Optimal Load Balancing Policies for Heterogeneous Server Farms

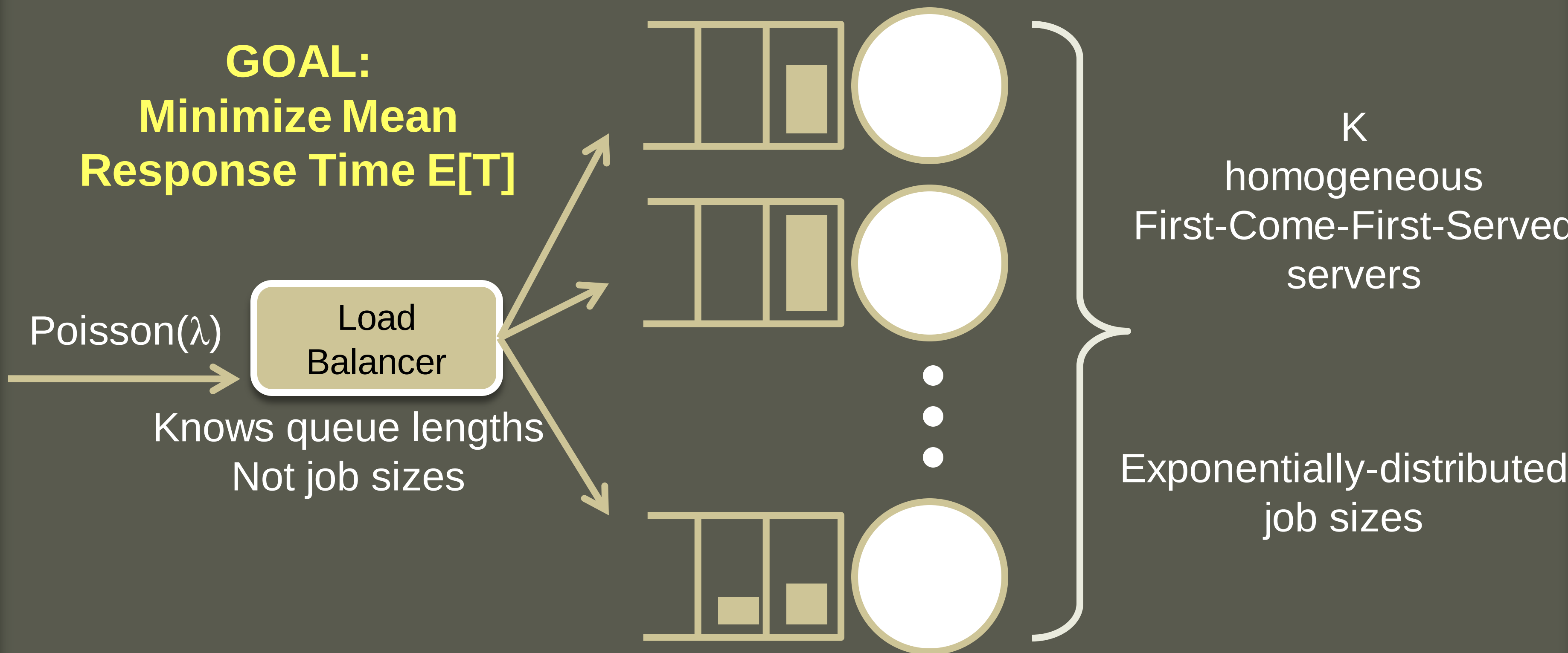
VARUN GUPTA
Carnegie Mellon University

With:

Mor Harchol-Balter
(CMU)

Warm up

GOAL:
Minimize Mean Response Time E[T]

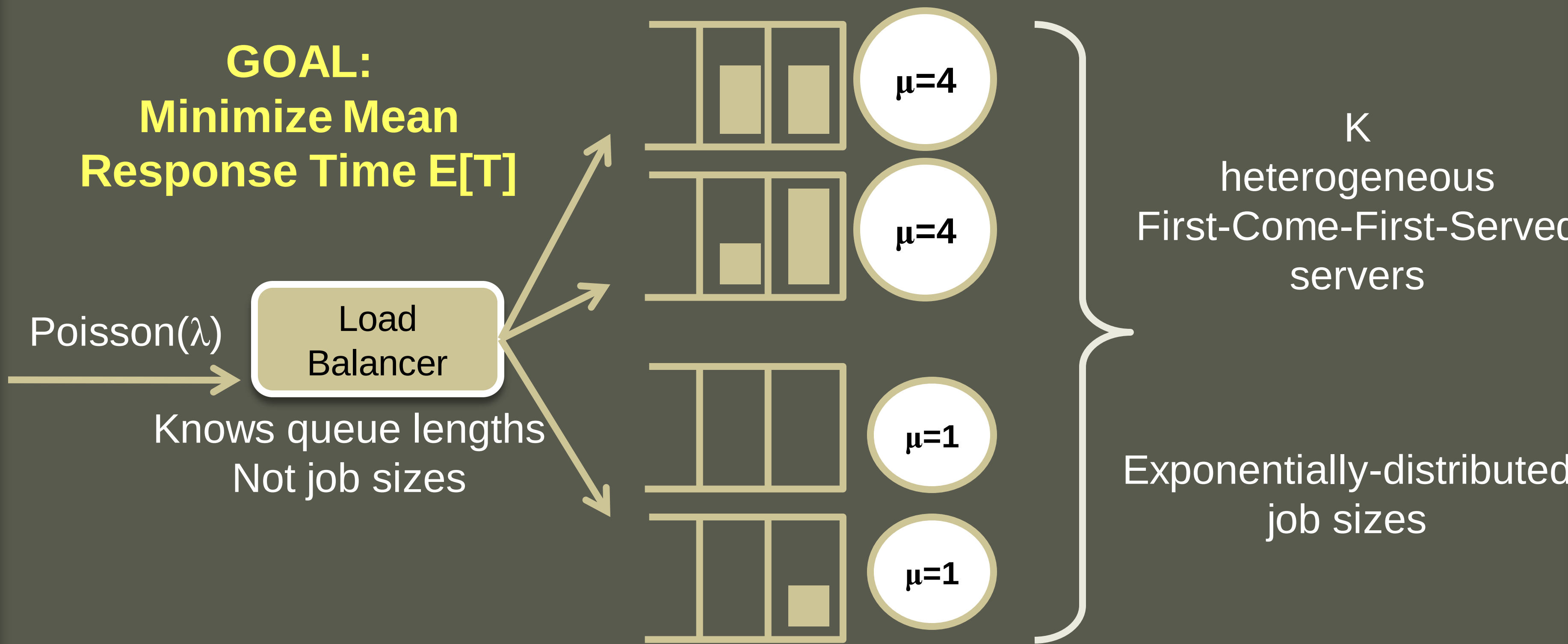


Q: What is the optimal load balancing policy?
A: **J**oin-the-**S**hortest-**Q**ueue

Q: Why?
A: JSQ = **M**inimize **E**xpected **R**esponse time of arrival

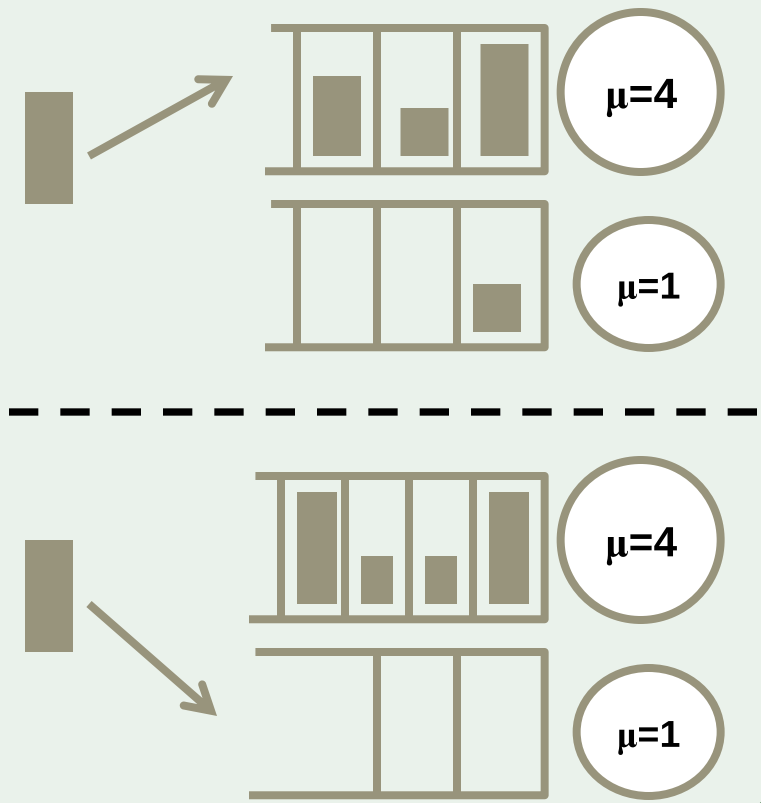
This Talk

GOAL:
Minimize Mean Response Time E[T]

