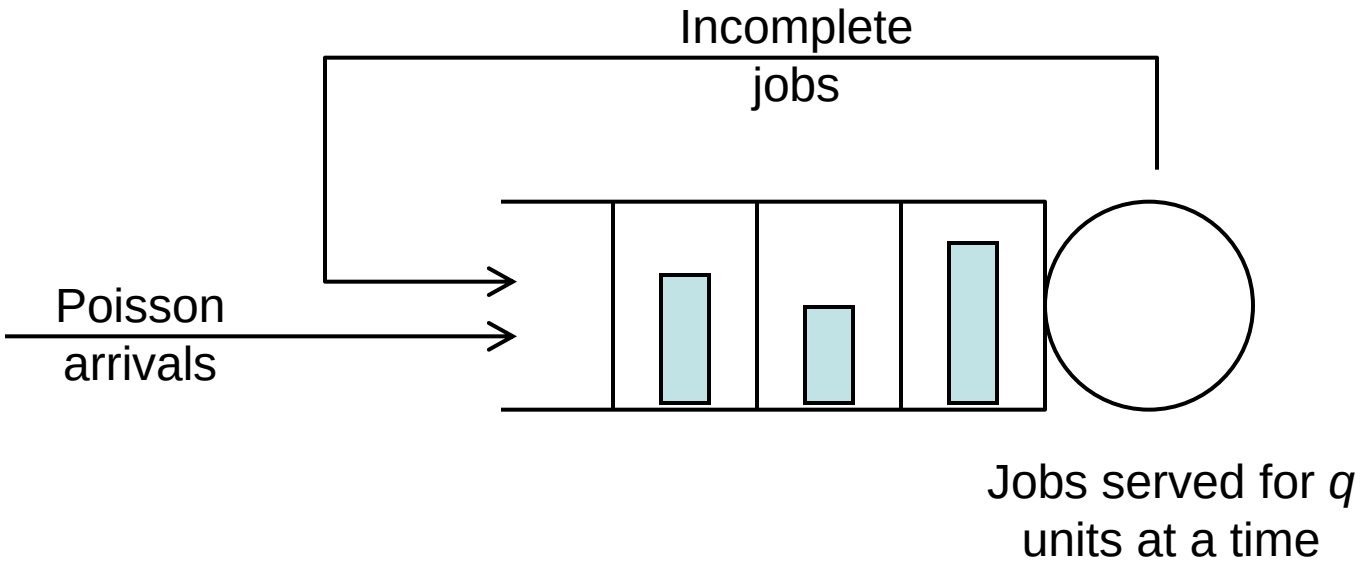
FINDING THE OPTIMAL QUANTUM SIZE

Revisiting the M/G/1 Round-Robin Queue

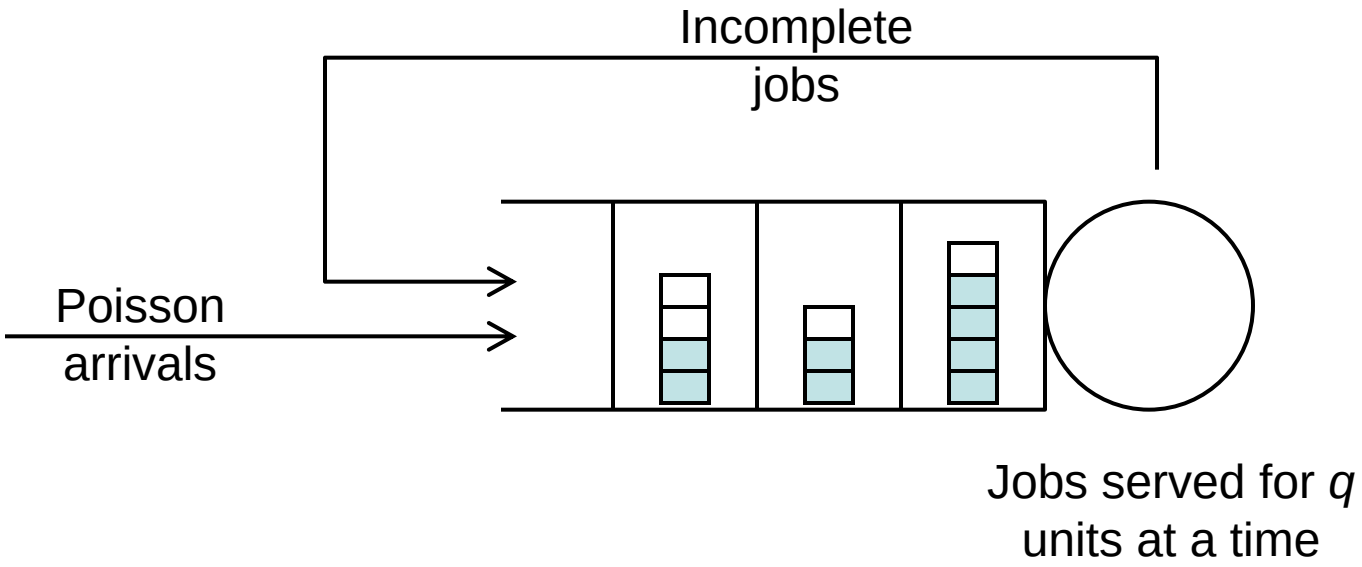
VARUN GUPTA

Carnegie Mellon University

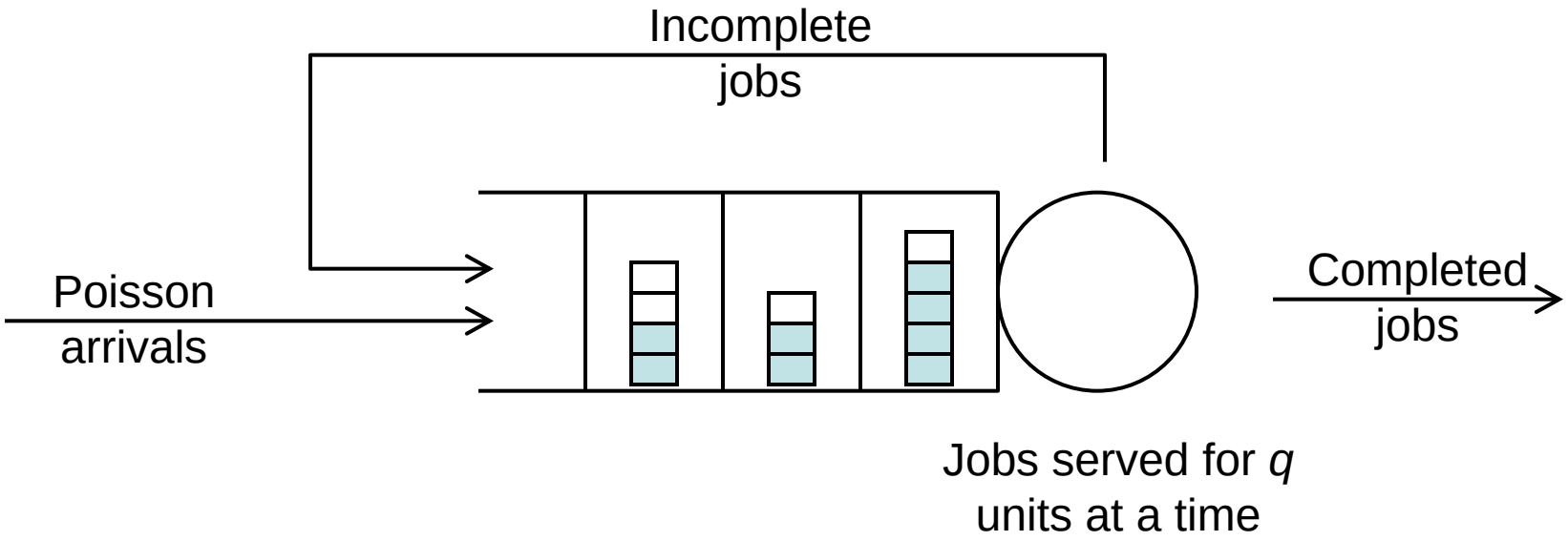
M/G/1/RR


Incomplete
jobs
Poisson
arrivals
Jobs served for *q*
units at a time

M/G/1/RR


Incomplete
jobs
Poisson
arrivals
Jobs served for *q*
units at a time

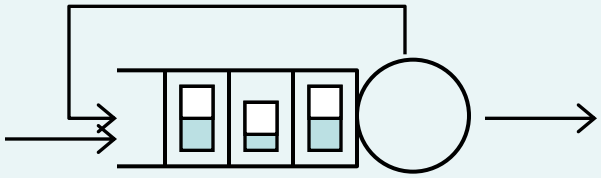
M/G/1/RR



- arrival rate = λ
- job sizes i.i.d. ~ S
- load $\rho = \lambda E[S]$
- $C^2 = \frac{var(S)}{E[S]^2}$

Squared coefficient of variability (SCV) of job sizes: $C^2 \geq 0$

M/G/1/RR

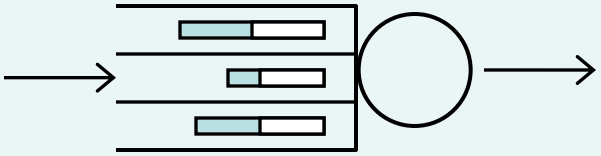


q

$q \to 0$

$q = \infty$

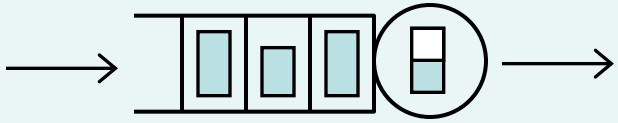
M/G/1/PS



$$E[T^{PS}] = \frac{E[S]}{1-\rho}$$

Preemptions cause overhead
(e.g. OS scheduling)

M/G/1/FCFS



$$E[T^{FCFS}] = E[T^{PS}]\left(1 + \rho\frac{C^2-1}{2}\right)$$

Variable job sizes cause long delays

GOAL

Optimal operating *q* for M/G/1/RR with *overheads* and *high* C^2

effect of preemption overheads

SUBGOAL

Sensitivity Analysis: Effect of *q* and C^2 on M/G/1/RR performance

(no switching overheads)

PRIOR WORK

- **Lots of *exact* analysis: [Wolff70], [Sakata et al.71], [Brown78]**

👎 **No closed-form solutions/bounds**

👎 **No simple expressions for interplay of *q* and C^2**

Outline

- Effect of q and C^2 on mean response time
  - Approximate analysis
  - Bounds for M/G/1/RR

