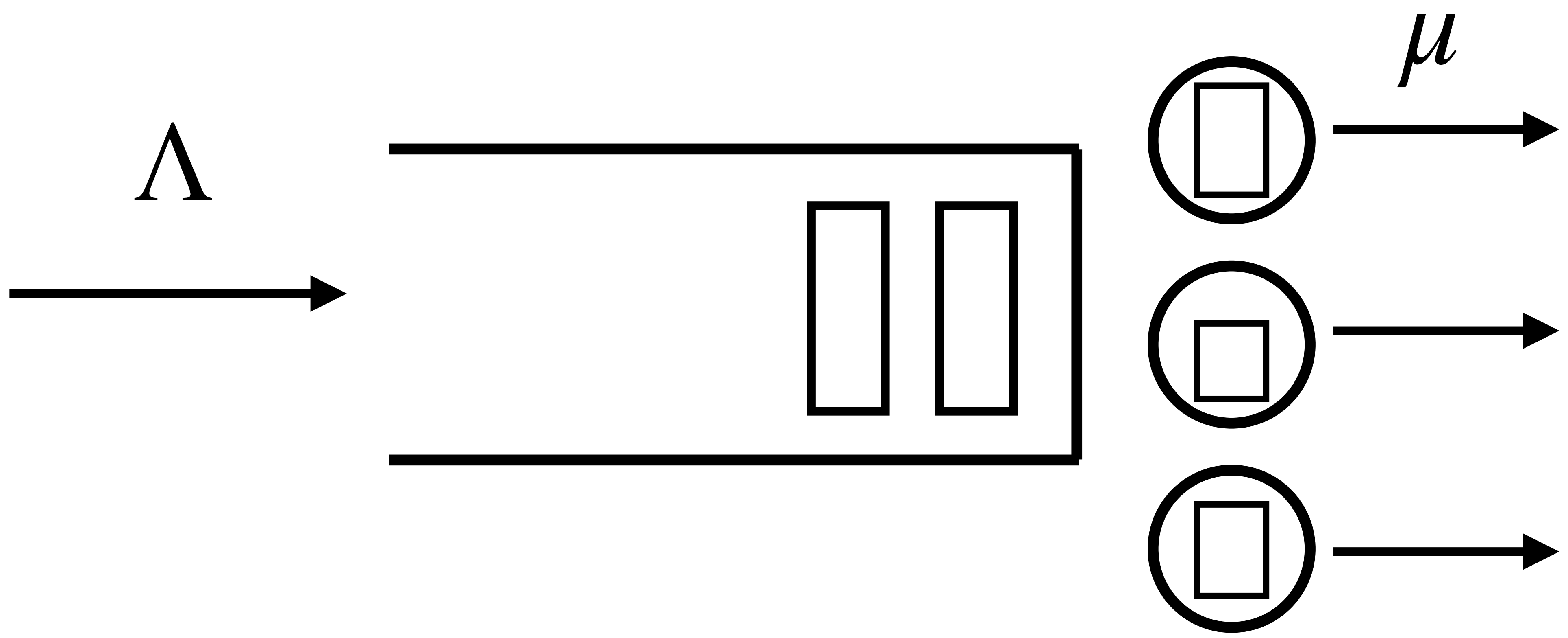
$$\mathbb{E}[Q]^{\text{M/GI/n}} \le \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

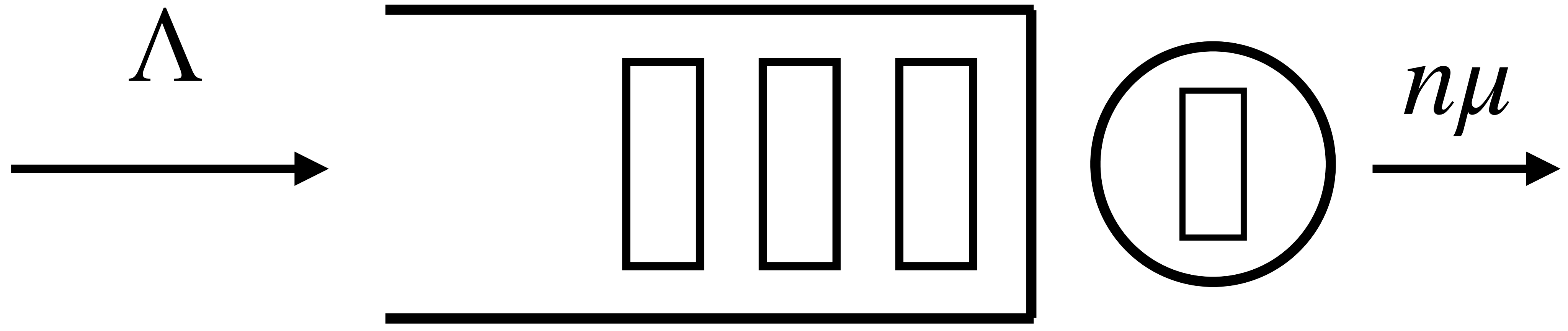
$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

"Palm exp" at completions

modified M/GI/n **M/GI/1**





$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

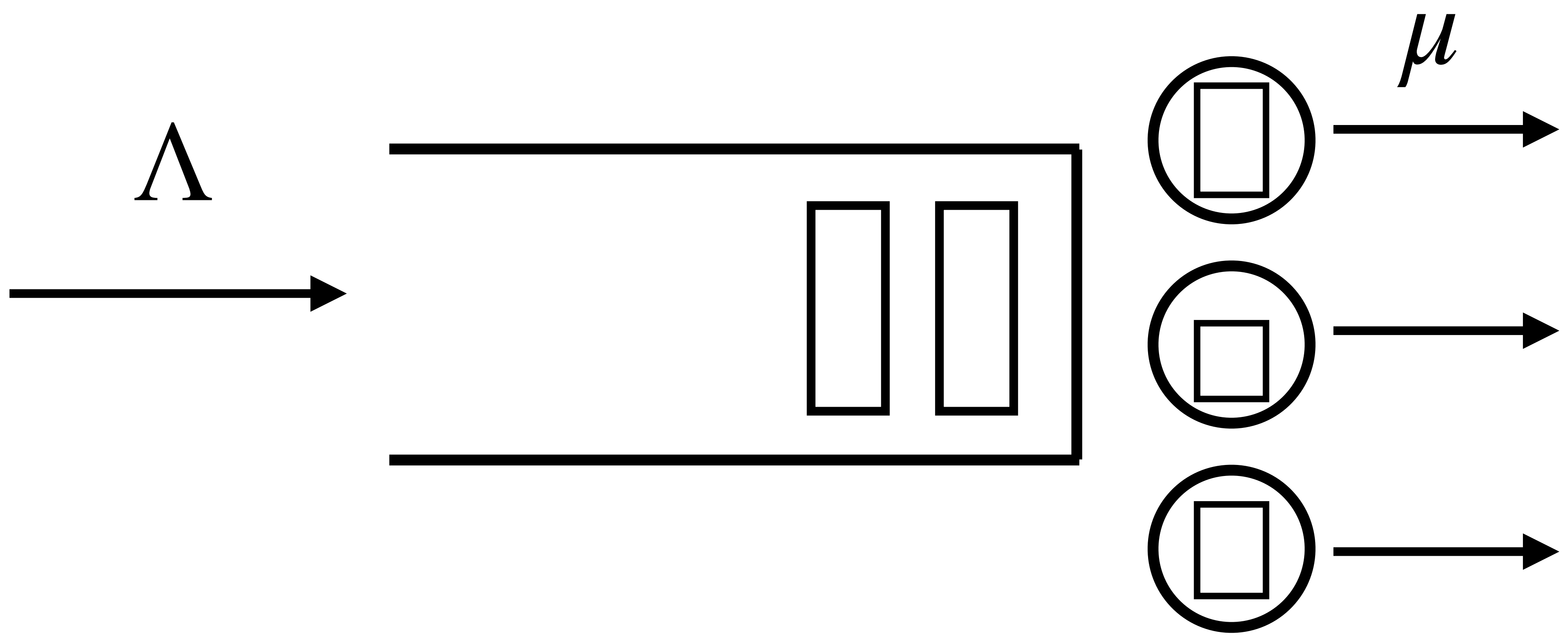
$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

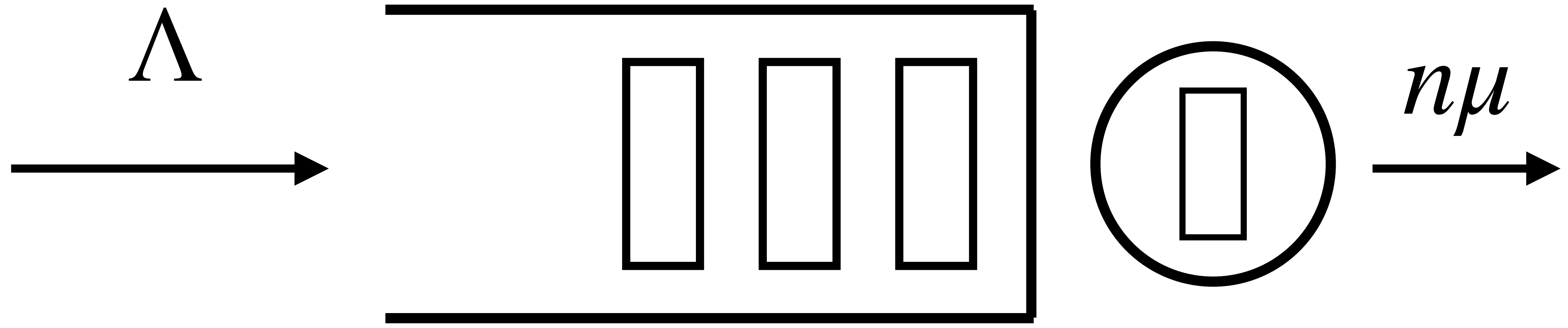
"Palm exp" at completions

res. service time

modified M/GI/n



M/GI/1

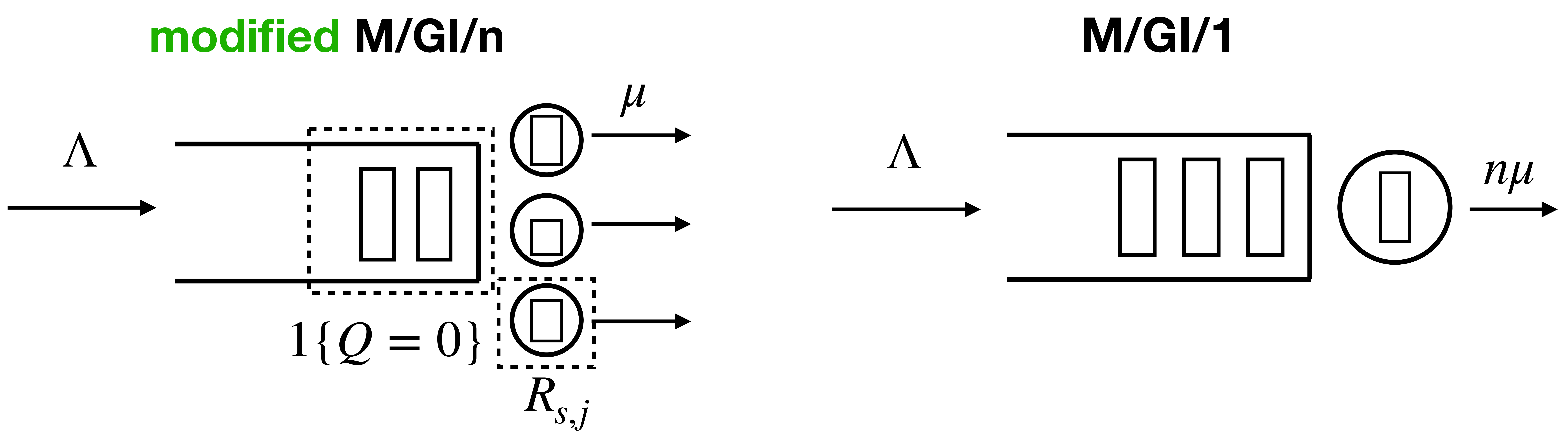


$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$



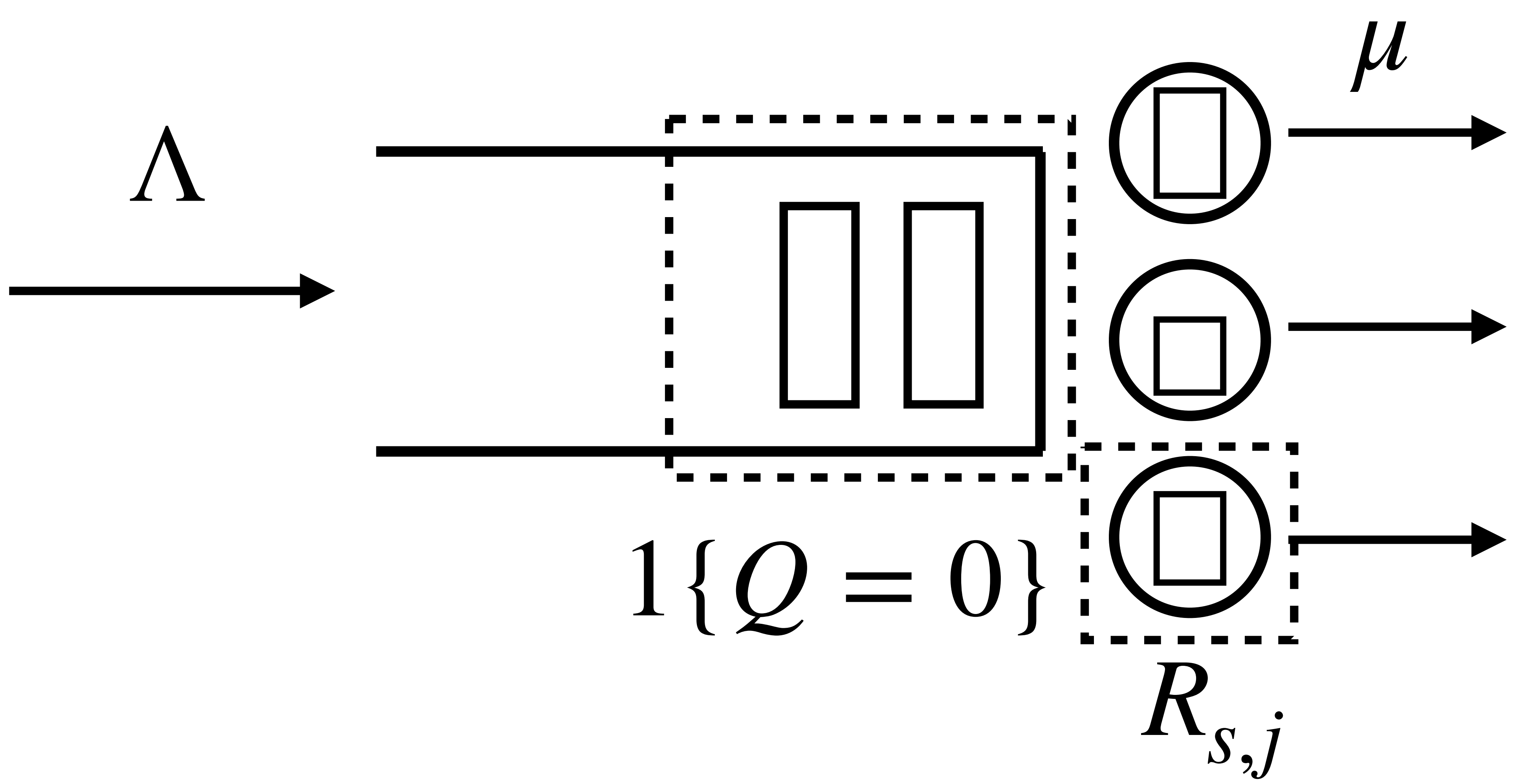
$$\mathbb{E}[Q]^{\text{M/GI/n}} \le \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \Gamma_{s,j}$$