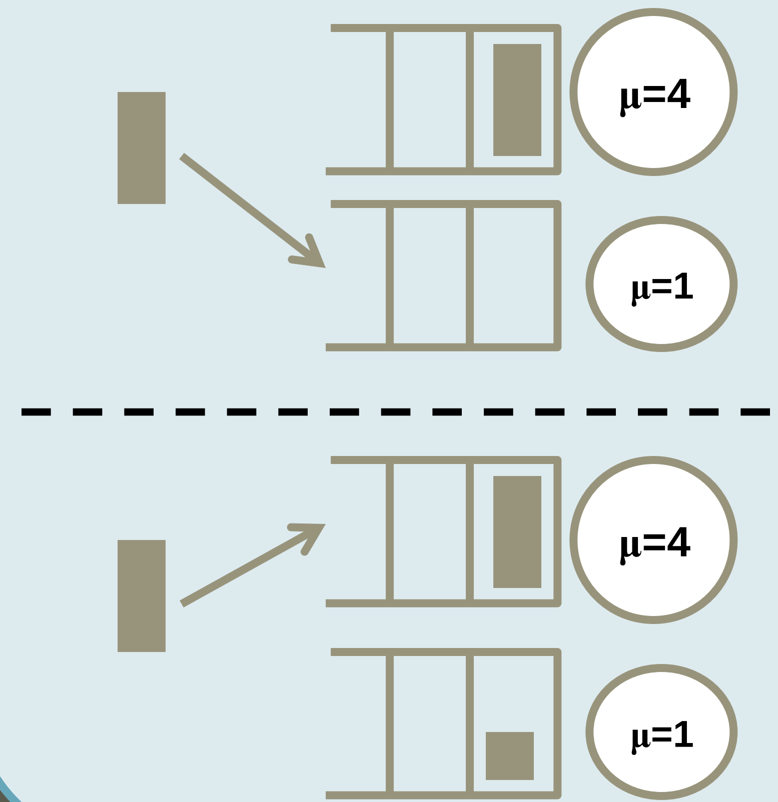


Q: What is the optimal load balancing policy?

MER = **M**inimum **E**xpected **R**esponse time



Smart-JSQ = **J**oin-**S**hortest-**Q**ueue (with **smart** tie breaks)



Q: Which is the better policy?
Q: What is the optimal policy?

Outline

Many-servers limit: $K \to \infty$

Light-traffic regime

$$\frac{\lambda}{\text{capacity}} \to \text{constant}$$

$\Rightarrow$ Partial characterization of the optimal policy

Heavy-traffic regime

$$\text{capacity} - \lambda \to \text{constant}$$

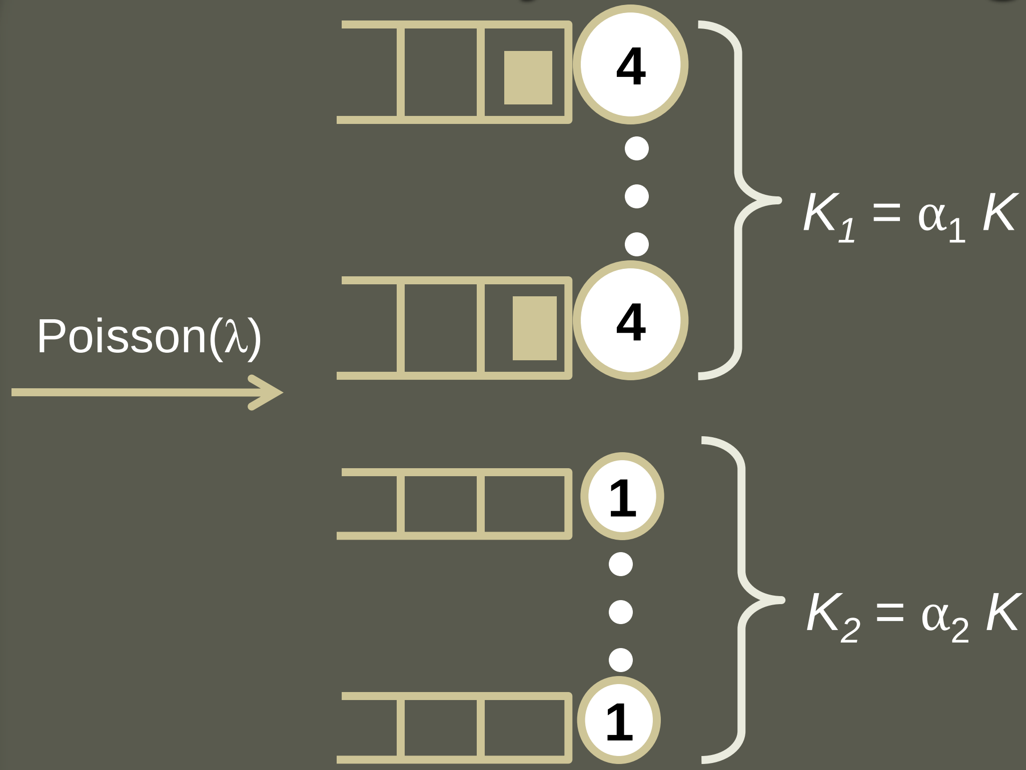
$\Rightarrow$ Complete characterization of optimal policies

$\Rightarrow$ First asymptotic approximations

Simulation Results

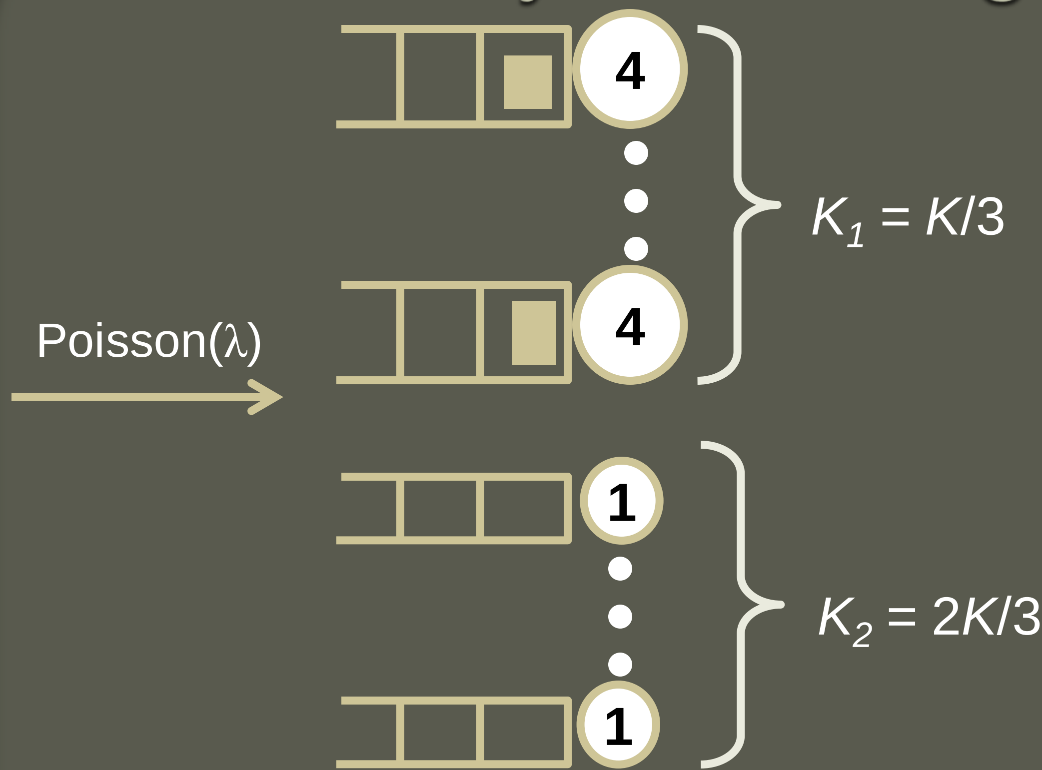
- Effect of K
- Effect of arrival rate (λ)
- Effect of degree of heterogeneity