No preemption overheads

- Choosing the optimal quantum size

Effect of preemption overheads

Approximate sensitivity analysis of M/G/1/RR

Approximation assumption 1:

Service quantum ~ Exp(1/*q*)

Approximation assumption 2:

$$\text{Job size distribution} \sim \begin{cases} 0 & \text{w.p. } p \\ \text{Exp}(\mu) & \text{w.p. } 1-p \end{cases}$$

$$E[T^{RR*}] = E[T^{PS}] \left[1 + \frac{C^2-1}{C^2+1} \cdot \frac{\lambda}{\frac{1}{q} + \frac{2}{C^2+1}\frac{1}{E[S]}} \right]$$

Approximate sensitivity analysis of M/G/1/RR

$$E[T^{RR*}] = E[T^{PS}]\left[1 + \frac{C^2-1}{C^2+1} \cdot \frac{\lambda}{\frac{1}{q}+\frac{2}{C^2+1}\frac{1}{E[S]}}\right]$$

- Monotonic in q
  - Increases from $E[T^{PS}] \rightarrow E[T^{FCFS}]$
- Monotonic in C^2
  - Increases from $E[T^{PS}] \rightarrow E[T^{PS}](1+\lambda q)$

For high C^2: $E[T^{RR*}] \approx E[T^{PS}](1+\lambda q)$

Outline

- Effect of q and C^2 on mean response time
  - ✓ Approximate analysis
  - Bounds for M/G/1/RR
- Choosing the optimal quantum size

M/G/1/RR bounds

Assumption: job sizes $\in \{0, q, \ldots, Kq\}$

THEOREM:

$$\left[\frac{1-\rho}{2} + \frac{1}{2} \cdot \frac{1+\rho}{1+\lambda q}\right] \leq \frac{E[T^{RR}]}{E[T^{PS}]} \leq \left[1 + \frac{(1+\rho)\lambda q}{2} - \frac{\rho(1+\rho)}{2K}\right]$$

Lower bound is TIGHT:
E[S] = iq

$$E[T^{RR}] = E[T^{PS}]\left[\frac{1-\rho}{2} + \frac{1}{2} \cdot \frac{1+\rho}{1+\lambda q}\right]$$

Upper bound is TIGHT within $(1+\rho/K)$:
Job sizes $\in \{0, Kq\}$

$$E[T^{RR}] = \frac{E[T^{PS}]}{1+\rho/K}\left[1 + \frac{(1+\rho)\lambda q}{2} - \frac{\rho^2}{K}\right]$$

As $K \to \infty$: $\sup E[T^{RR}] = E[T^{PS}]\left[1 + \frac{(1+\rho)\lambda q}{2}\right]$

Outline

- ✓ Effect of q and C^2 on mean response time
  - ✓ Approximate analysis
  - ✓ Bounds for M/G/1/RR
- ❑ Choosing the optimal quantum size

Optimizing q

Preemption overhead = h

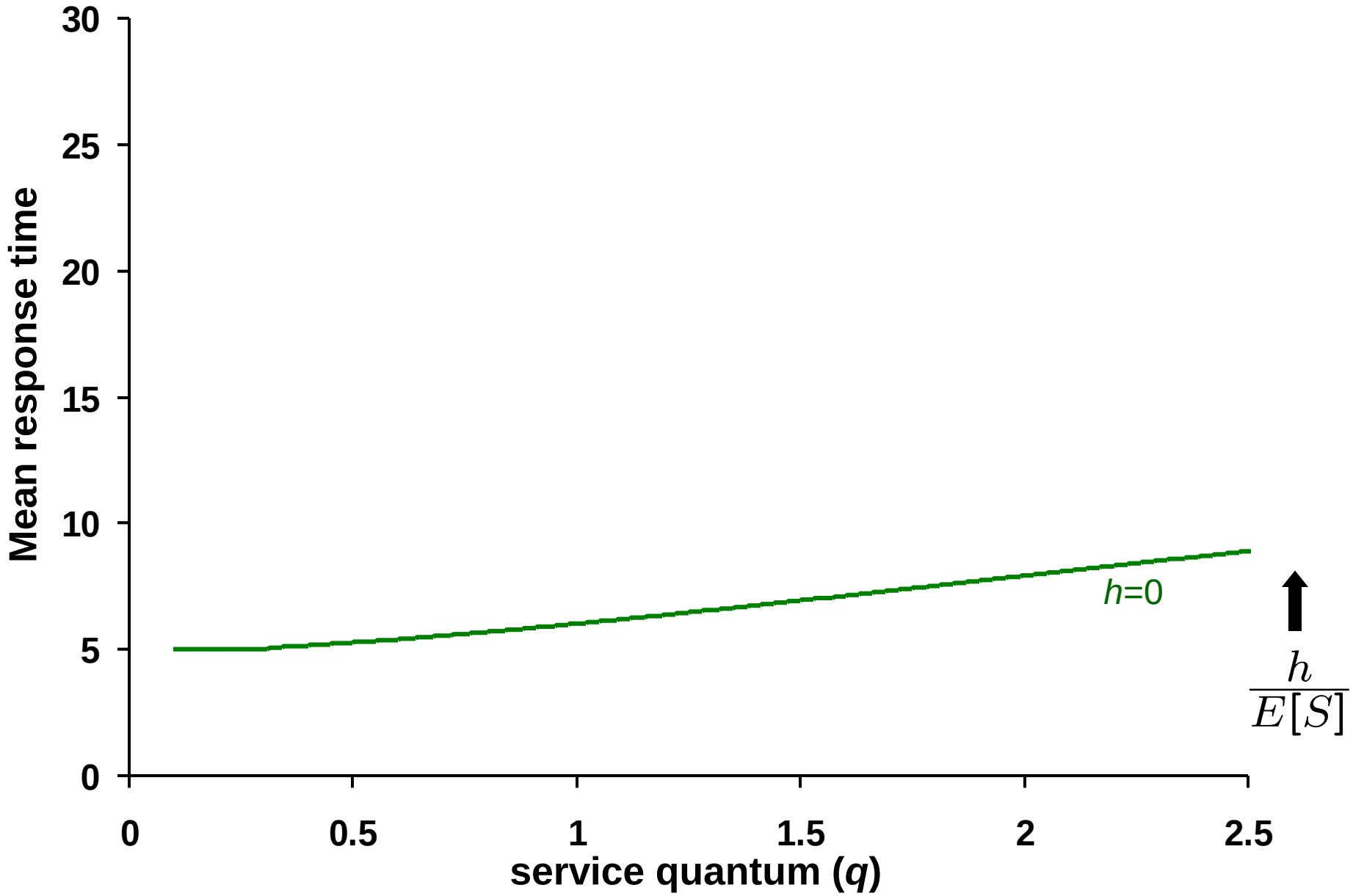
1.
$$\begin{aligned} q' &= q + h \\ E[S]' &= E[S]\left(1 + \frac{h}{q}\right) \\ \rho' &= \lambda E[S]' \end{aligned}$$