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

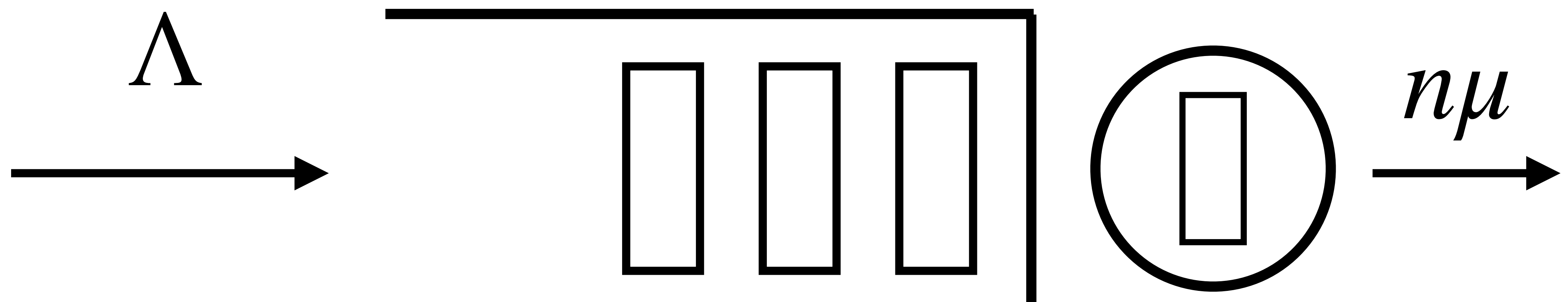
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

modified M/GI/n



M/GI/1



$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n}\Gamma_{s,j}$$

By "Basic Adjoint Relationship" (BAR)

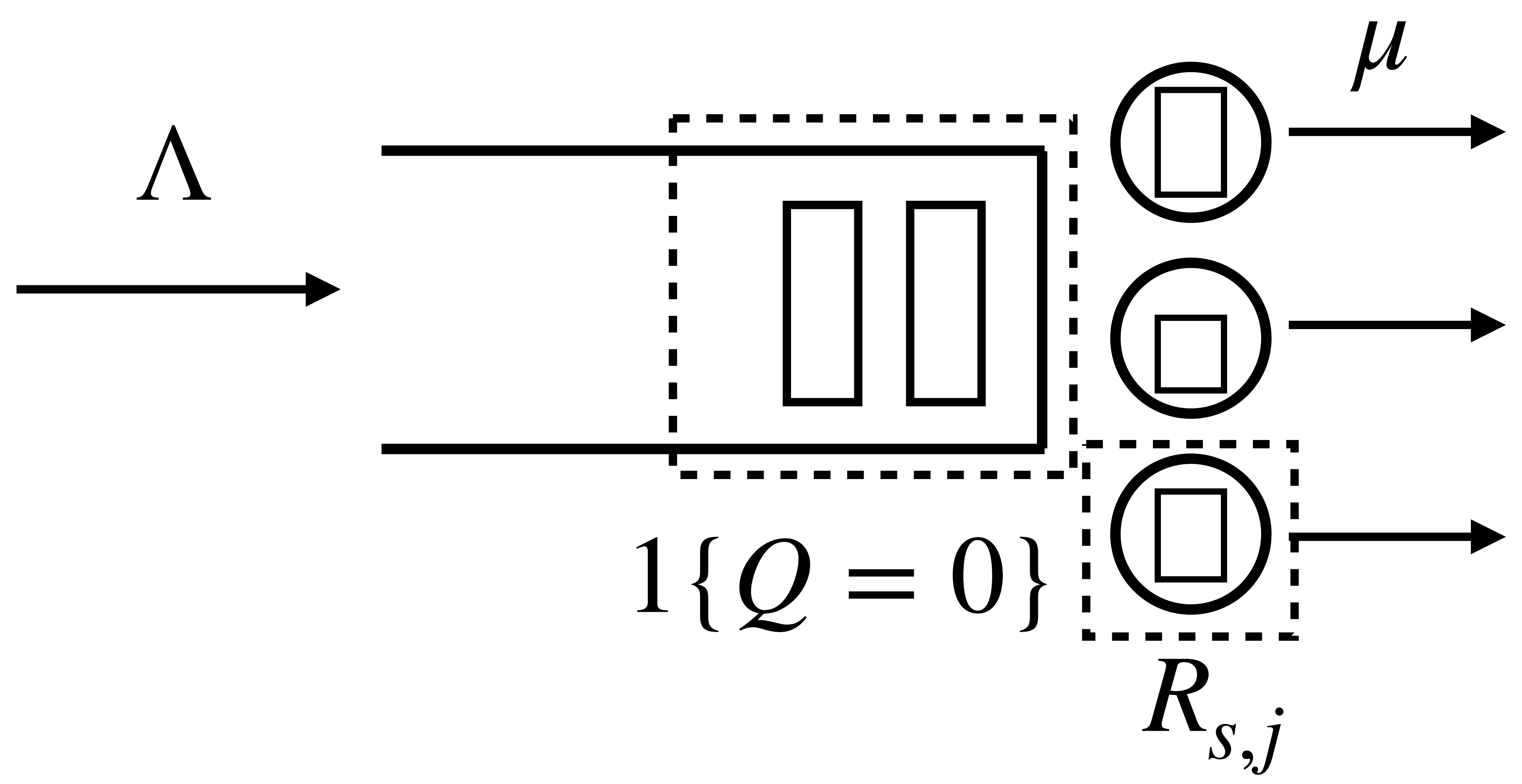
see, e.g., [Miyazawa 94, 15]
[Braverman et al. 2017, 2024]

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

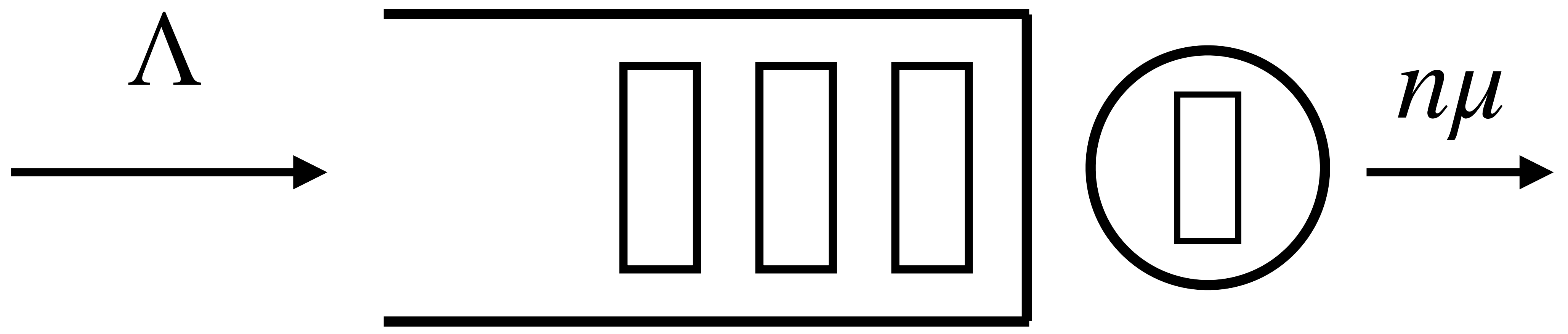
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

modified M/GI/n



M/GI/1



$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\boxed{\sum_{j=1}^{n}\Gamma_{s,j}}$$

By “Basic Adjoint Relationship” (BAR)

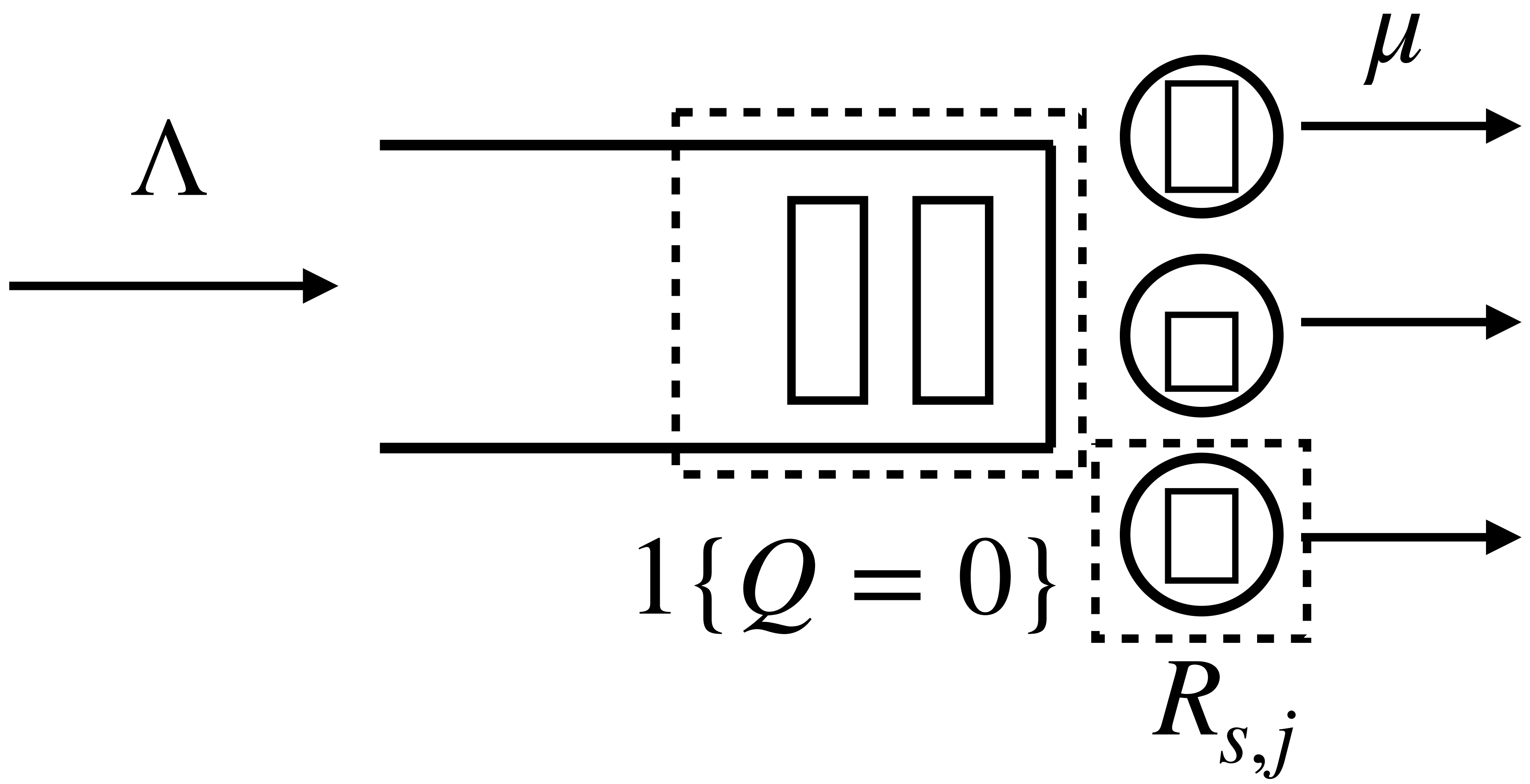
see, e.g., [Miyazawa 94, 15] [Braverman et al. 2017, 2024]

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

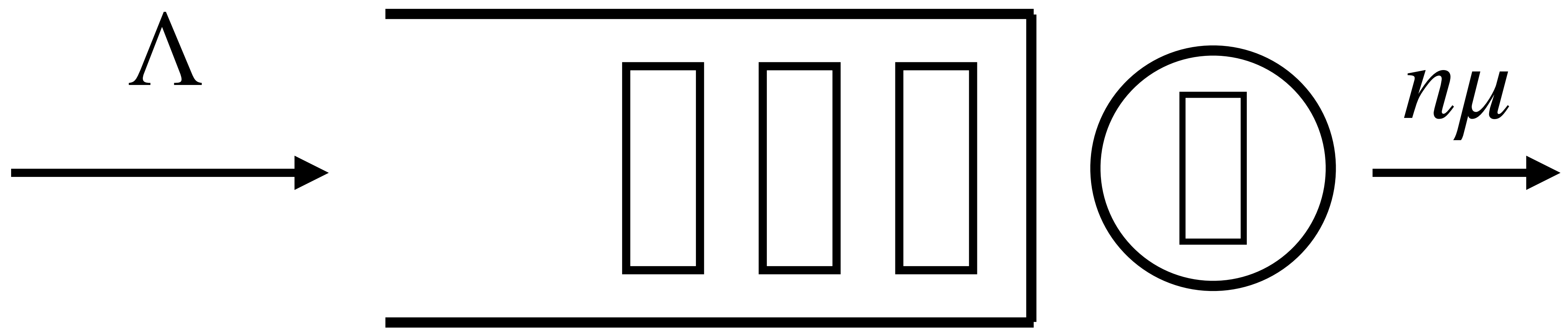
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$\approx$ **Covariance**$(1\{Q=0\}, R_{s,j})$

modified M/GI/n



M/GI/1



$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\boxed{\sum_{j=1}^{n}\Gamma_{s,j}}$$

Bounded by a constant ?!

By “Basic Adjoint Relationship” (BAR)

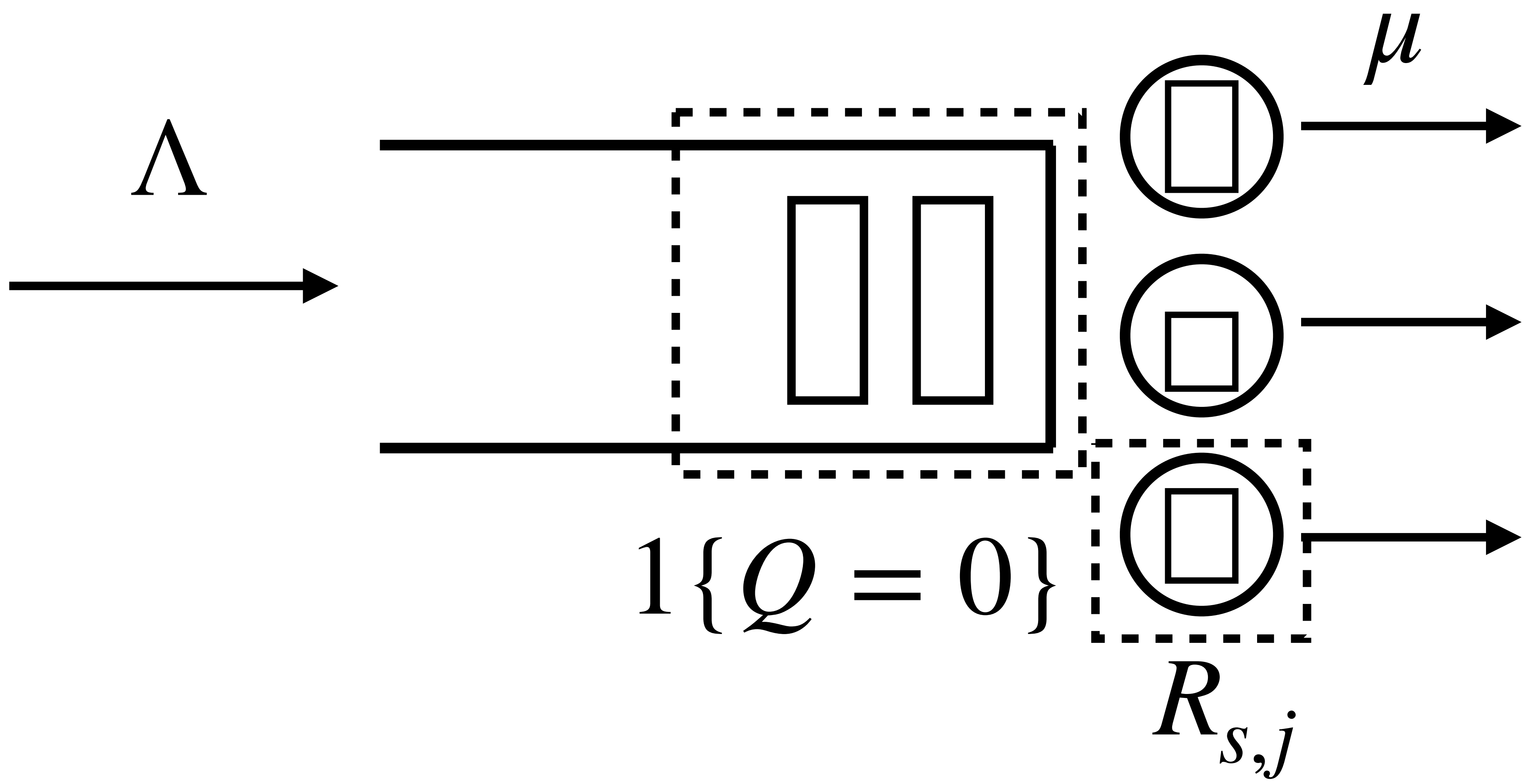
see, e.g., [Miyazawa 94, 15]
[Braverman et al. 2017, 2024]

$$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$$

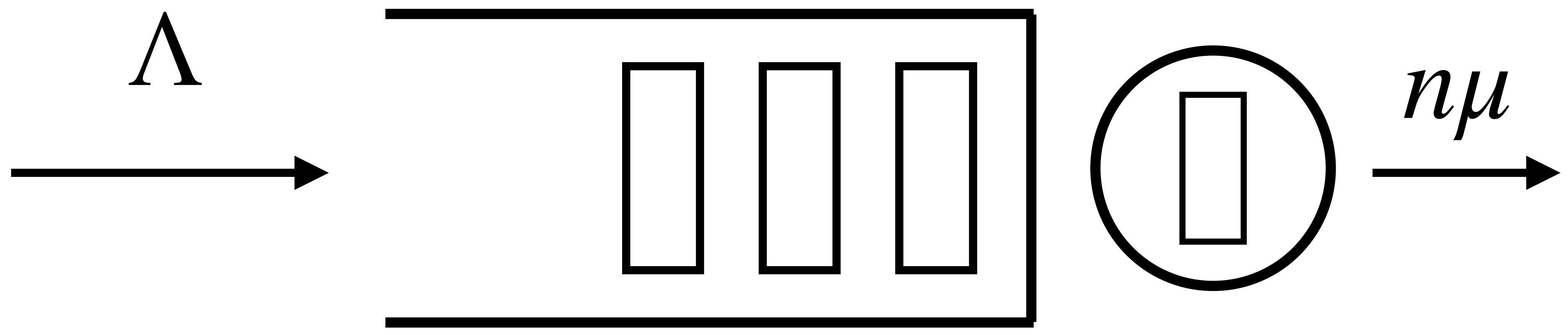
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

modified M/GI/n



M/GI/1



$$\mathbb{E}[Q]^{\text{M/GI/n}} \le \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**