Many-servers light-traffic limit



$$
\begin{aligned}
K &\to \infty \\
\frac{\lambda}{K} &\to \beta \\
\alpha_1, \alpha_2, \beta &\to \text{constant}
\end{aligned}
$$

Many-servers light-traffic limit



$$\begin{aligned} K &\to \infty \\ \frac{\lambda}{K} &\to \beta \\ \alpha_1, \alpha_2, \beta &\to \text{constant} \end{aligned}$$

Q: Performance of **MER**

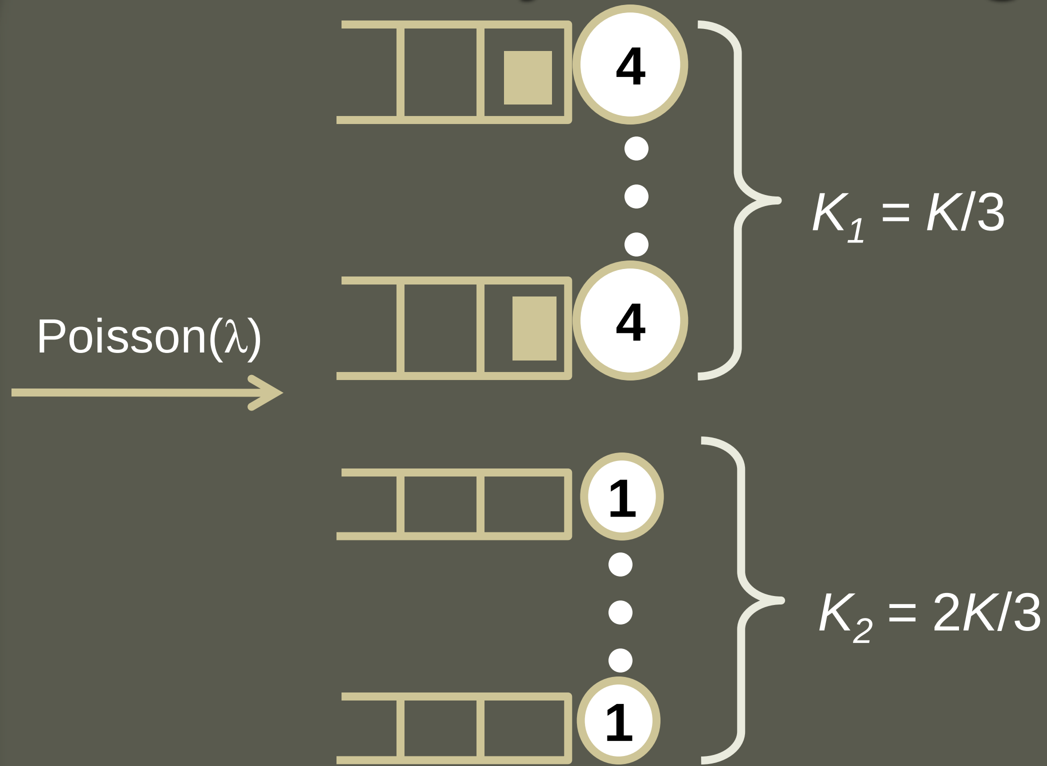
Case 1: $\lambda < 4K/3$

- Fast can handle λ
- Arrivals find at least one fast idle
- $\Rightarrow$ **E**[T] = 1/4

Case 2: $\lambda > 4K/3$

- Fast **can not** handle λ
- Can not use slow until each fast has 3 jobs !

Many-servers light-traffic limit



$$
\begin{array}{rcl}
K & \to & \infty \\
\dfrac{\lambda}{K} & \to & \beta \\
\alpha_1, \alpha_2, \beta & \to & \text{constant}
\end{array}
$$

Q: Performance of Smart-JSQ

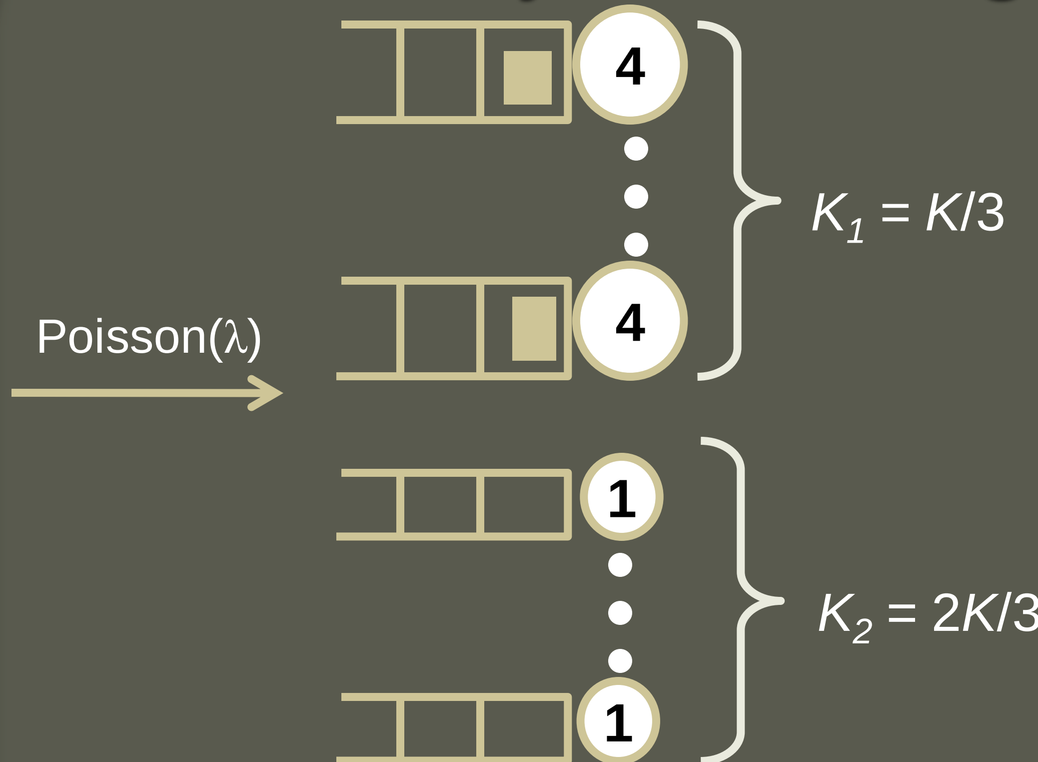
Case 1: $\lambda < 4K/3$

- Fast can handle λ
- Arrivals find at least one fast idle
- $\Rightarrow$ **E**[T] = 1/4

Case 2: $\lambda > 4K/3$

Use slow as soon as each fast has 1 job !