2. Minimize $E[T^{RR}]$ upper bound from Theorem:

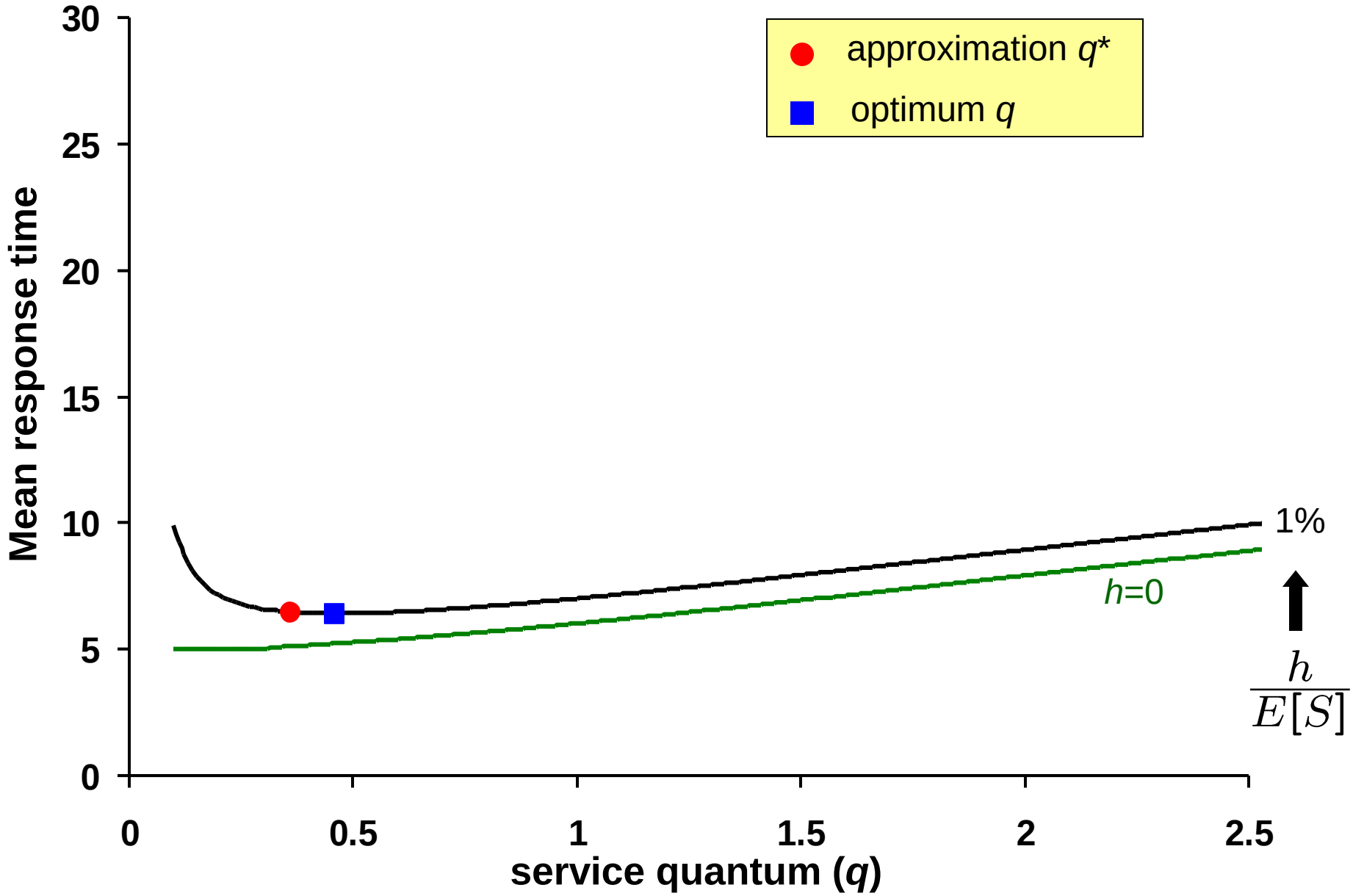
$$q^* = \text{argmin}_q \; \frac{E[S]'}{1-\rho'}\left[1 + \frac{(1+\rho')\lambda q'}{2}\right]$$