By "Basic Adjoint Relationship" (BAR)

see, e.g., [Miyazawa 94, 15] [Braverman et al. 2017, 2024]

$\approx \mathbb{E}[Q]^{\text{M/GI/1}}$

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$\approx$ **Covariance**$(1\{Q=0\}, R_{s,j})$

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**

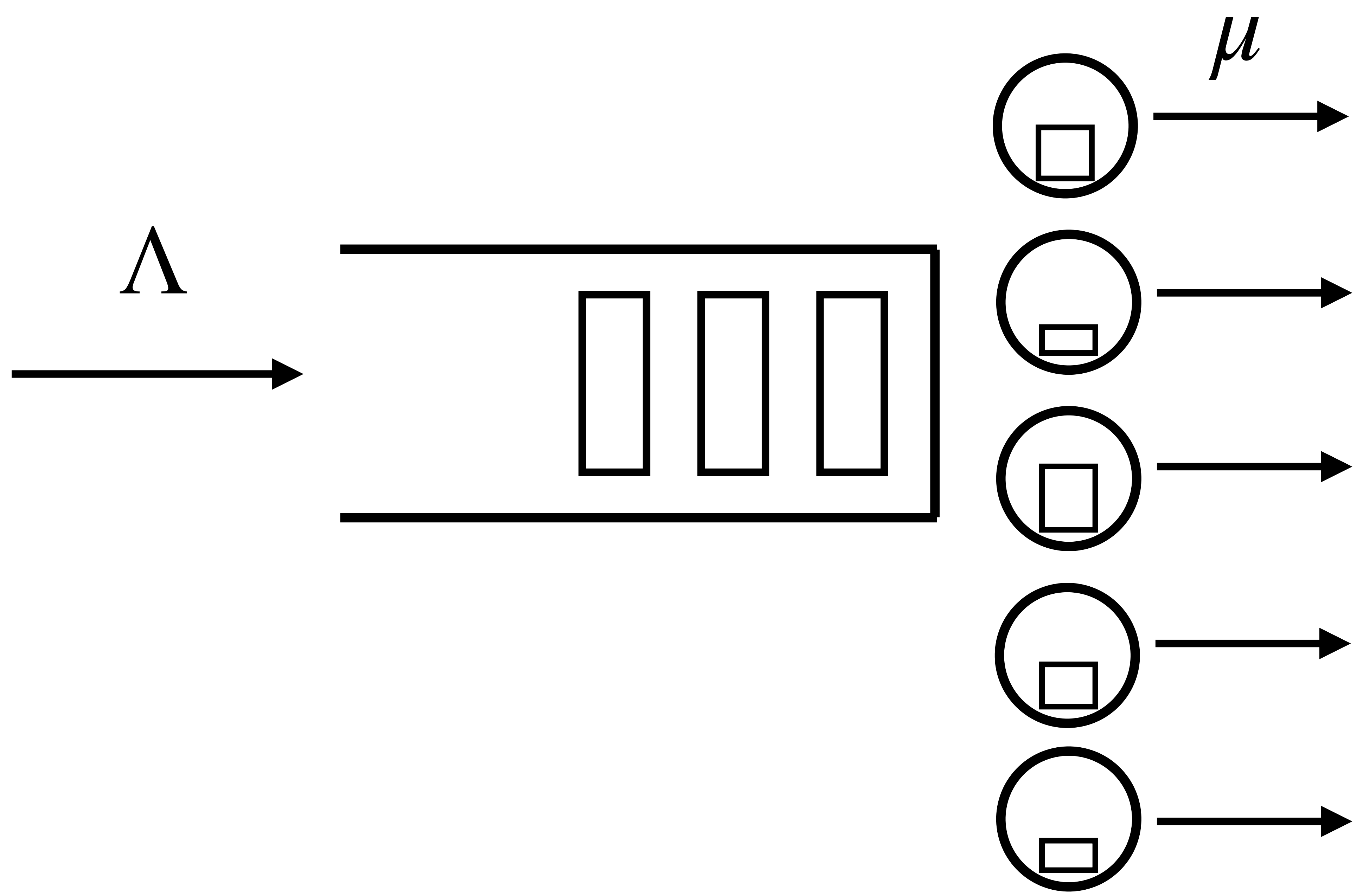
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**



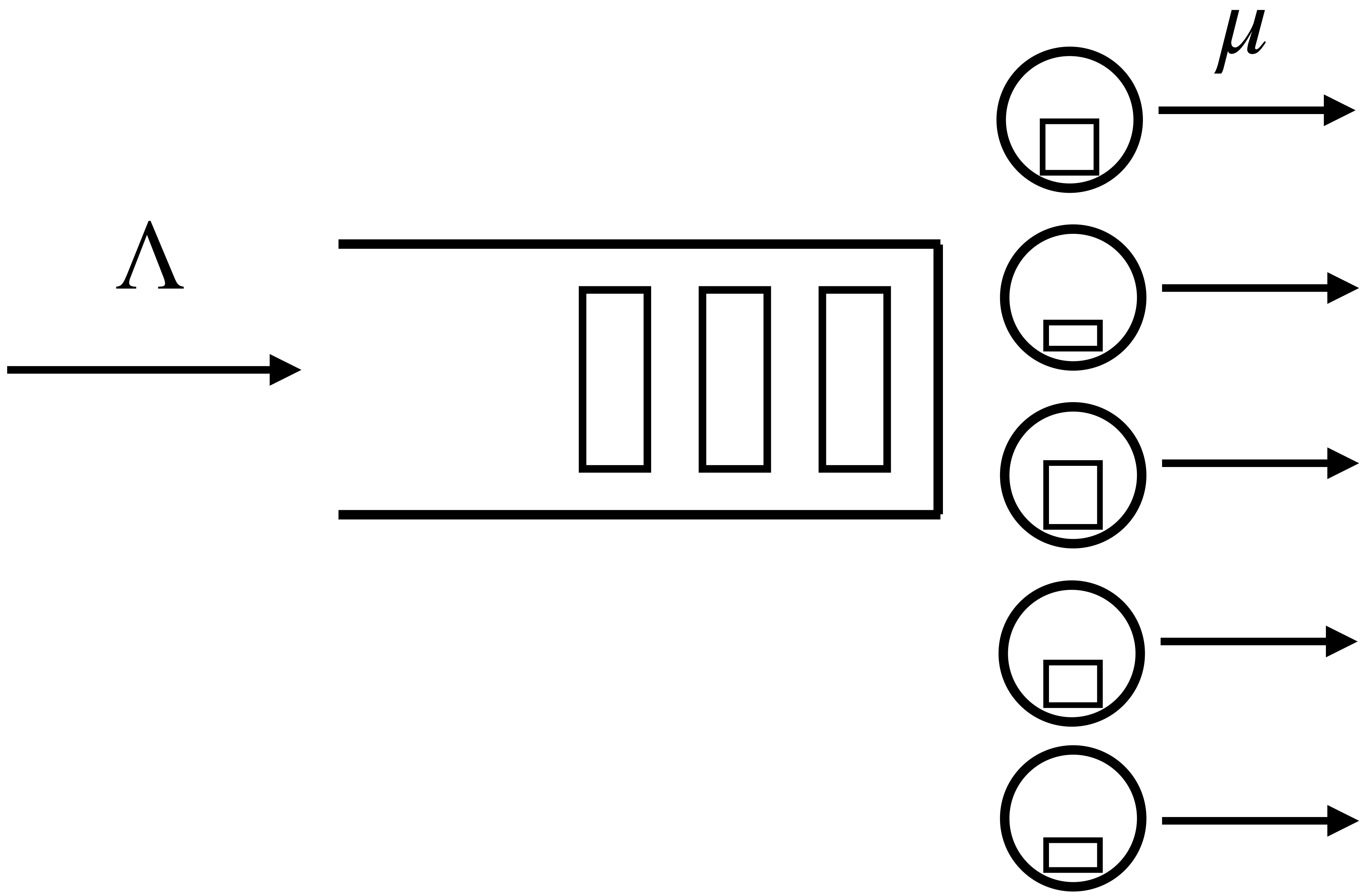
$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**



$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

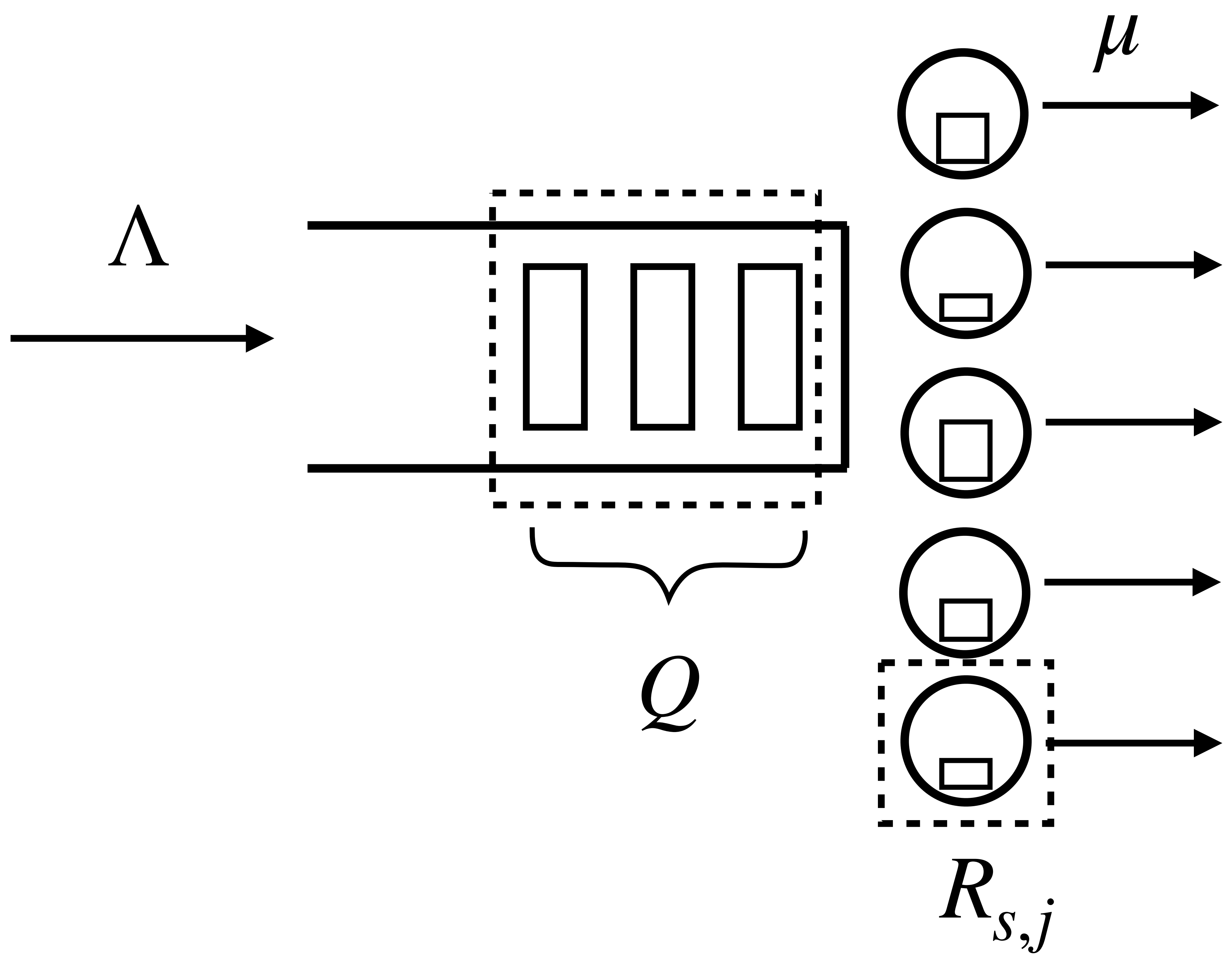
$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**



$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

Key challenge: understanding covariance

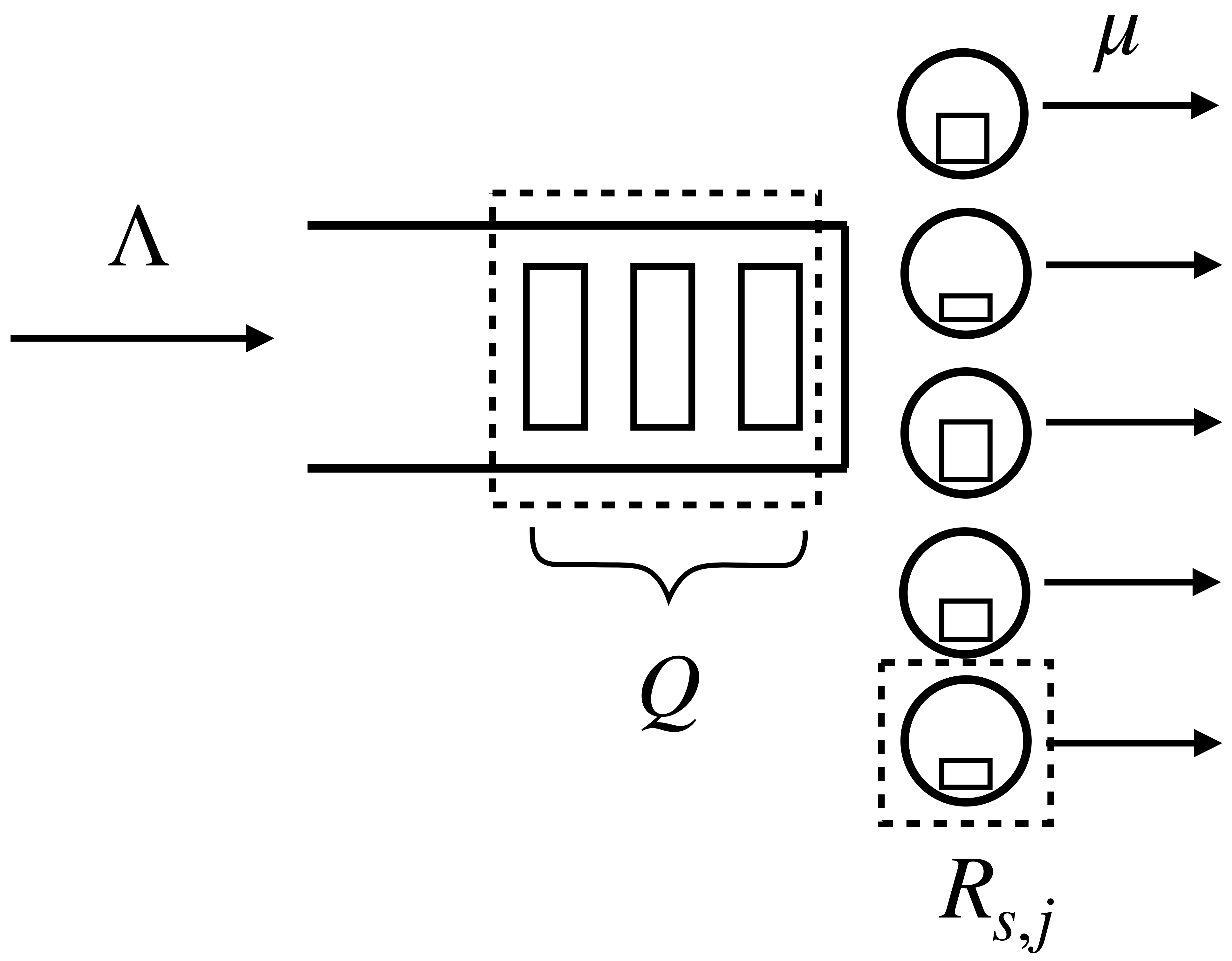
$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**