Many-servers light-traffic limit

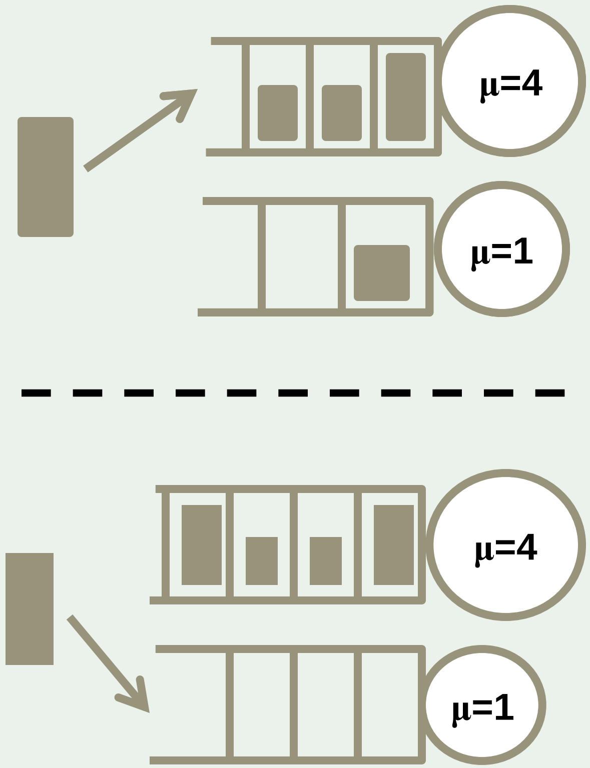


$$
\begin{array}{rcl}
K & \to & \infty \\
\dfrac{\lambda}{K} & \to & \beta \\
\alpha_1, \alpha_2, \beta & \to & \text{constant}
\end{array}
$$

Smart-JSQ better than MER!

…but any policy which sends to slow when all fast are busy is identical in light-traffic

MER

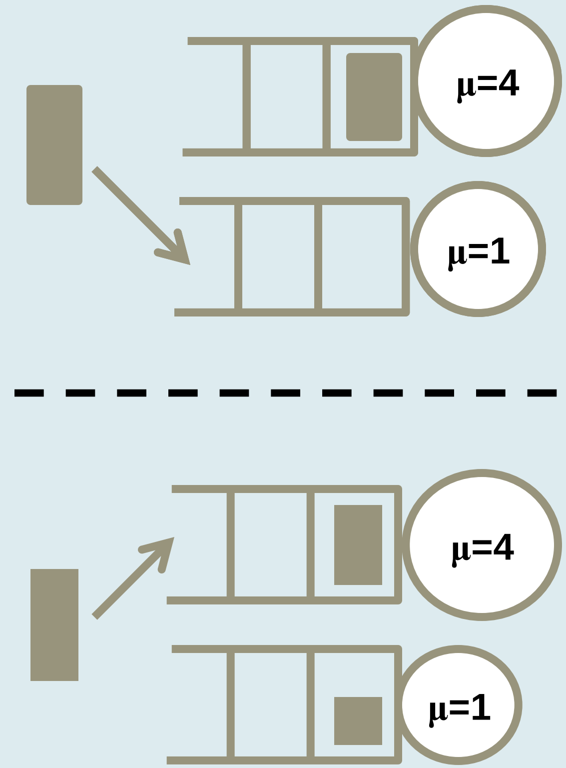


HYBRID (smart-JSQ+MER)

smart-JSQ when some server idle

MER when all busy

Smart-JSQ



Light-traffic $\Rightarrow$ HYBRID = Smart-JSQ

Outline

Many-servers limit: $K \to \infty$

Light-traffic regime

$$\frac{\lambda}{\text{capacity}} \to \text{constant}$$

$\Rightarrow$ Partial characterization of the optimal policy

Heavy-traffic regime

$$\text{capacity} - \lambda \to \text{constant}$$

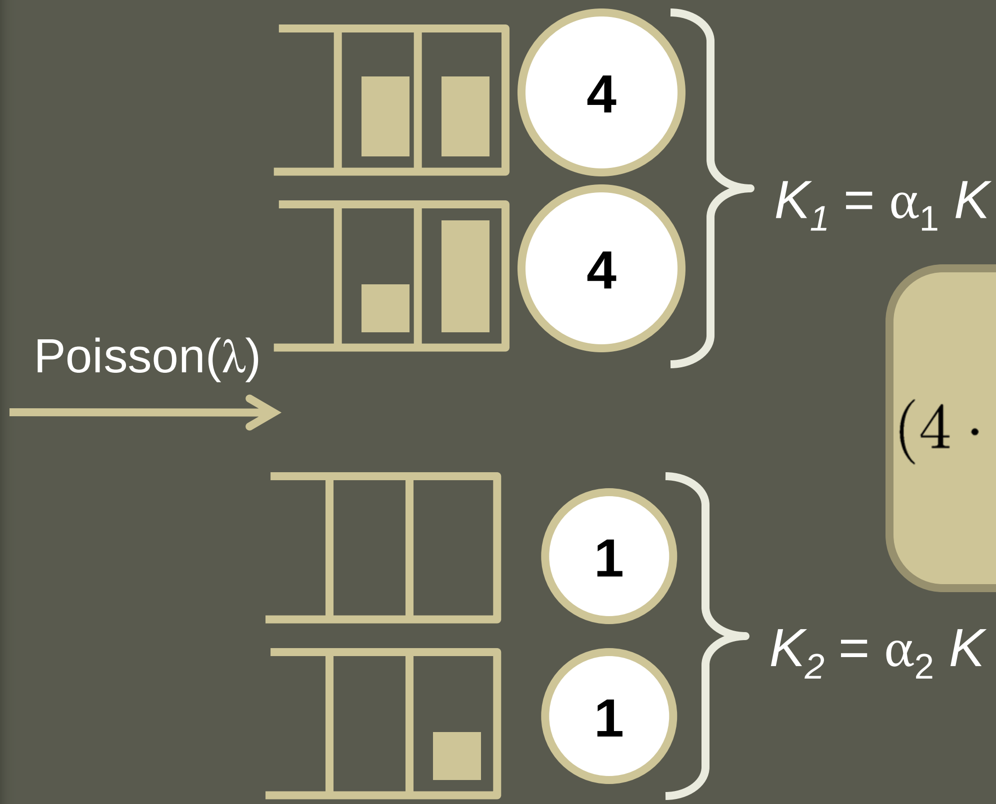
$\Rightarrow$ Complete characterization of optimal policies

$\Rightarrow$ First asymptotic approximations

Simulation Results

- Effect of K
- Effect of arrival rate (λ)
- Effect of degree of heterogeneity

Many-servers heavy-traffic limit



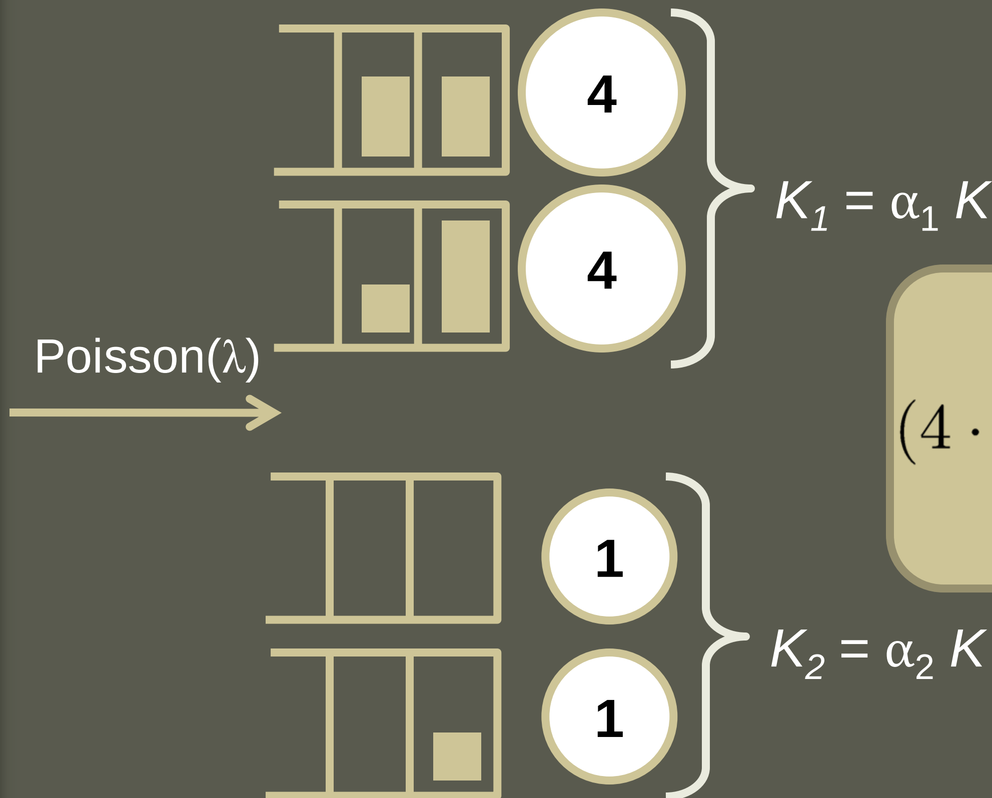
$$\begin{aligned} K &\to \infty \\ (4 \cdot \alpha_1 + 1 \cdot \alpha_2)K - \lambda &\to \theta \\ \alpha_1, \alpha_2, \theta &\to \text{const} \end{aligned}$$