Common case: $\dfrac{h}{E[S]} \ll (1-\rho)$

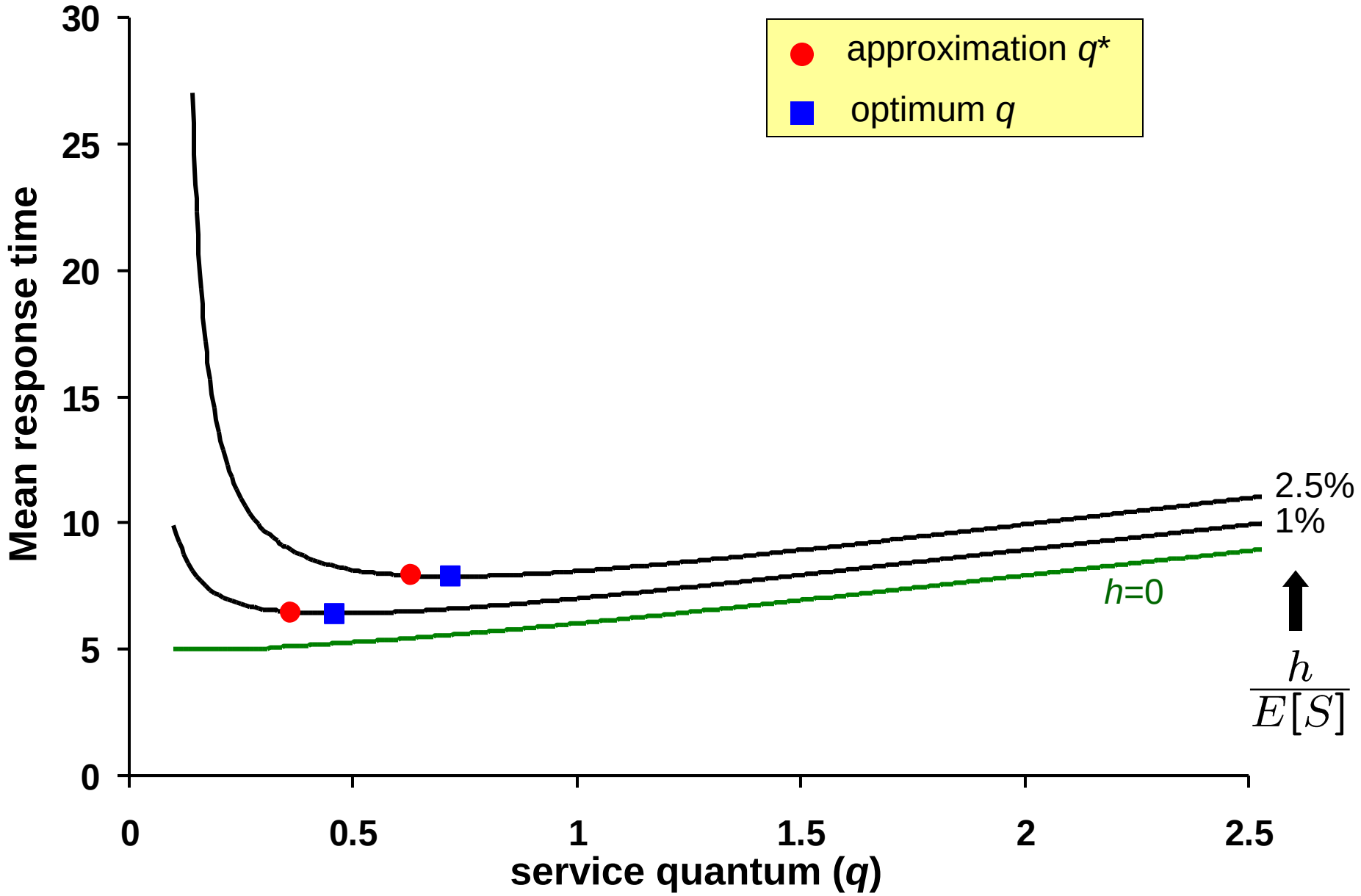
$$q^* \approx \alpha(\rho)\sqrt{hE[S]}$$