$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

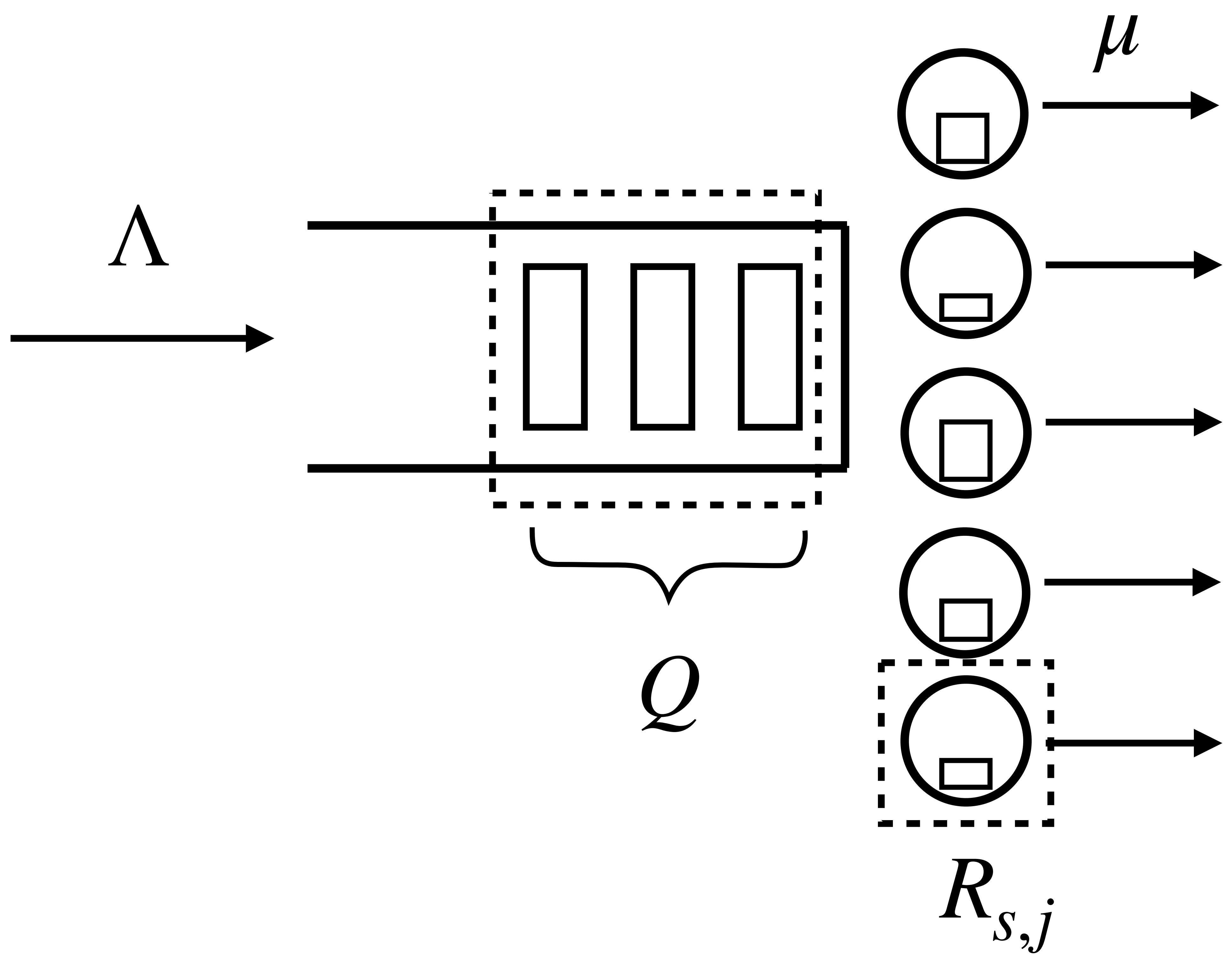


How much correlation?

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**



$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

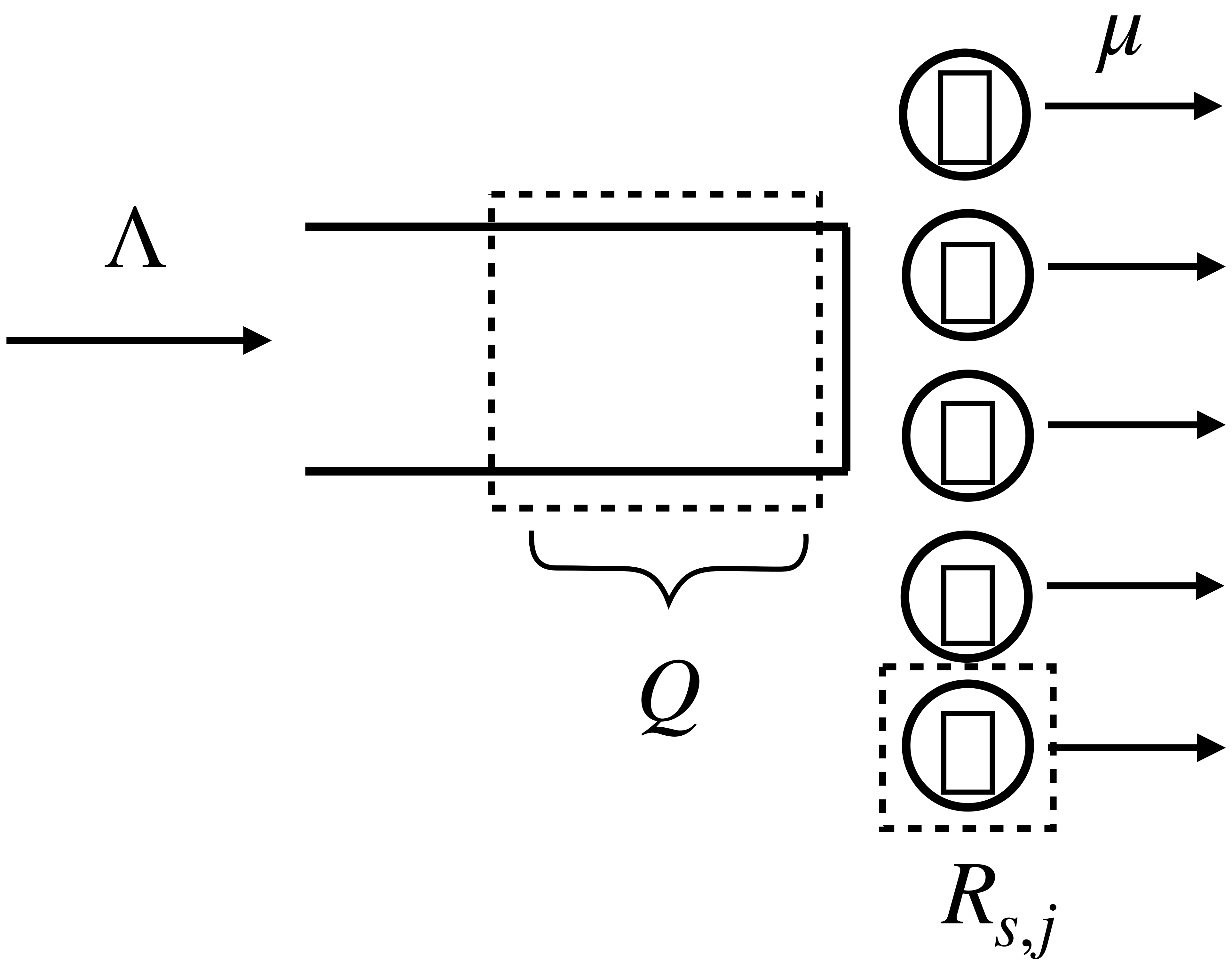
How much correlation?

Could $R_{s,j}$ be much larger when $Q = 0$?

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ **?**



$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

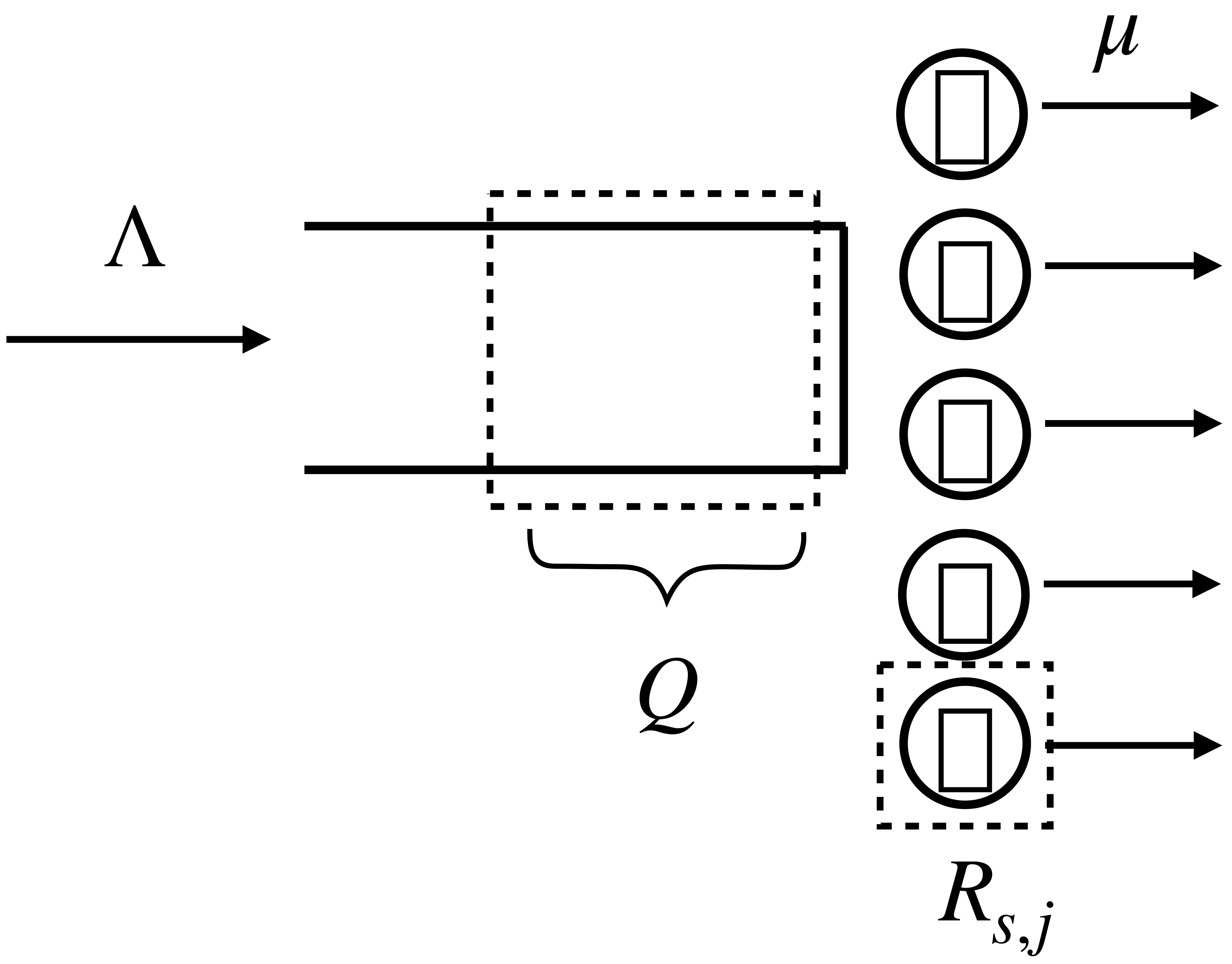
How much correlation?

Could $R_{s,j}$ be much larger when $Q = 0$?

Key challenge: understanding covariance

$$\mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{1-\rho}\sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$?



$$\Gamma_{s,j} = \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$\approx \textbf{Covariance}(1\{Q=0\}, R_{s,j})$$

independent?
correlated?

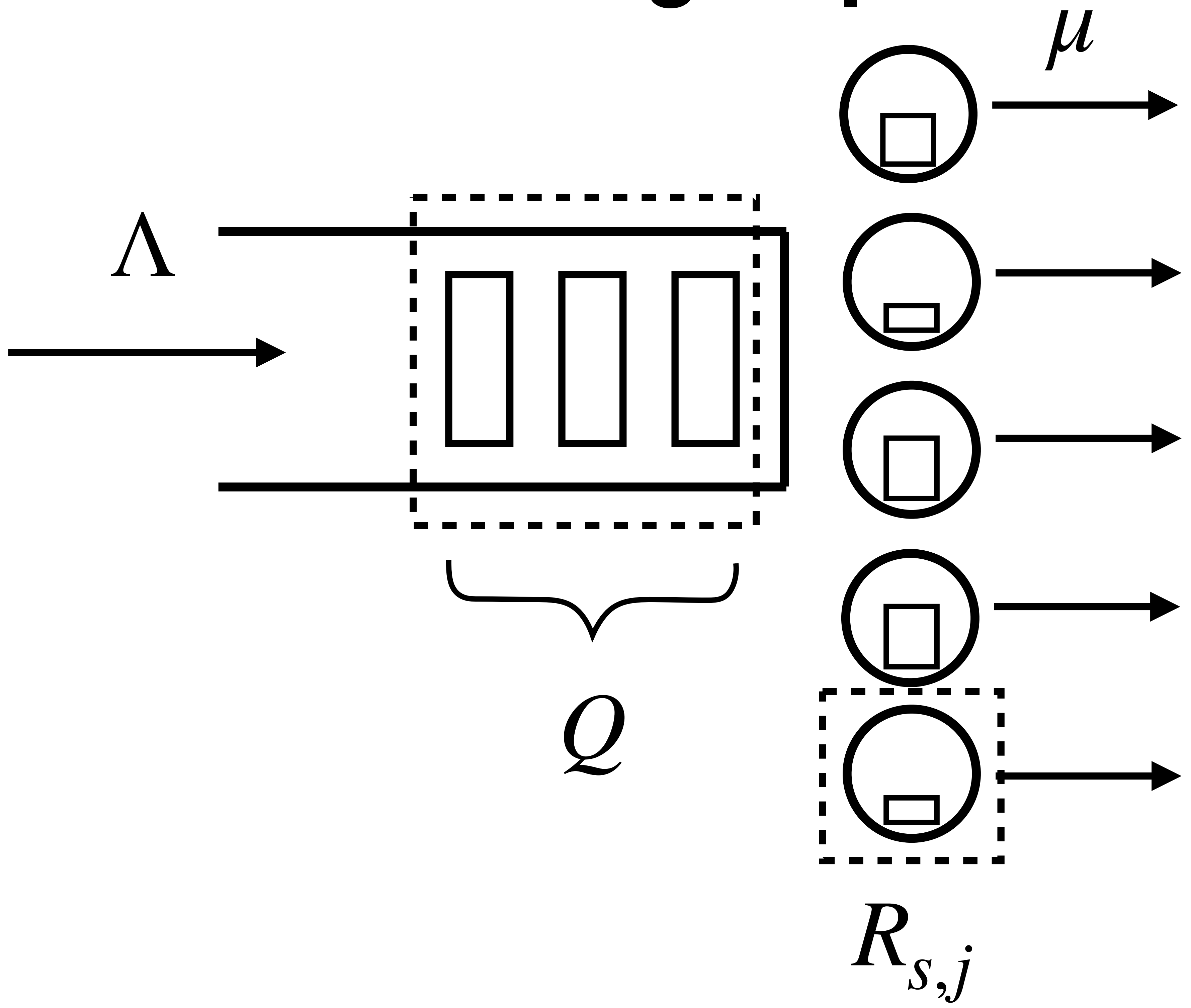
How much correlation?