Analysis of JSQ for homogeneous server

GOAL

Analysis of policies for heterogeneous servers

Many-servers heavy-traffic limit



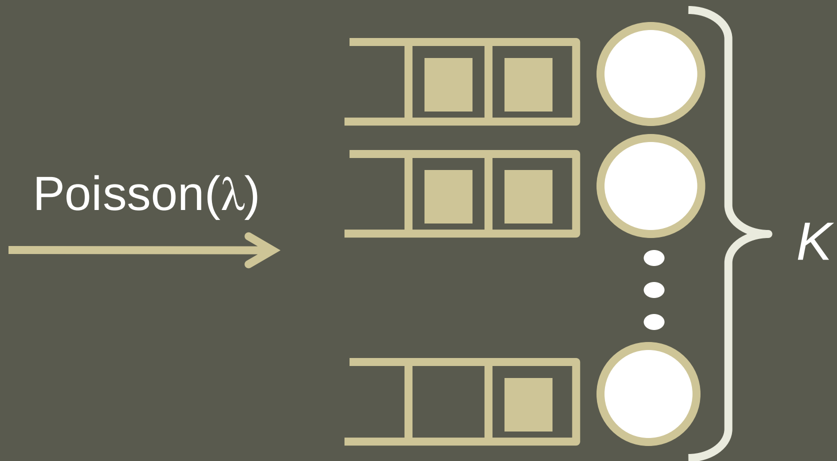
$$
\begin{aligned}
K &\to \infty \\
(4 \cdot \alpha_1 + 1 \cdot \alpha_2)K - \lambda &\to \theta \\
\alpha_1, \alpha_2, \theta &\to \text{const}
\end{aligned}
$$

Analysis of JSQ for homogeneous server

GOAL

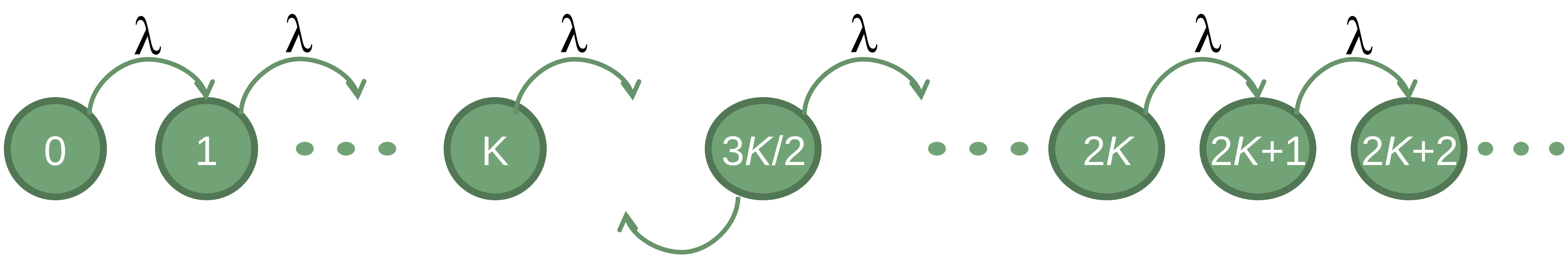
Analysis of policies for heterogeneous servers

Many-servers heavy-traffic analysis for homogenous JSQ



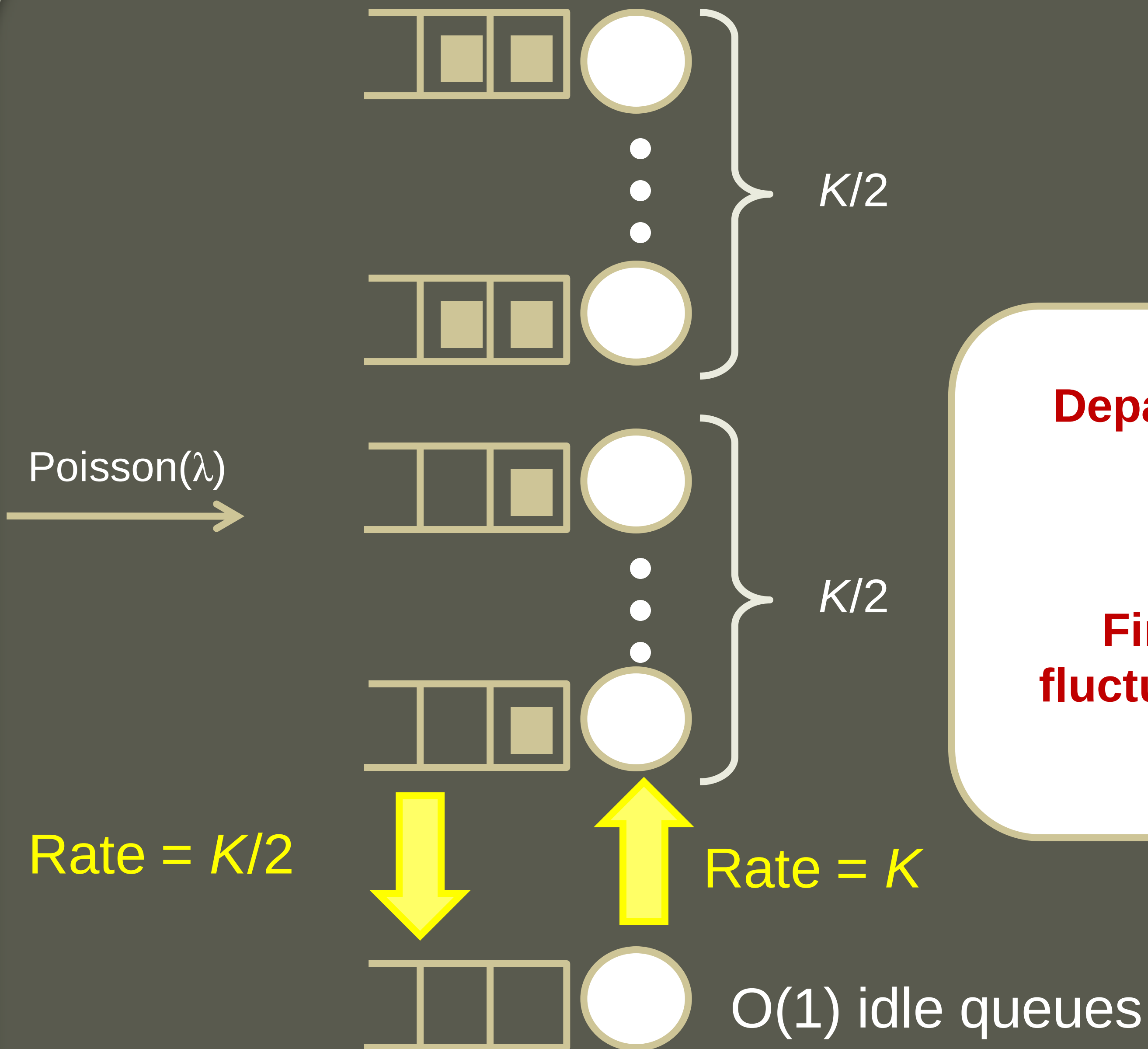
$$
\begin{aligned}
K &\rightarrow \infty \\
K - \lambda &\rightarrow \theta \\
\theta &\rightarrow \text{const}
\end{aligned}
$$

Analysis technique: Markov chain for total jobs in system



?