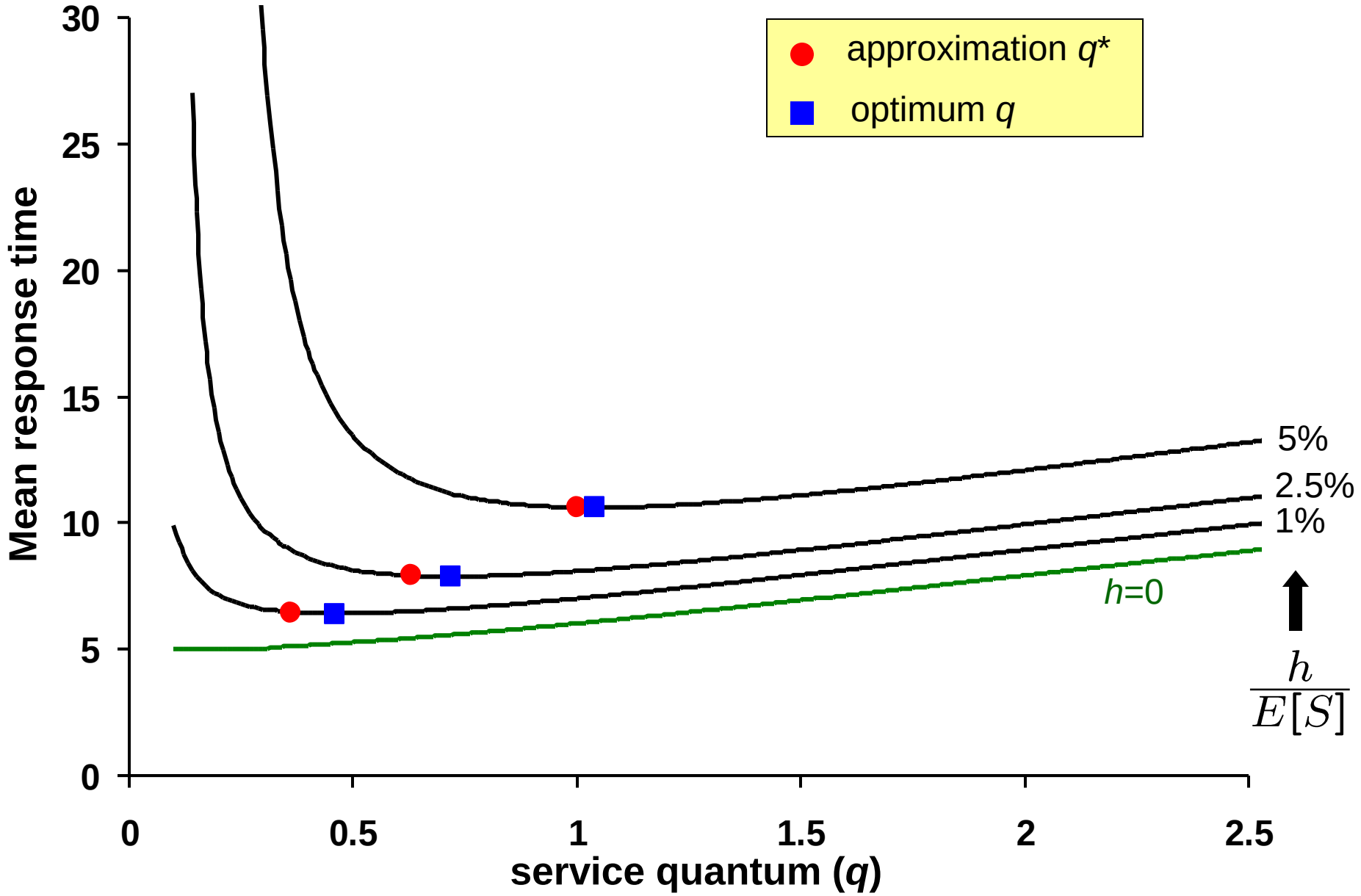
Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$