Could $R_{s,j}$ be much larger when $Q=0$?

——Let's do a "controlled experiment"

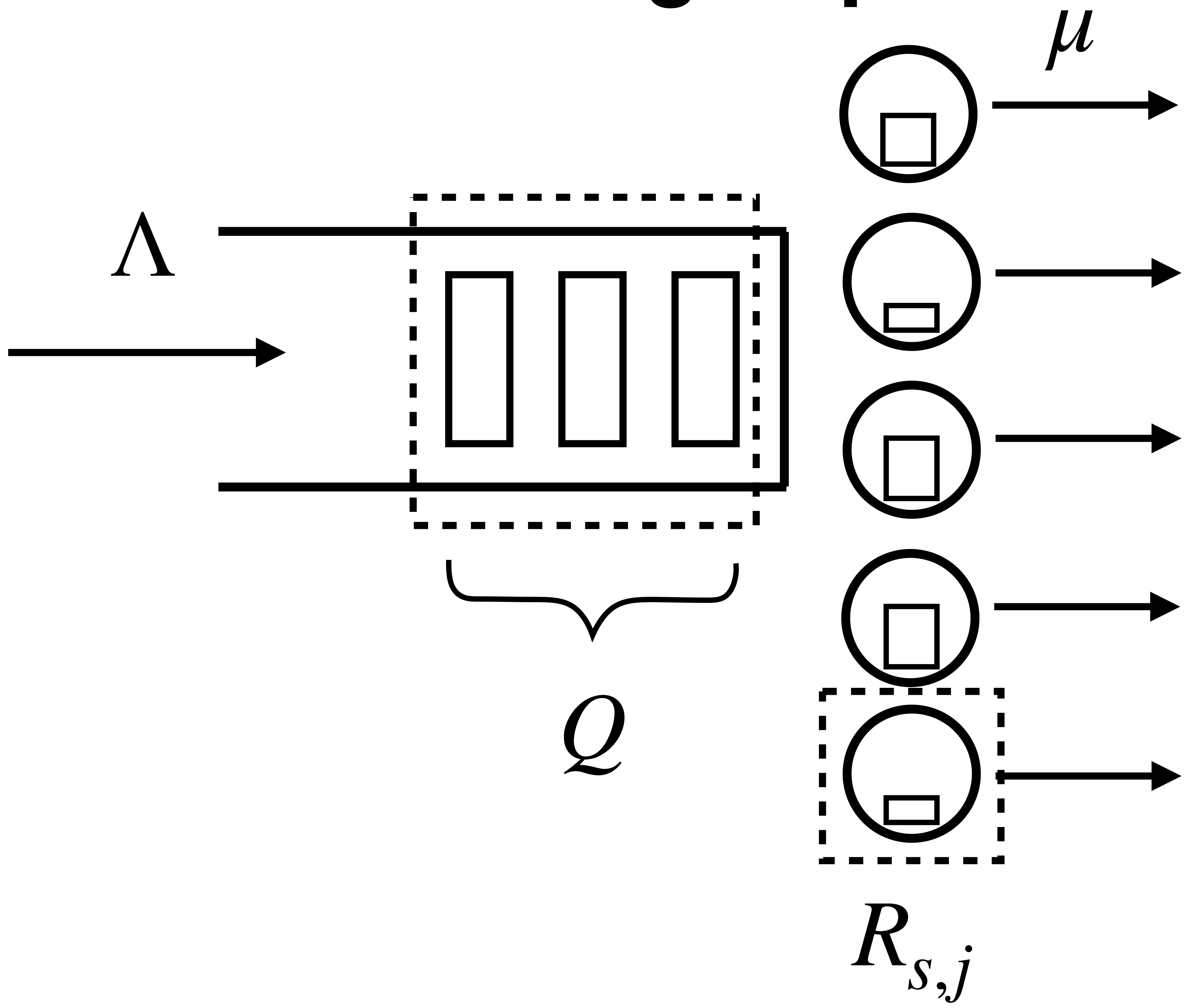
Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q = 0\}, R_{s,j}) = O(1/n)$

“Controlled group”

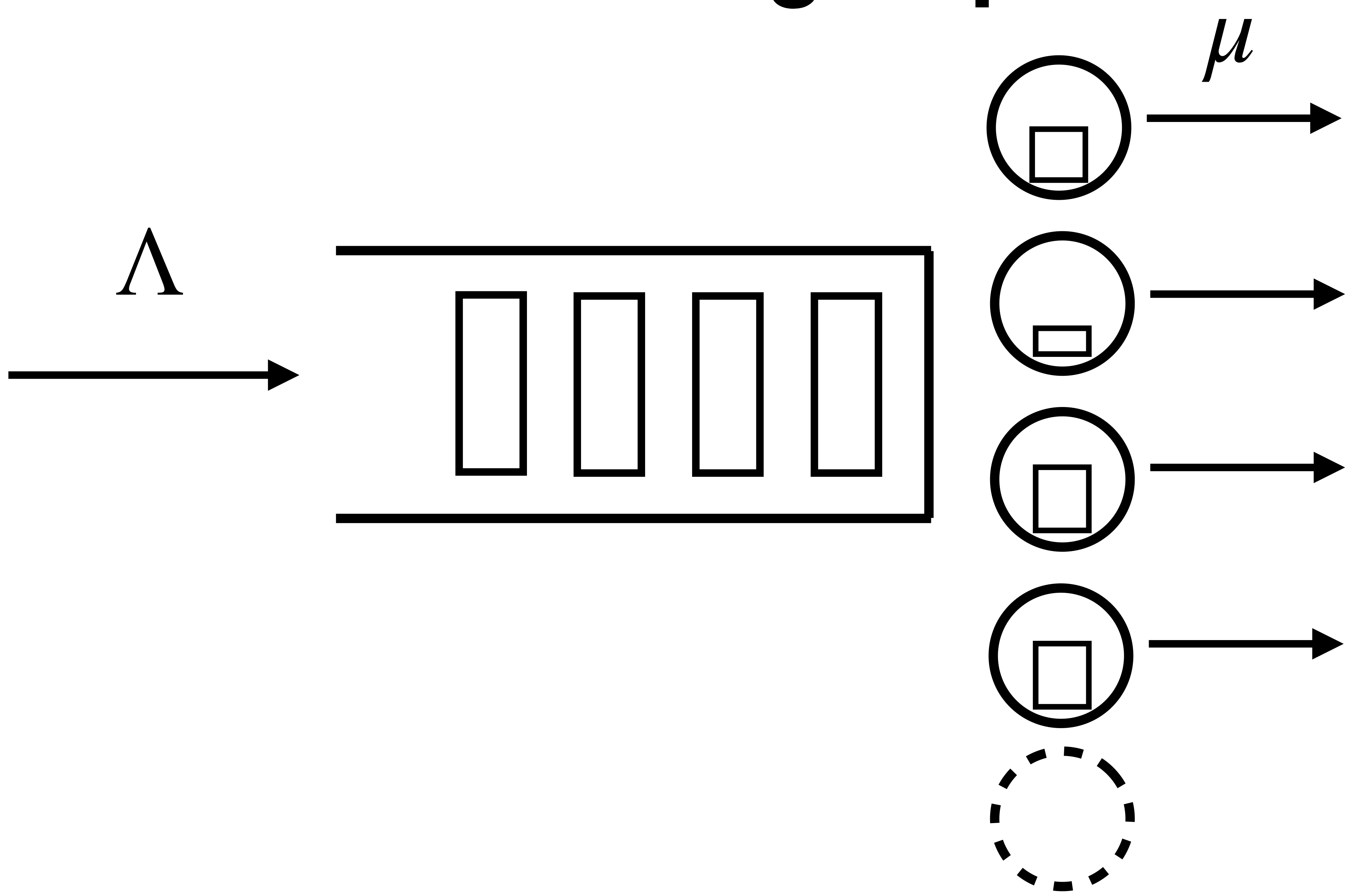


Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$

"Controlled group"

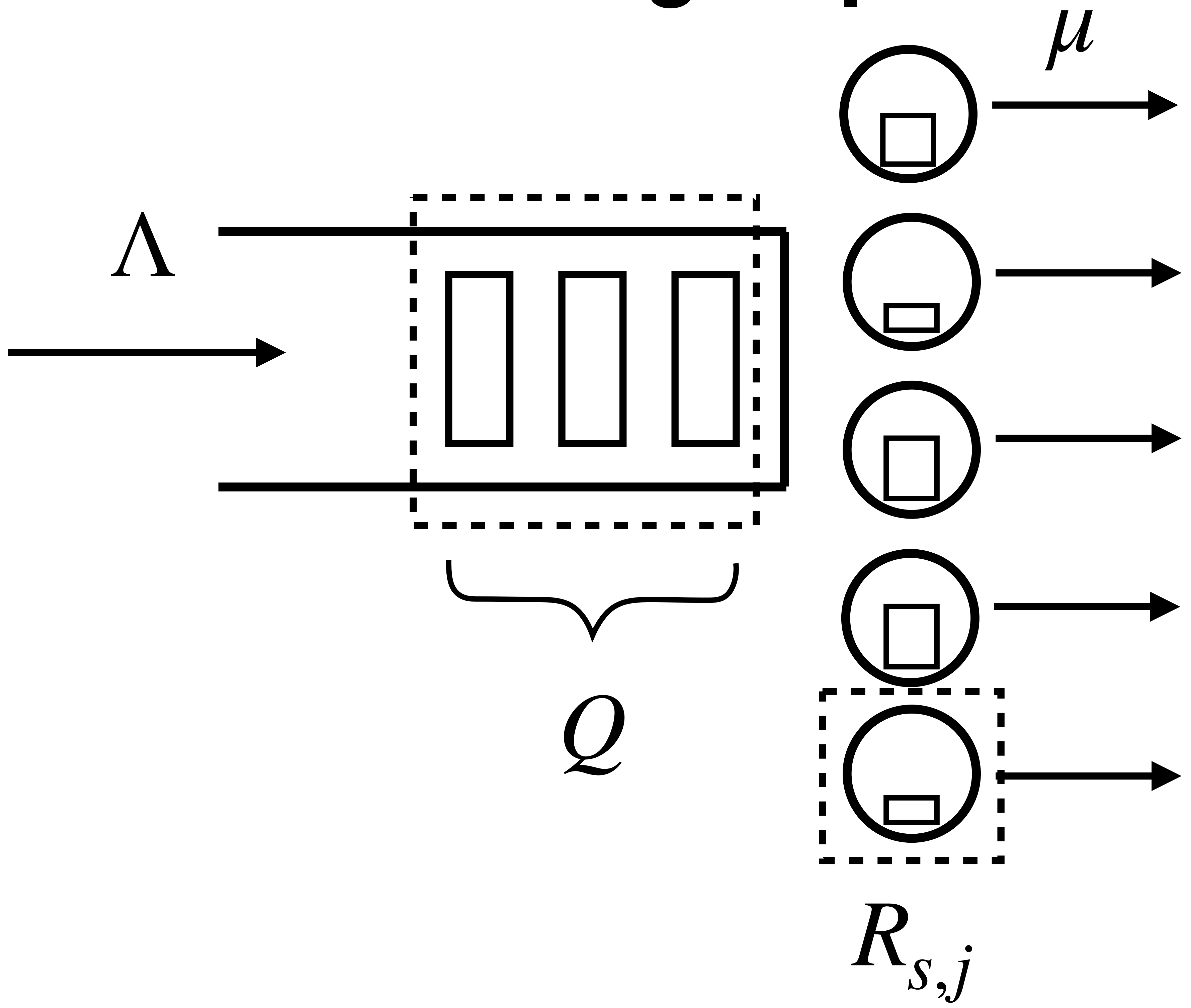


"Treatment group"

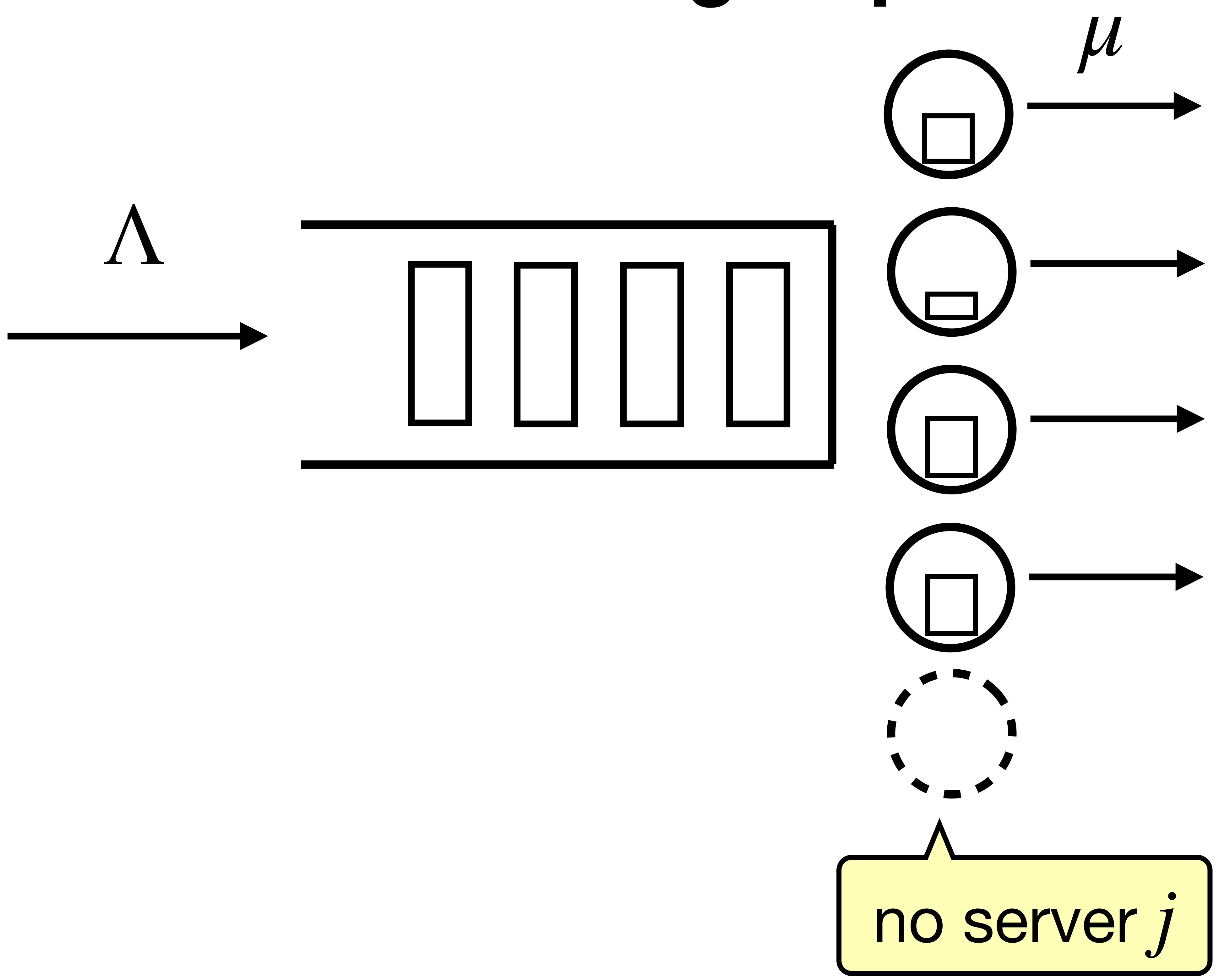


Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$

"Controlled group"