= mean departure rate **given** 3K/2 jobs

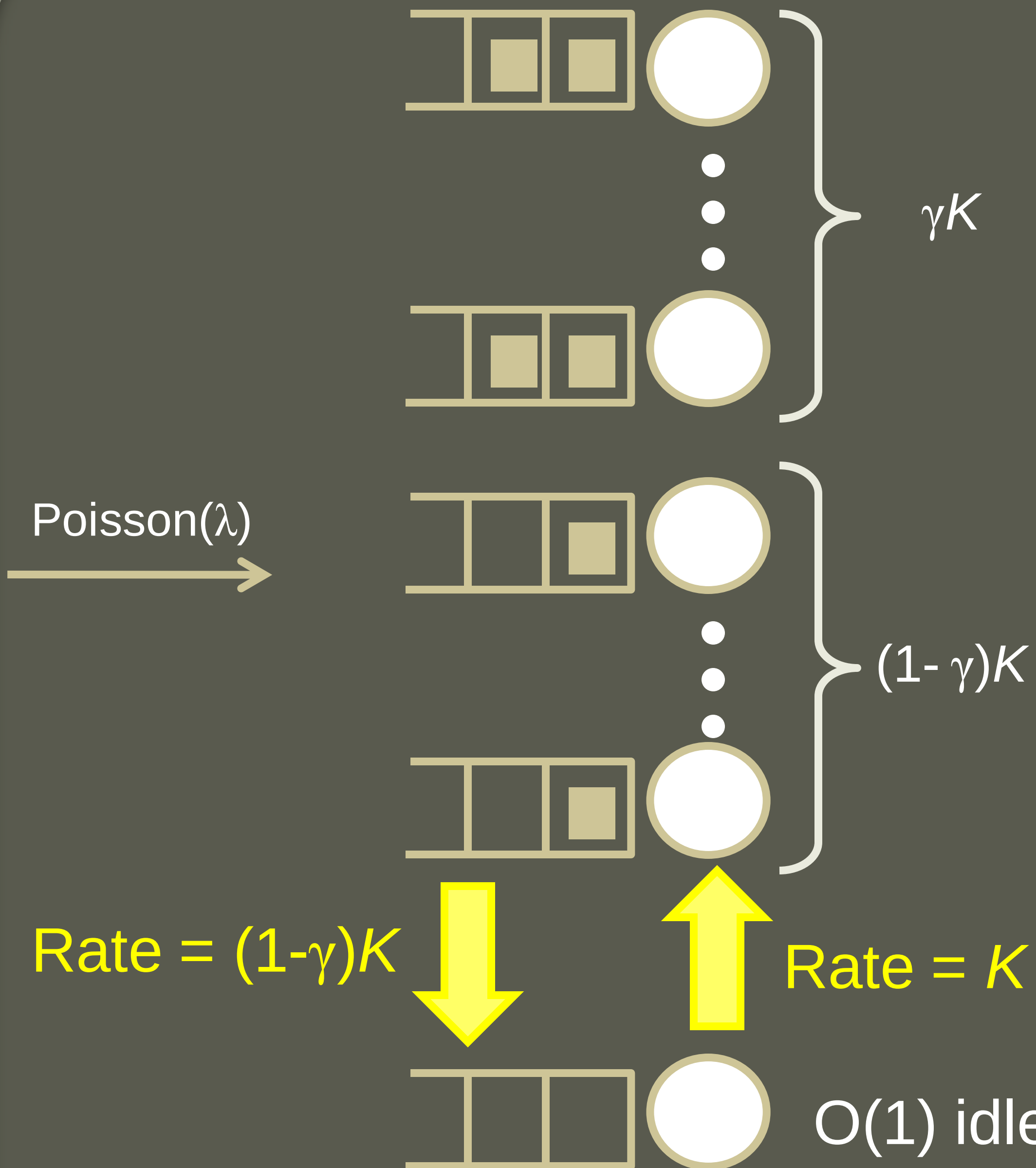


$N = 3K/2$

Departure rate = K–1

(not K)

Finding the O(1) fluctuations critical to analysis



$N = (1+\gamma)K$

$(0<\gamma<1)$

Departure rate = K–(1-γ)/ γ

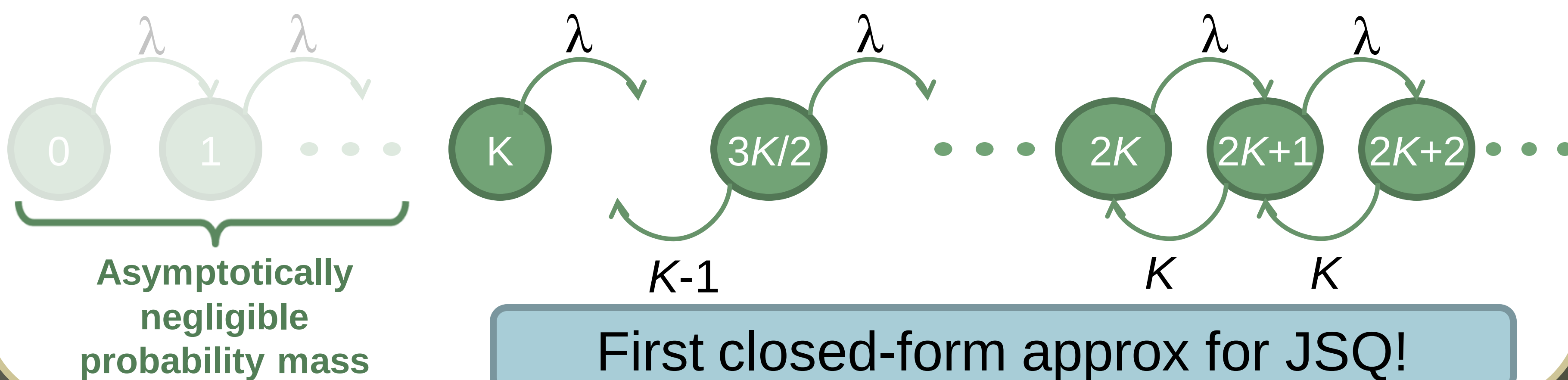
(not K)

Finding the O(1) fluctuations critical to analysis

Many-servers heavy-traffic analysis for homogenous JSQ

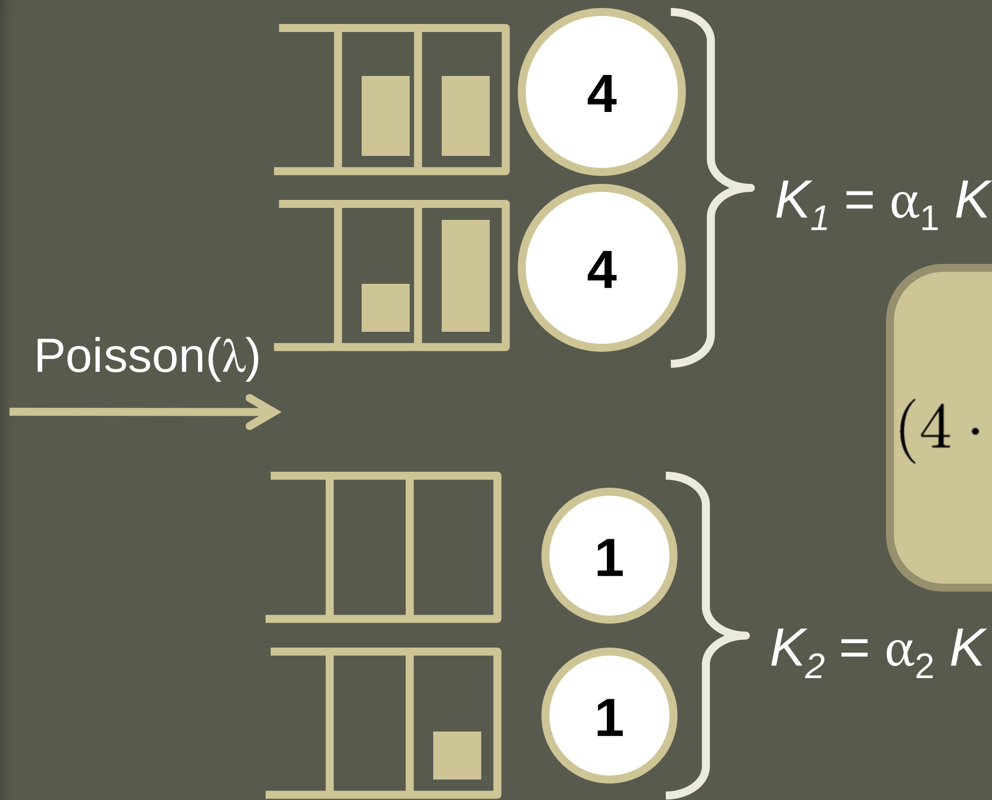
$$
\begin{aligned}
K &\to \infty \\
K - \lambda &\to \theta \\
\theta &\to \text{const}
\end{aligned}
$$

Analysis technique: Markov chain for total jobs in system



First closed-form approx for JSQ!