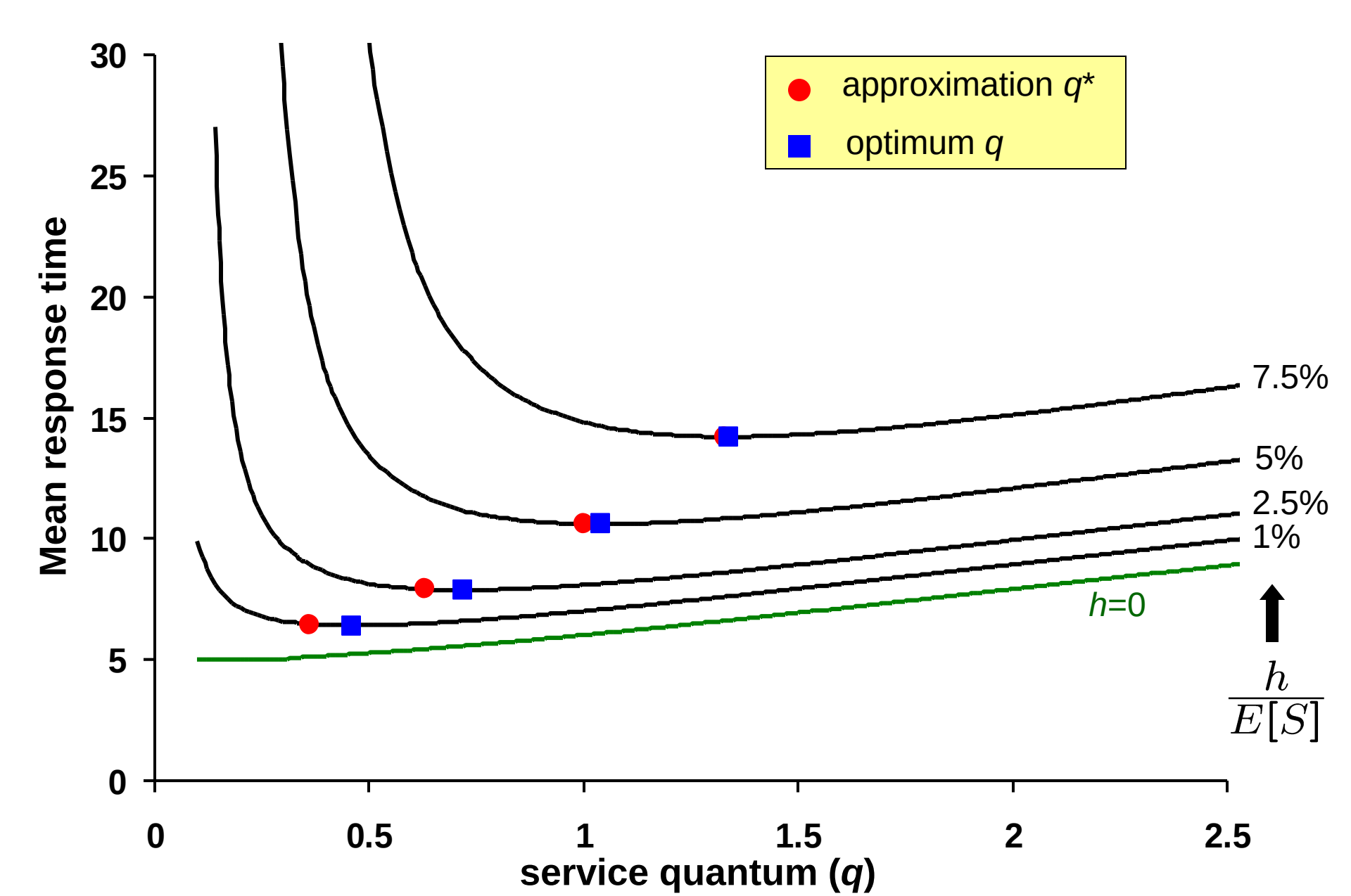
Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$



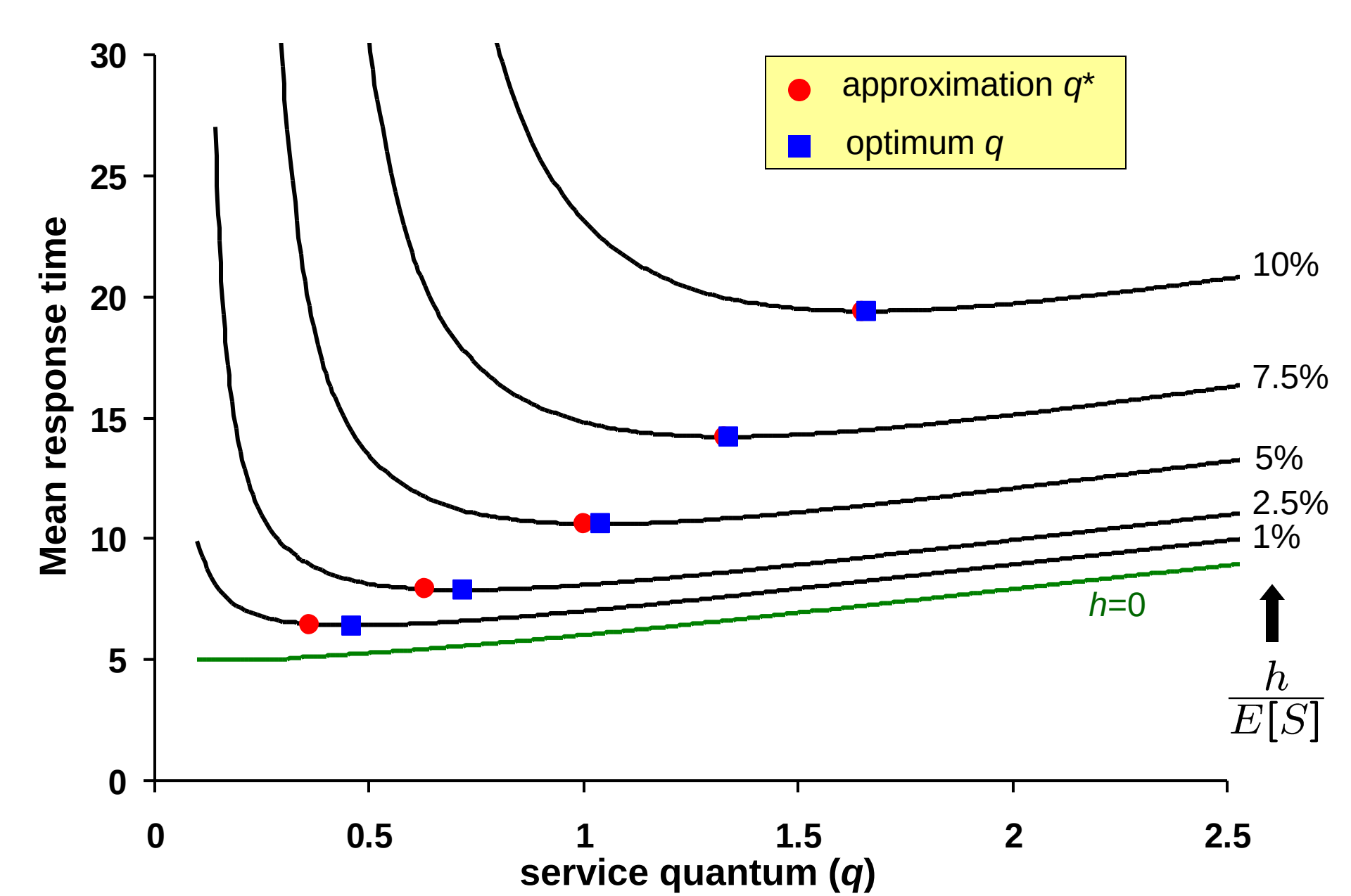
Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$



Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$



Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$



Job size distribution: H_2 (balanced means) $E[S] = 1$, $C^2 = 19$, $\rho = 0.8$

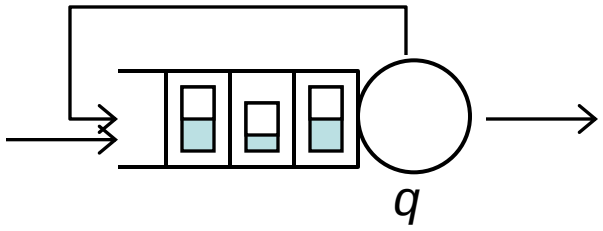
Outline

- ✓ Effect of q and C^2 on mean response time
  - ✓ Approximate analysis
  - ✓ Bounds for M/G/1/RR
- ✓ Choosing the optimal quantum size

Conclusion/Contributions

- Simple approximation and bounds for M/G/1/RR

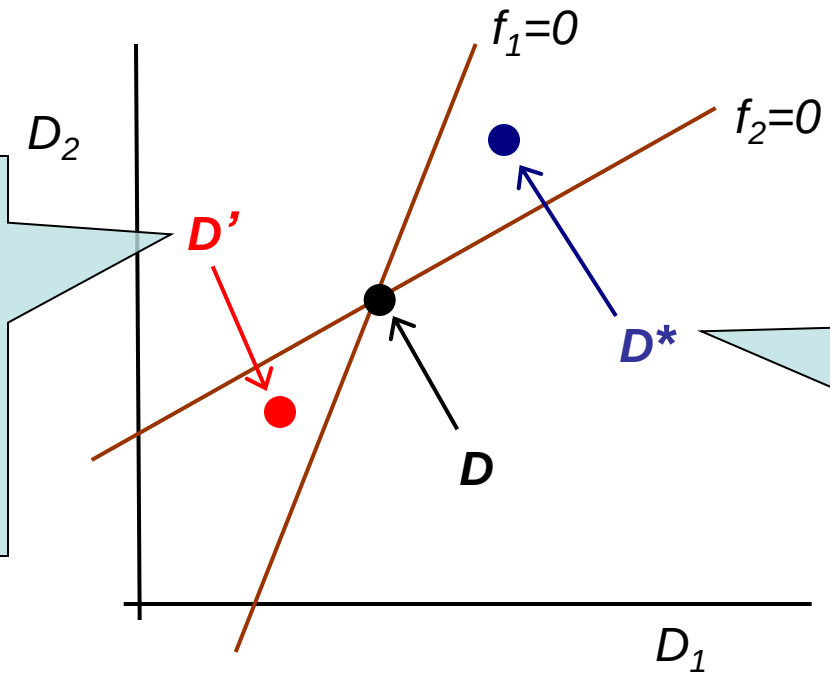
- Optimal quantum size for handling highly variable job sizes under preemption overheads



Bounds – Proof outline

- D_i = mean delay for i^{th} quantum of service
- $\boldsymbol{D} = [D_1\ D_2\ \ldots\ D_K]$
- $\boldsymbol{D}$ is the fixed point of a monotone linear system:

$$\boldsymbol{D}^T = \boldsymbol{A}_P\boldsymbol{D}^T + \boldsymbol{b}$$



Sufficient condition for lower bound:

$\boldsymbol{D'}^T \leq \boldsymbol{A}_P\boldsymbol{D'}^T + \boldsymbol{b}$

for all P

Sufficient condition for upper bound:

$\boldsymbol{D^*}^T \geq \boldsymbol{A}_P\boldsymbol{D^*}^T + \boldsymbol{b}$

for all P

Optimizing *q*

- Preemption overhead = *h*
- *q'* = *q*+*h*, E[*S*]' → E[*S*] (*1*+*h*/*q*), ρ' = λ *E*[*S*]'
- $\min_q \frac{E[S]'}{1-\rho'}\left[1+\frac{(1+\rho')\lambda q'}{2}\right]$

Heavy traffic: $\frac{1}{1-\rho} \gg \frac{E[S]}{h}$

$$q^* \approx \frac{2h}{1-\rho}$$

Small overhead: $\frac{1}{1-\rho} \ll \frac{E[S]}{h}$

$$q^* \approx K(\rho)\sqrt{hE[S]}$$