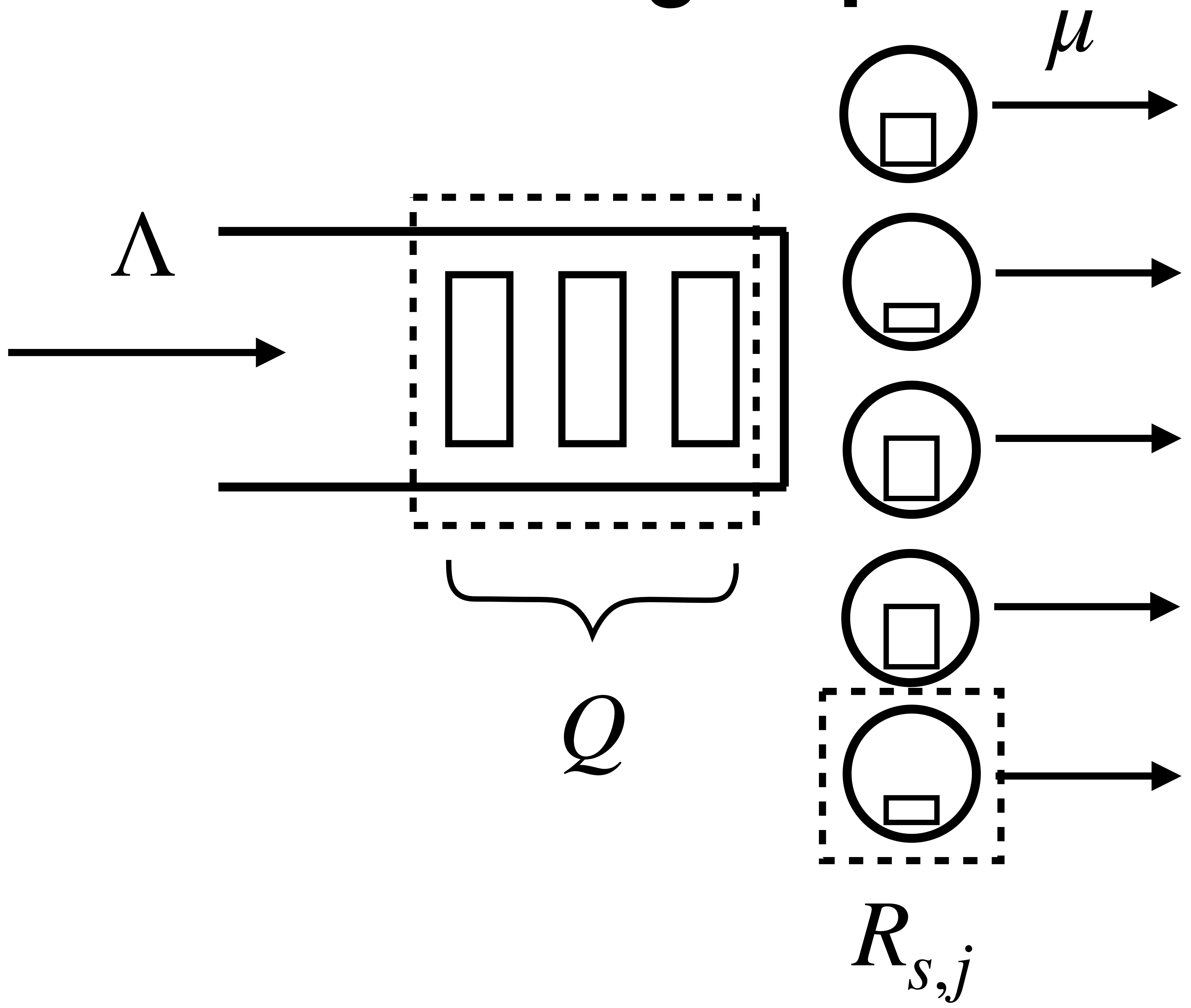


"Treatment group"

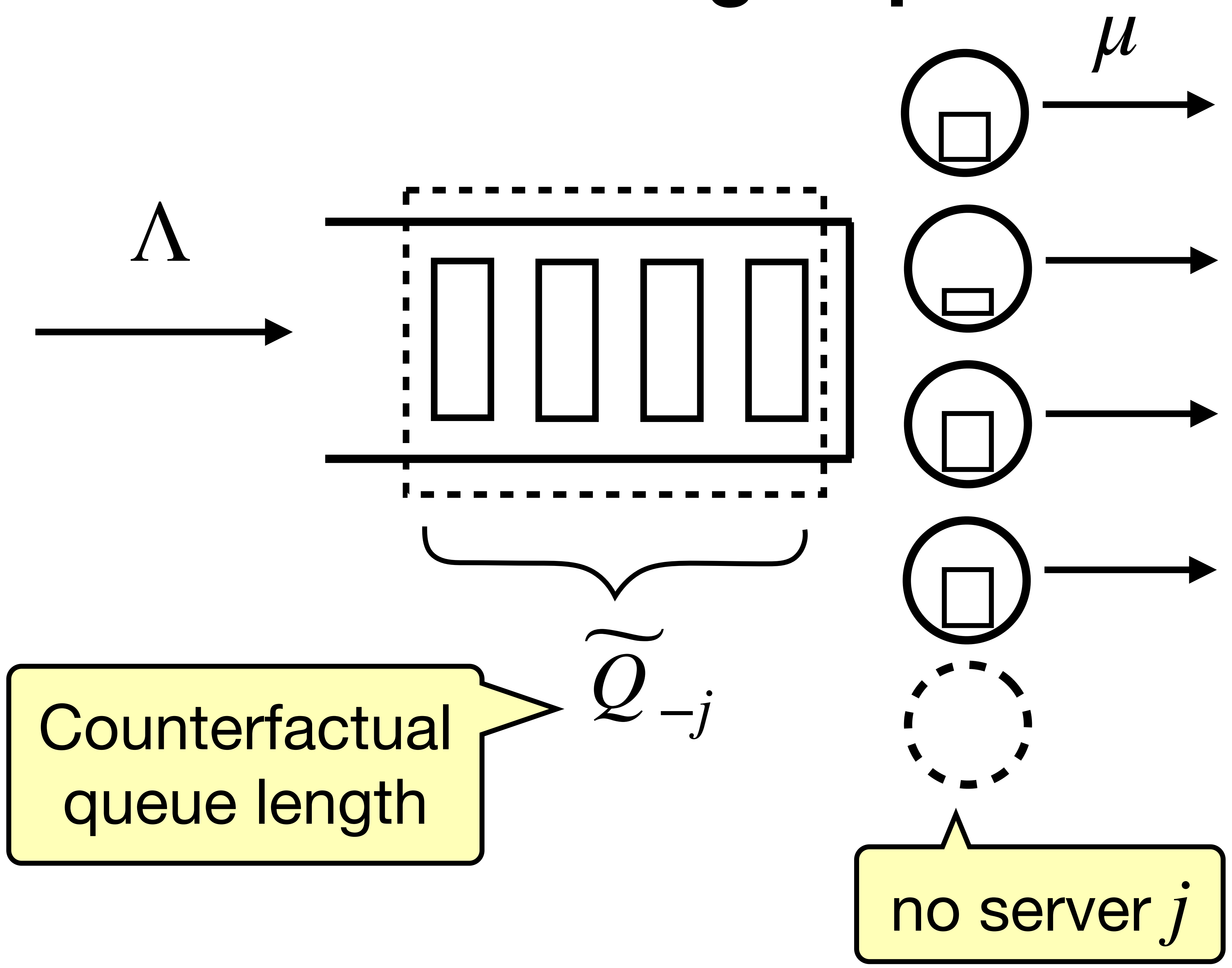


Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$

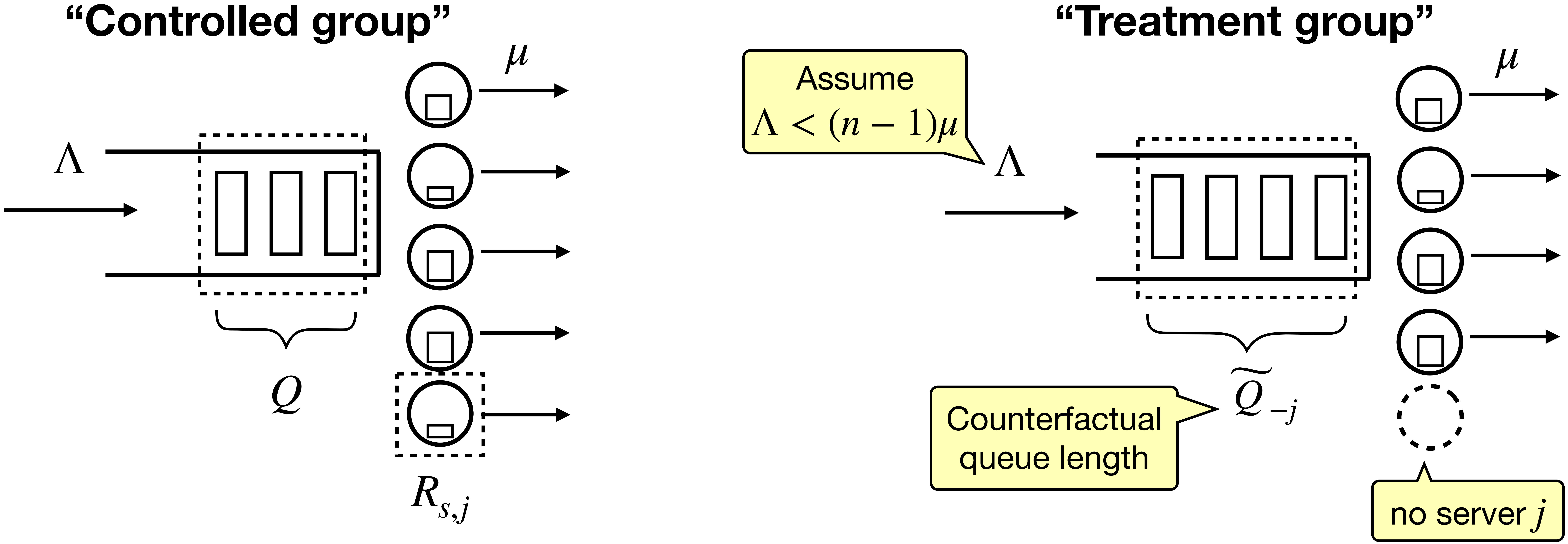
"Controlled group"



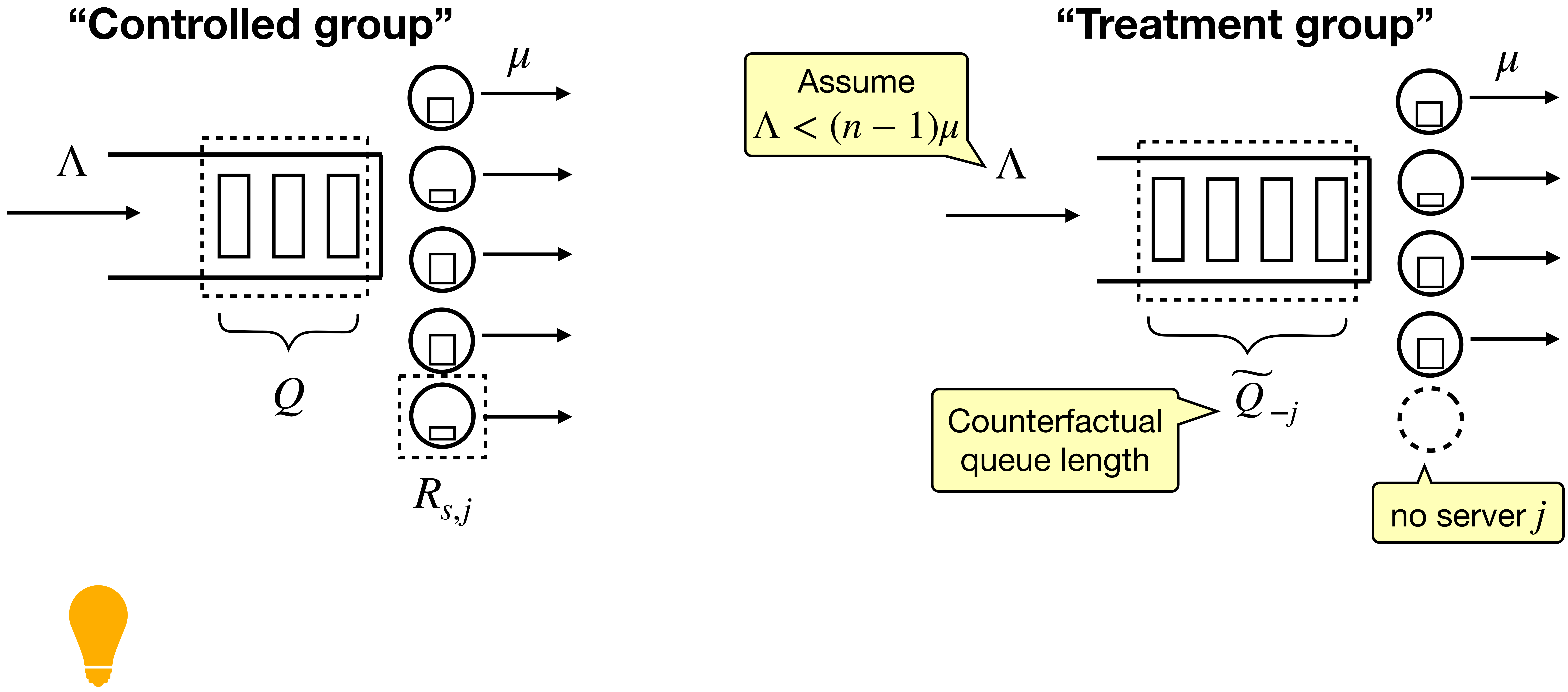
"Treatment group"



Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q = 0\}, R_{s,j}) = O(1/n)$

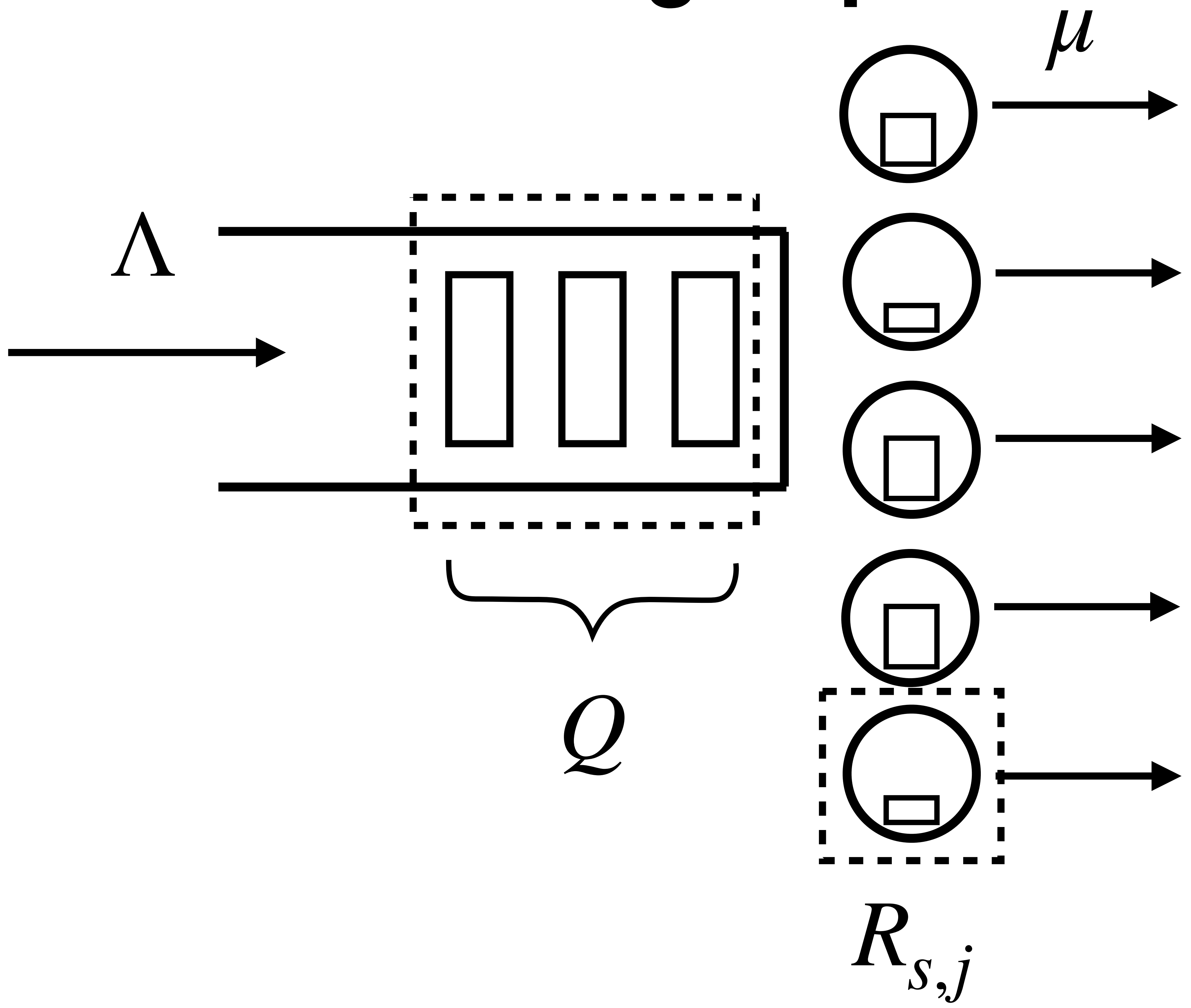


Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q = 0\}, R_{s,j}) = O(1/n)$

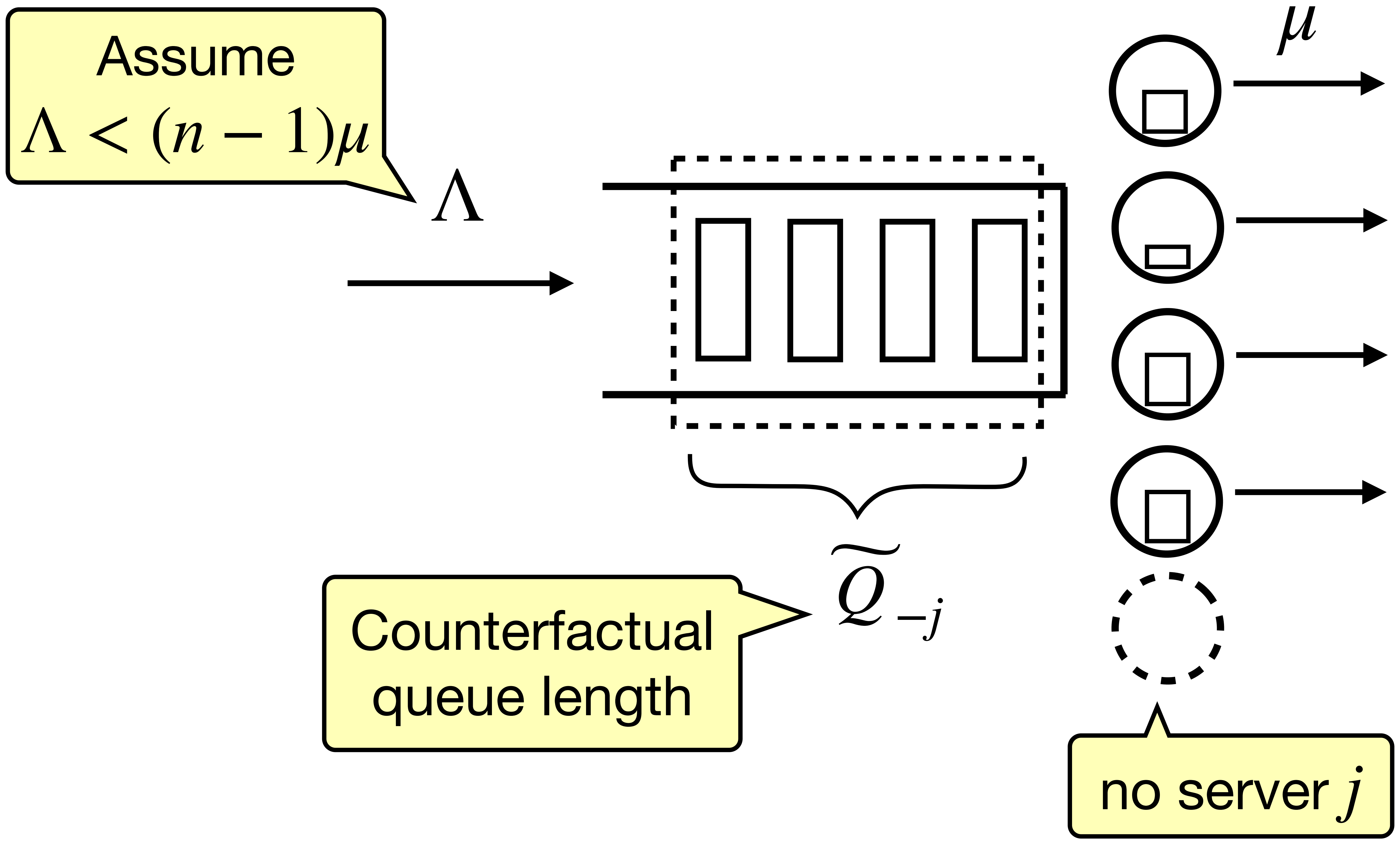


Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$

"Controlled group"

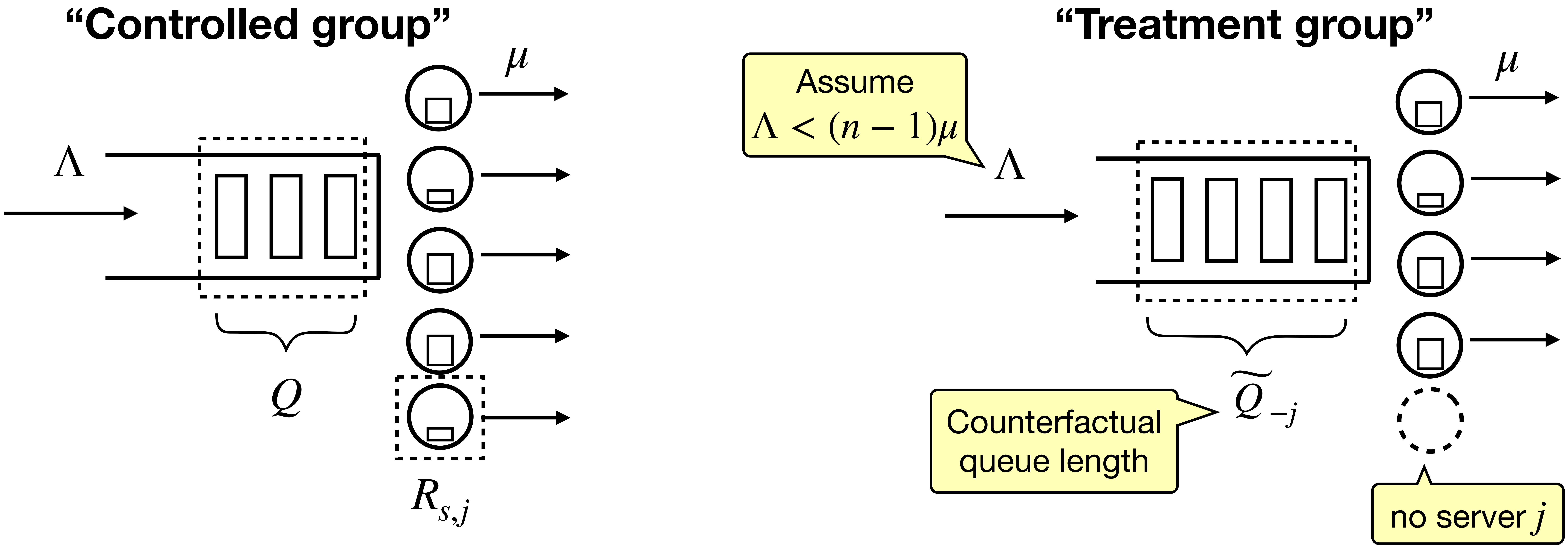


"Treatment group"



$$\mathbb{E}_s\left[1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right] \leq O(1/n)$$

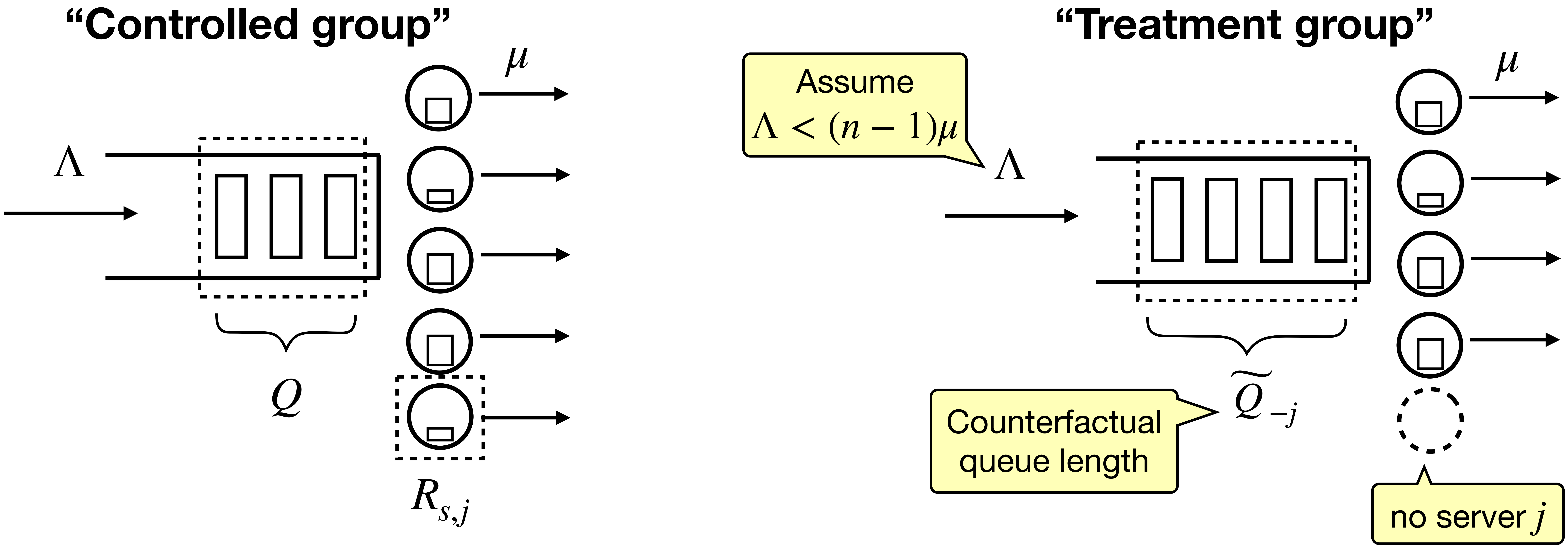
Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$



$$\mathbb{E}_s\left[1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right] \le O(1/n)$$

$$1\{Q=0\} \ge 1\{\widetilde{Q}_{-j}=0\} \text{ a.s.}$$

Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$

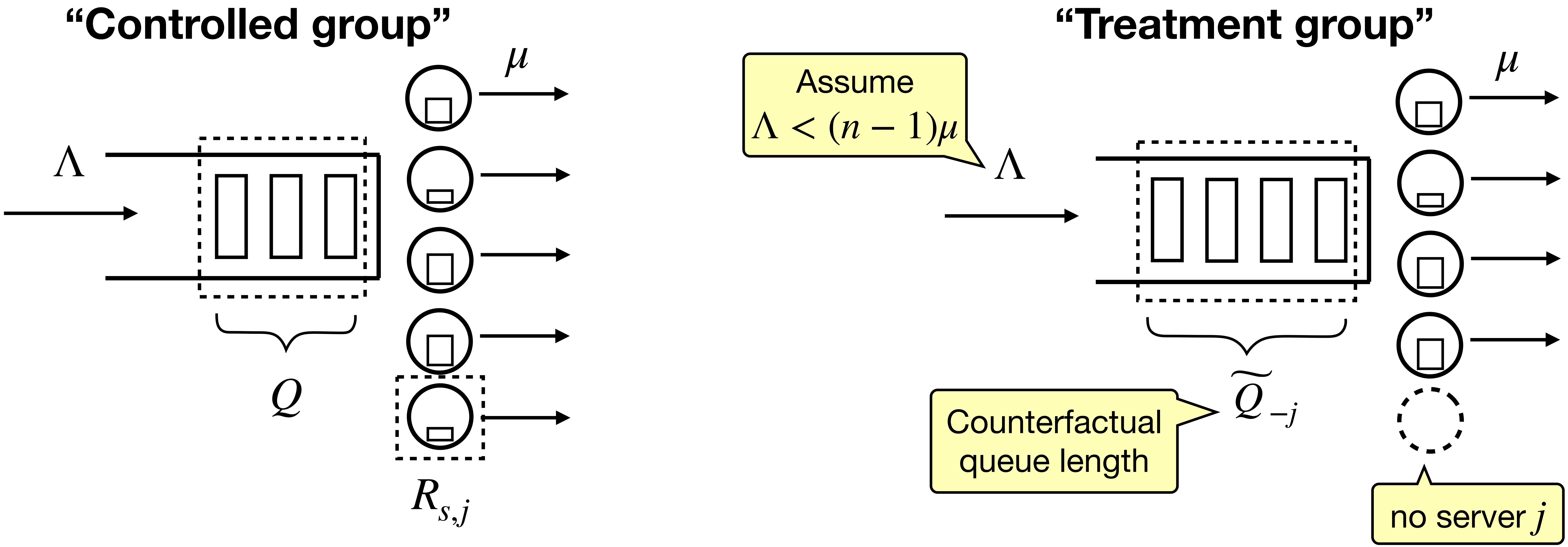


$$\mathbb{E}_s\left[1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right] \leq O(1/n)$$

$$1\{Q=0\} \geq 1\{\widetilde{Q}_{-j}=0\} \text{ a.s.}$$

and $\widetilde{Q}_{-j} \perp R_{s,j}$

Goal: $\Gamma_{s,j} \approx$ **Covariance**$(1\{Q=0\}, R_{s,j}) = O(1/n)$ 😆



$$\mathbb{E}_s\left[1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right] \le O(1/n)$$

$$1\{Q=0\} \ge 1\{\widetilde{Q}_{-j}=0\} \text{ a.s.}$$

and $\widetilde{Q}_{-j} \perp R_{s,j}$

Completing the proof

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)} \sum_{j=1}^{n} \Gamma_{s,j}$$

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)} \sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)} \sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

(Details of bounding $\Gamma_{s,j}$)

$$\Gamma_{s,j} \triangleq \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$= \underbrace{\mathbb{E}_s\left[\left(1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right)R_{s,j}\right]}_{\leq R_s^{\max}/n} + \underbrace{\mathbb{P}_s\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_s\left[R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]}_{= -\mathbb{E}[S^2]/(2n)}$$

$$= \left(R_s^{\max} - \mathbb{E}[S^2]/2\right)/n$$

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \le \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)} \sum_{j=1}^{n} \boxed{\Gamma_{s,j}}$$

$= O(1/n)$ ✔

(Details of bounding $\Gamma_{s,j}$)

$$\Gamma_{s,j} \triangleq \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$= \underbrace{\mathbb{E}_s\left[\left(1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right)R_{s,j}\right]}_{\le R_s^{\max}/n} + \underbrace{\mathbb{P}_s\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_s\left[R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]}_{= -\mathbb{E}[S^2]/(2n)}$$

$$= \left(R_s^{\max} - \mathbb{E}[S^2]/2\right)/n$$

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)} \sum_{j=1}^{n} \boxed{\Gamma_{s,j}} \leq \frac{R_s^{\max}}{1-\rho}$$

$= O(1/n)$ ✔

(Details of bounding $\Gamma_{s,j}$)

$$\begin{aligned}
\Gamma_{s,j} &\triangleq \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right] \\
&= \underbrace{\mathbb{E}_s\left[\left(1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right)R_{s,j}\right]}_{\leq R_s^{\max}/n} + \underbrace{\mathbb{P}_s\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_s\left[R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]}_{= -\mathbb{E}[S^2]/(2n)} \\
&= \left(R_s^{\max} - \mathbb{E}[S^2]/2\right)/n
\end{aligned}$$

Completing the proof

$$\mathbb{E}[Q]^{\text{M/GI/n}} \leq \mathbb{E}[Q]^{\text{modified M/GI/n}} = \frac{\mathbb{E}[S^2]}{2(1-\rho)} + \frac{1}{(1-\rho)}\sum_{j=1}^{n} \underbrace{\boxed{\Gamma_{s,j}}}_{= O(1/n)\ ✔} \leq \frac{R_s^{\max}}{1-\rho} \qquad \square$$

(Details of bounding $\Gamma_{s,j}$)

$$\Gamma_{s,j} \triangleq \mathbb{E}_s\left[1\{Q=0\}R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]$$

$$= \underbrace{\mathbb{E}_s\left[\left(1\{Q=0\} - 1\{\widetilde{Q}_{-j}=0\}\right)R_{s,j}\right]}_{\leq R_s^{\max}/n} + \underbrace{\mathbb{P}_s\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_s\left[R_{s,j}\right] - (1-\rho)\mathbb{E}\left[R_{s,j}\right]}_{= -\mathbb{E}[S^2]/(2n)}$$

$$= \left(R_s^{\max} - \mathbb{E}[S^2]/2\right)/n$$

Summary

Thank you!

GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{\mathrm{Var}(\Lambda A) + 2R_s^{\max} - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t\geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$ and $R_a^{\min} = \inf_{t \geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$

— —hold for any number of servers n and any load $\rho < 1$

Yige Hong, 5th-year PhD student in Carnegie Mellon University

Looking for an academic job!

Summary

Thank you!

GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \le \frac{\mathrm{Var}(\Lambda A) + 2R_s^{\max} - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \le \frac{R_s^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t\ge 0} \mathbb{E}[\mu S - t \,|\, \mu S \ge t]$ and $R_a^{\min} = \inf_{t\ge 0} \mathbb{E}[\Lambda A - t \,|\, \Lambda A \ge t]$

——hold for any number of servers n and any load $\rho < 1$

Yige Hong, 5th-year PhD student in Carnegie Mellon University

Looking for an academic job!

Summary

Thank you!

GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{\mathrm{Var}(\Lambda A) + 2R_s^{\max} - \rho}{2(1-\rho)} + \left(\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}\right)$$

M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{R_s^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t\geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$ and $R_a^{\min} = \inf_{t\geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$

— —hold for any number of servers n and any load $\rho < 1$

Yige Hong, 5th-year PhD student in Carnegie Mellon University

Looking for an academic job!

Backup slides

My Harris ergodicity assumptions

Assumption 1: the service time distribution should be *non-lattice*, i.e., its support is not a subset of $\{0, \delta, 2\delta, 3\delta, \ \dots\}$ for any $\delta > 0$

Assumption 2:

- Modified GI/GI/n is positive Harris recurrent with finite $\mathbb{E}[Q]$ whenever $\rho < 1$
- Each leave-one-out system is positive Harris recurrent whenever $\rho < 1 - 1/n$

My Harris ergodicity assumptions

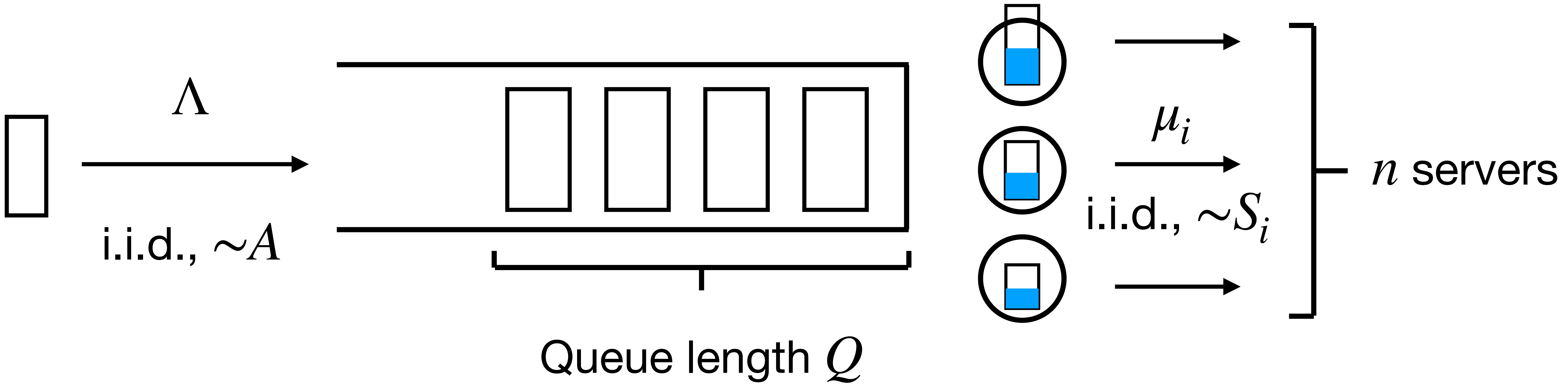
Assumption 2:

- Modified GI/GI/n is Harris ergodic with $\mathbb{E}[Q] < \infty$ whenever $\rho < 1$
- Each leave-one-out system is Harris ergodic whenever $\rho < 1 - 1/n$

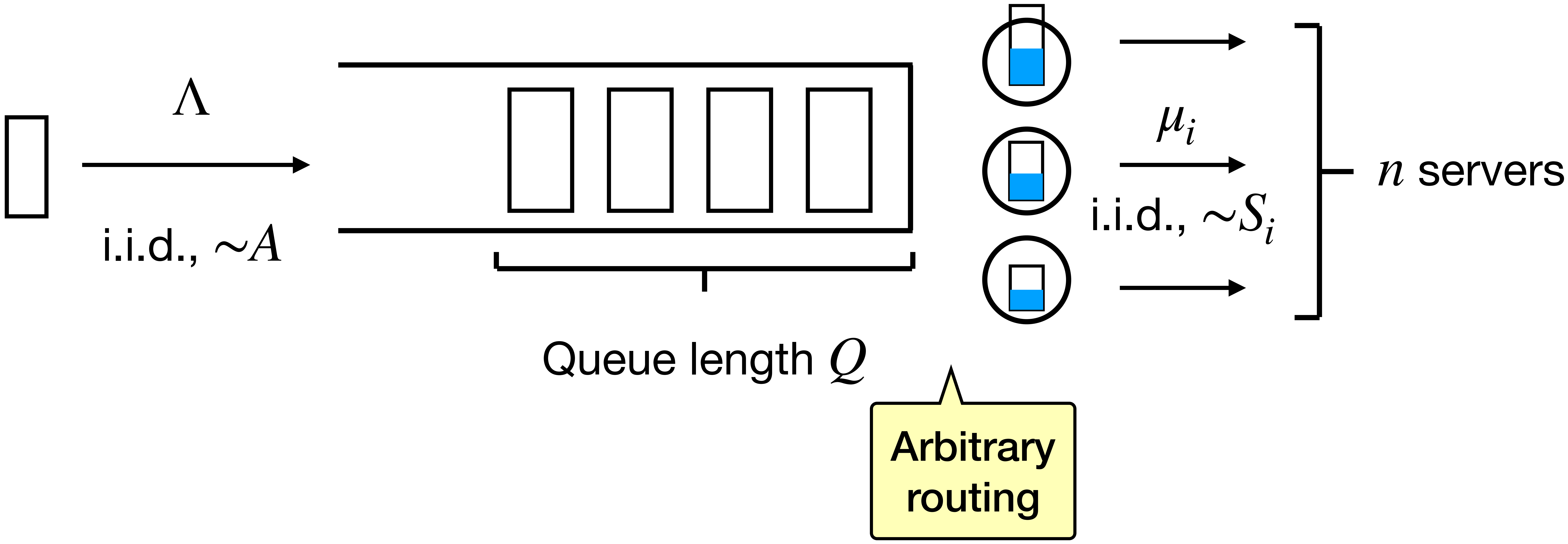
Remarks:

- Artifact of proof technique; require separate efforts
- GI/GI/n is Harris ergodic with $\mathbb{E}[Q] < \infty$ whenever $\rho < 1$ if service times and interarrival times have finite second moments (see Asmussen 2003, Section XII)
- Intuitively, modified G/G/n should not be worse, but we may need A and S to be *spread-out* (roughly, have a continuous component), and A being unbounded.
  - Essentially, only need to show Harris ergodicity; $\mathbb{E}[Q] < \infty$ follows from Li and Goldberg (2025)'s proof.
  - Alternatively, prove using fluid limit methods (Dai and Meyn 1995)

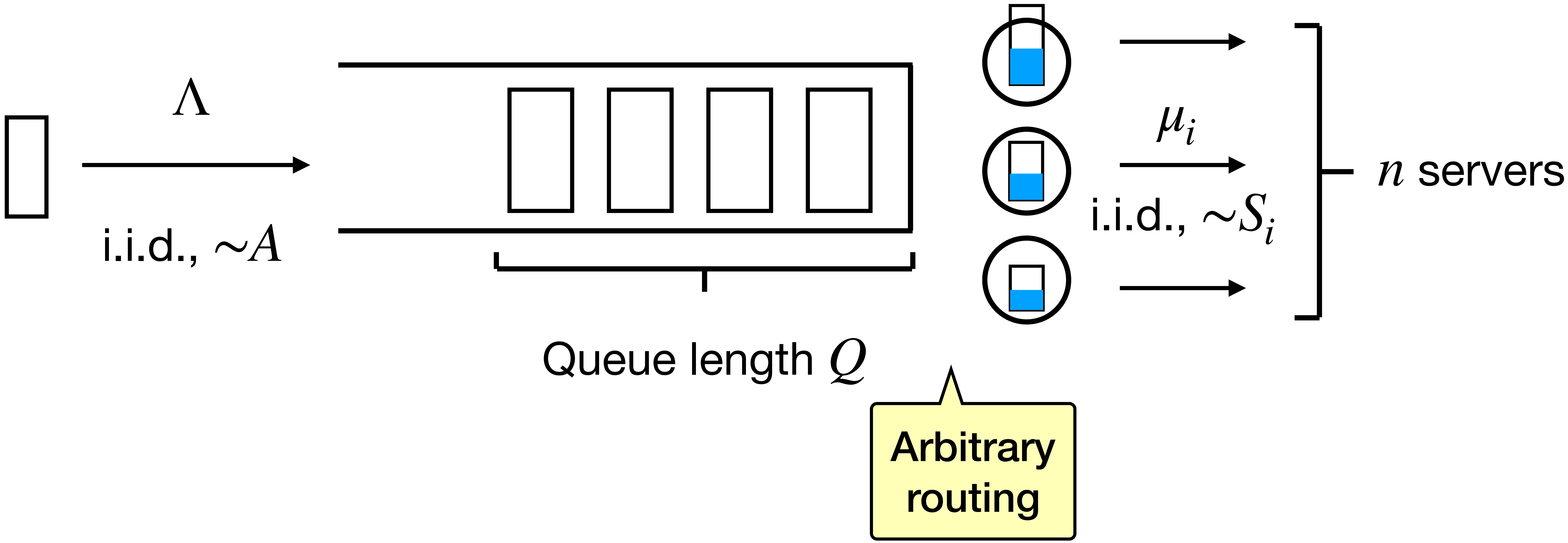
GI/GI/n with heterogeneous servers



GI/GI/n with heterogeneous servers



GI/GI/n with heterogeneous servers



$$\mu_\Sigma = \sum_{i=1}^{n} \mu_i \text{ and } \rho = \frac{\Lambda}{\mu_\Sigma}$$

Results for GI/GI/n with heterogeneous servers

GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2\sum_{i=1}^{n}(\mu_i/\mu_\Sigma)R_{s,i}^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(R_a^{\min} - \mathbb{E}[(\Lambda A)^2]/2\right)$$

M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{\sum_{i=1}^{n}(\mu_i/\mu_\Sigma)R_{s,i}^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t\geq 0} \mathbb{E}[\mu S - t \,|\, \mu S \geq t]$ and $R_a^{\min} = \inf_{t\geq 0} \mathbb{E}[\Lambda A - t \,|\, \Lambda A \geq t]$

Results for GI/GI/n with heterogeneous servers

GI/GI/n

Assume $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$, $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{2\sum_{i=1}^{n}(\mu_i/\mu_\Sigma)R_{s,i}^{\max} + \mathrm{Var}(\Lambda A) - \rho}{2(1-\rho)} + \left(R_a^{\min} - \mathbb{E}[(\Lambda A)^2]/2\right)$$

M/GI/n

Assume $R_s^{\max} < \infty$ and $\rho < 1$:

$$\mathbb{E}[Q] \leq \frac{\sum_{i=1}^{n}(\mu_i/\mu_\Sigma)R_{s,i}^{\max}}{1-\rho}$$

where $R_s^{\max} = \sup_{t\geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$ and $R_a^{\min} = \inf_{t\geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$

(Omit the Harris ergodicity assumptions)

Basic Adjoint Relationship

How it works:

$\mathbb{E}[f(X(1)) - f(X(0))] = 0$ for any “regular” test function f, if $X(0) \sim$ stationary distr. π

[see, e.g., Braverman et al. 25]

Basic Adjoint Relationship

How it works:

$\mathbb{E}[f(X(1)) - f(X(0))] = 0$ for any “regular” test function f, if $X(0) \sim$ stationary distr. π

[see, e.g., Braverman et al. 25]

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

Basic Adjoint Relationship

How it works:

$\mathbb{E}[f(X(1)) - f(X(0))] = 0$ for any "regular" test function f, if $X(0) \sim$ stationary distr. π

[see, e.g., Braverman et al. 25]

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

What test function should we take?

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

M/M/1 $\longrightarrow$ M/GI/1 $\longrightarrow$ M/GI/n

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

$$\text{M/M/1} \longrightarrow \text{M/GI/1} \longrightarrow \text{M/GI/n}$$

$$f(X) = Q^2$$

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

M/M/1 $\longrightarrow$ M/GI/1 $\longrightarrow$ M/GI/n

$f(X) = Q^2$ $f(X) = (Q + R_s)^2$

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

M/M/1 $\longrightarrow$ M/GI/1 $\longrightarrow$ M/GI/n

$f(X) = Q^2$

$f(X) = (Q + R_s)^2$

$$f(X) = \left(Q + \sum_{i=1}^{n} R_{s,i}\right)^2$$

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

M/M/1 $\longrightarrow$ M/GI/1 $\longrightarrow$ M/GI/n

$f(X) = Q^2$

$f(X) = (Q + R_s)^2$

$$f(X) = \left(Q + \sum_{i=1}^{n} R_{s,i}\right)^2$$

Give exact formula for $\mathbb{E}[Q]$

Test functions

In modified M/GI/1, the state $X = (R_{s,1}, R_{s,2}, \ldots, R_{s,n}, Q)$

M/M/1 $\longrightarrow$ M/GI/1 $\longrightarrow$ M/GI/n

$f(X) = Q^2$ $\quad$ $f(X) = (Q + R_s)^2$ $\quad$ $f(X) = \left(Q + \sum_{i=1}^{n} R_{s,i}\right)^2$

Give exact formula for $\mathbb{E}[Q]$

- $f(X) = \mathbb{E}[\text{work} \mid X]^2$, similar to work^2 in [Grosof et al. 22]

Discussion of tightness

- Lack a lower bound
  - Modified system is an upper bound of the original system
- The bound $\mathbb{E}[Q] \leq \dfrac{R_s^{\max}}{1-\rho}$ is likely loose:
  - When $R_s^{\max}$ is large, arbitrarily worse
  - Recall that for M/M/n, $\mathbb{E}[Q] = \dfrac{\rho}{1-\rho} P_Q$, where $P_Q \to 0$ when $1-\rho = \omega\left(\dfrac{1}{\sqrt{n}}\right)$,

    i.e., sub-HW regime