Many-servers heavy-traffic limit

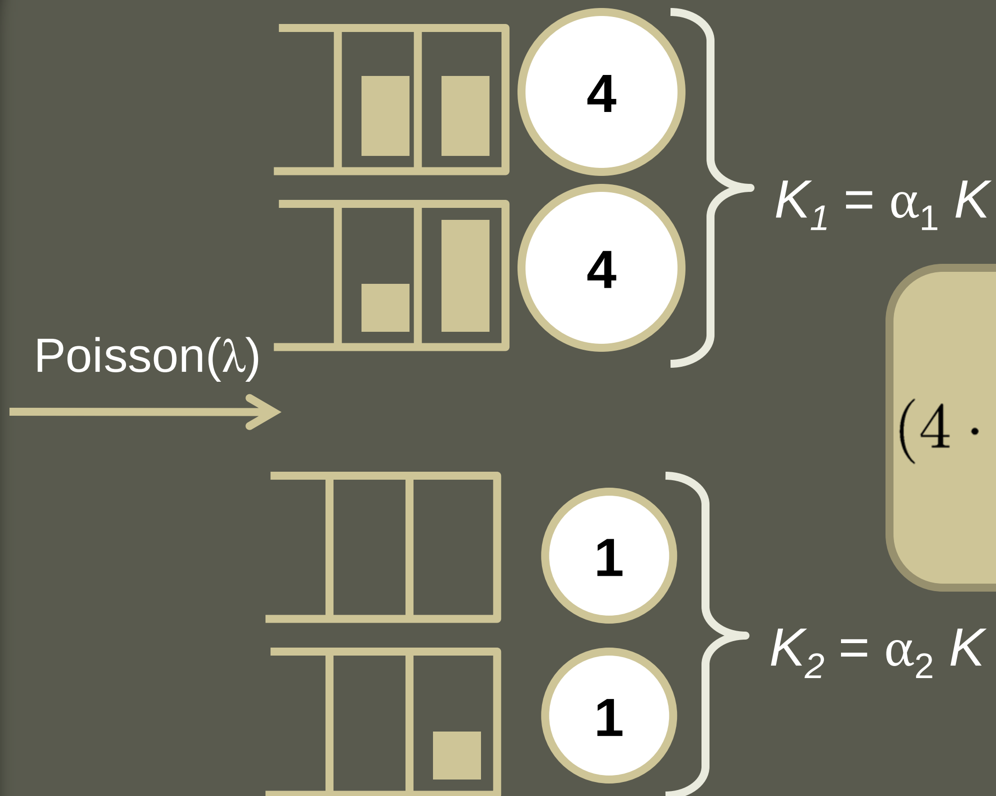


$$\begin{aligned} K &\to \infty \\ (4\cdot\alpha_1 + 1\cdot\alpha_2)K - \lambda &\to \theta \\ \alpha_1, \alpha_2, \theta &\to \text{const} \end{aligned}$$

Analysis of JSQ for homogeneous server

GOAL
Analysis of policies for heterogeneous servers

Many-servers heavy-traffic limit



$$\begin{aligned} K &\to \infty \\ (4\cdot\alpha_1 + 1\cdot\alpha_2)K - \lambda &\to \theta \\ \alpha_1, \alpha_2, \theta &\to \mathrm{const} \end{aligned}$$

OPT policy ⇒ maximize departure rate *for each* N ⇒ (preemptively) send jobs to slow servers even when they have 1 job and all fast servers have > 1

Smart-JSQ is optimal in many-servers

Outline

Many-servers limit: $K \to \infty$

Light-traffic regime

$$\frac{\lambda}{\text{capacity}} \to \text{constant}$$

$\Rightarrow$ Partial characterization of the optimal policy

Heavy-traffic regime

$$\text{capacity} - \lambda \to \text{constant}$$

$\Rightarrow$ Complete characterization of optimal policies

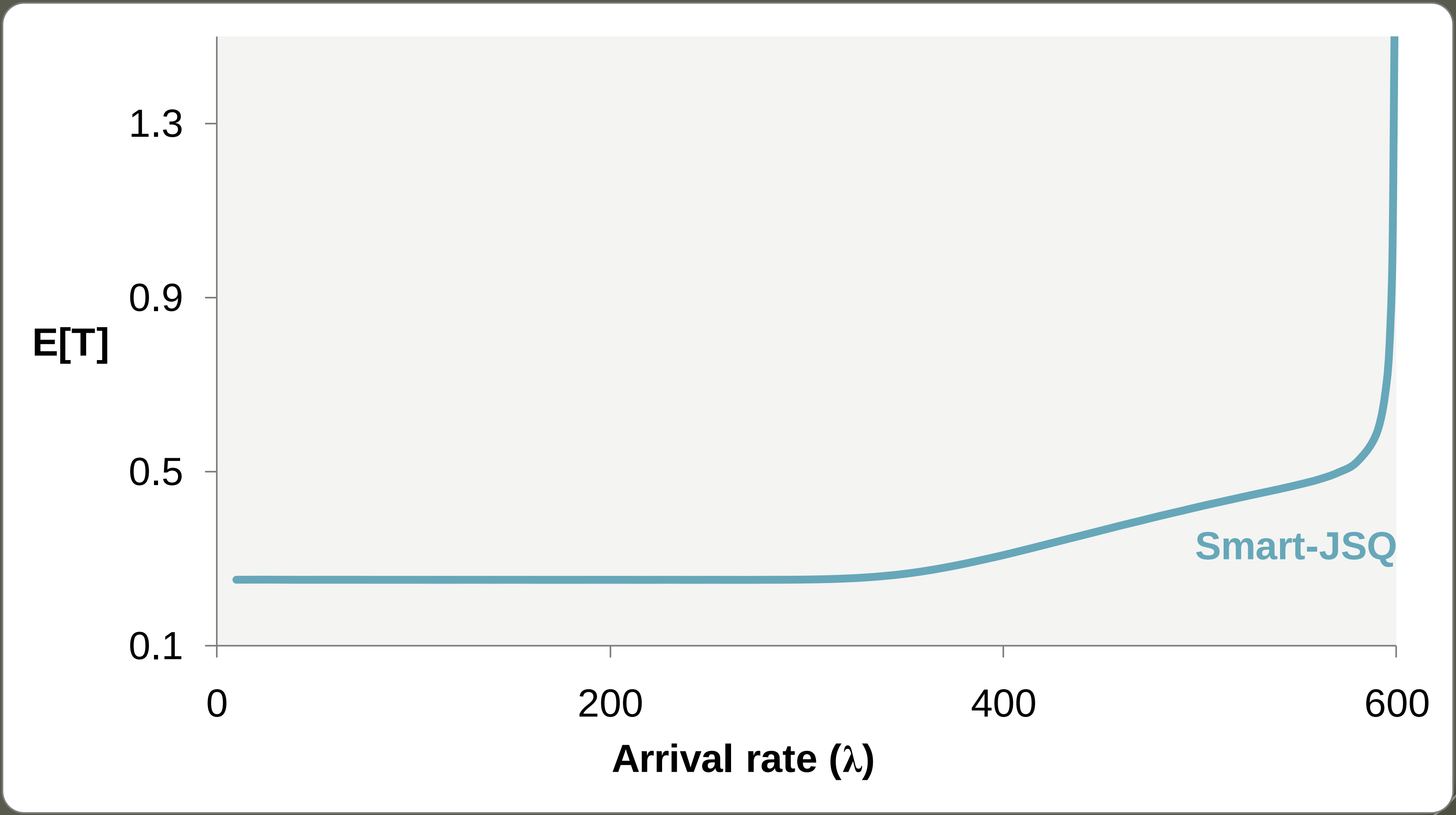
$\Rightarrow$ First asymptotic approximations

Simulation Results

- Effect of K
- Effect of arrival rate (λ)
- Effect of degree of heterogeneity

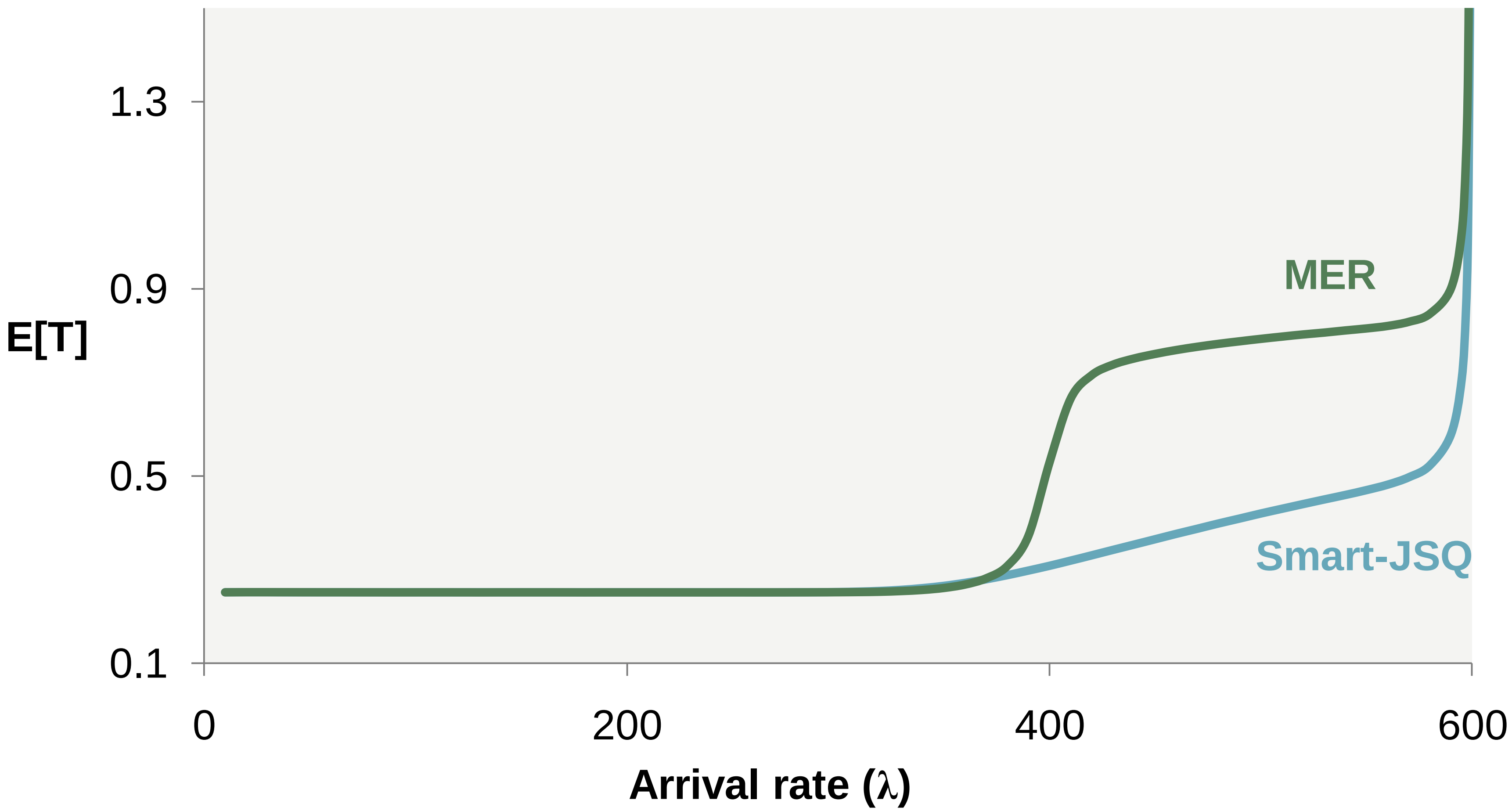
Many-servers light-traffic

$\mu_1=4,\ \mu_2=1,\ K_1=100,\ K_2=200$



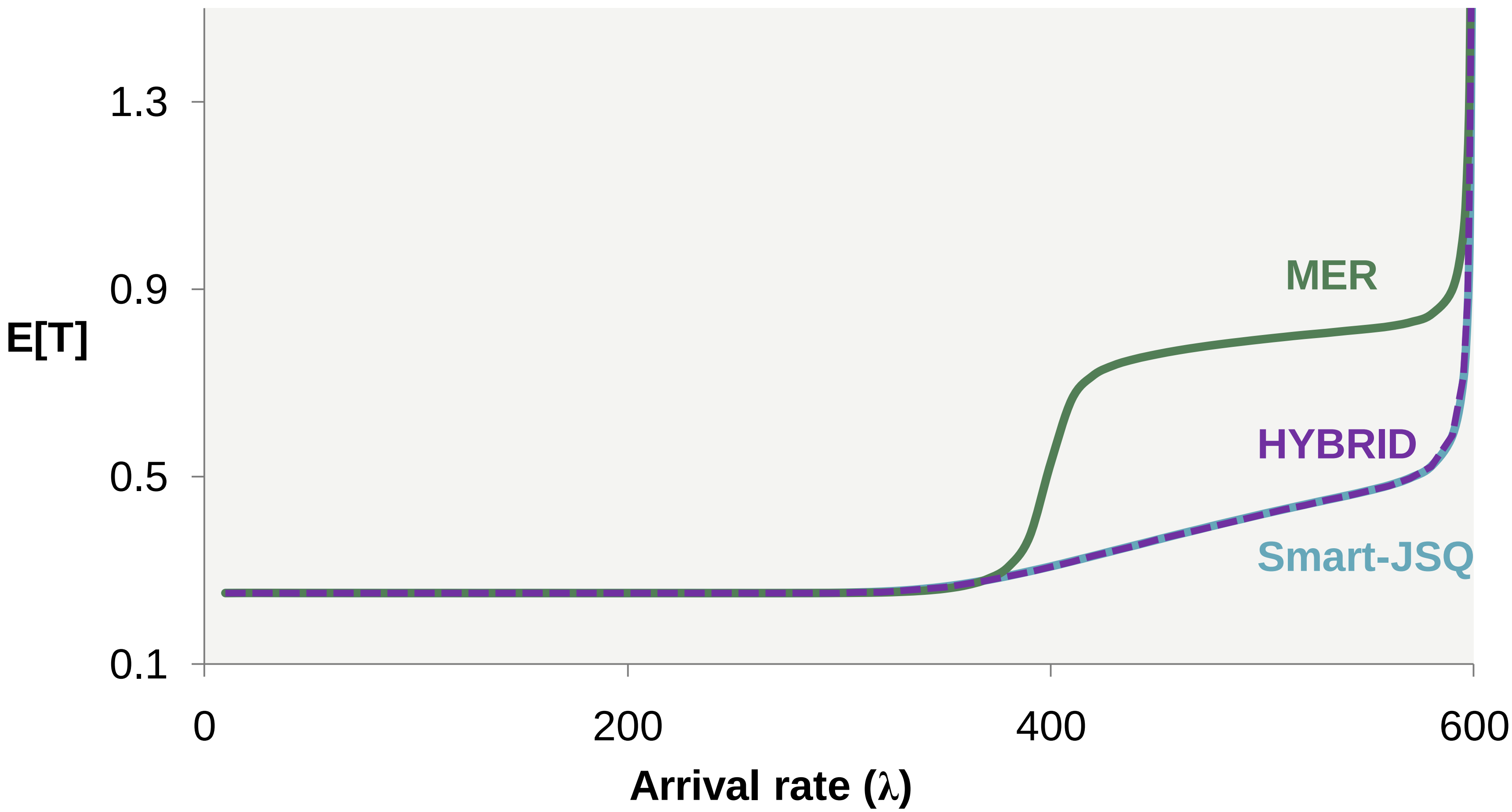
Many-servers light-traffic

$\mu_1=4, \mu_2=1, K_1=100, K_2=200$



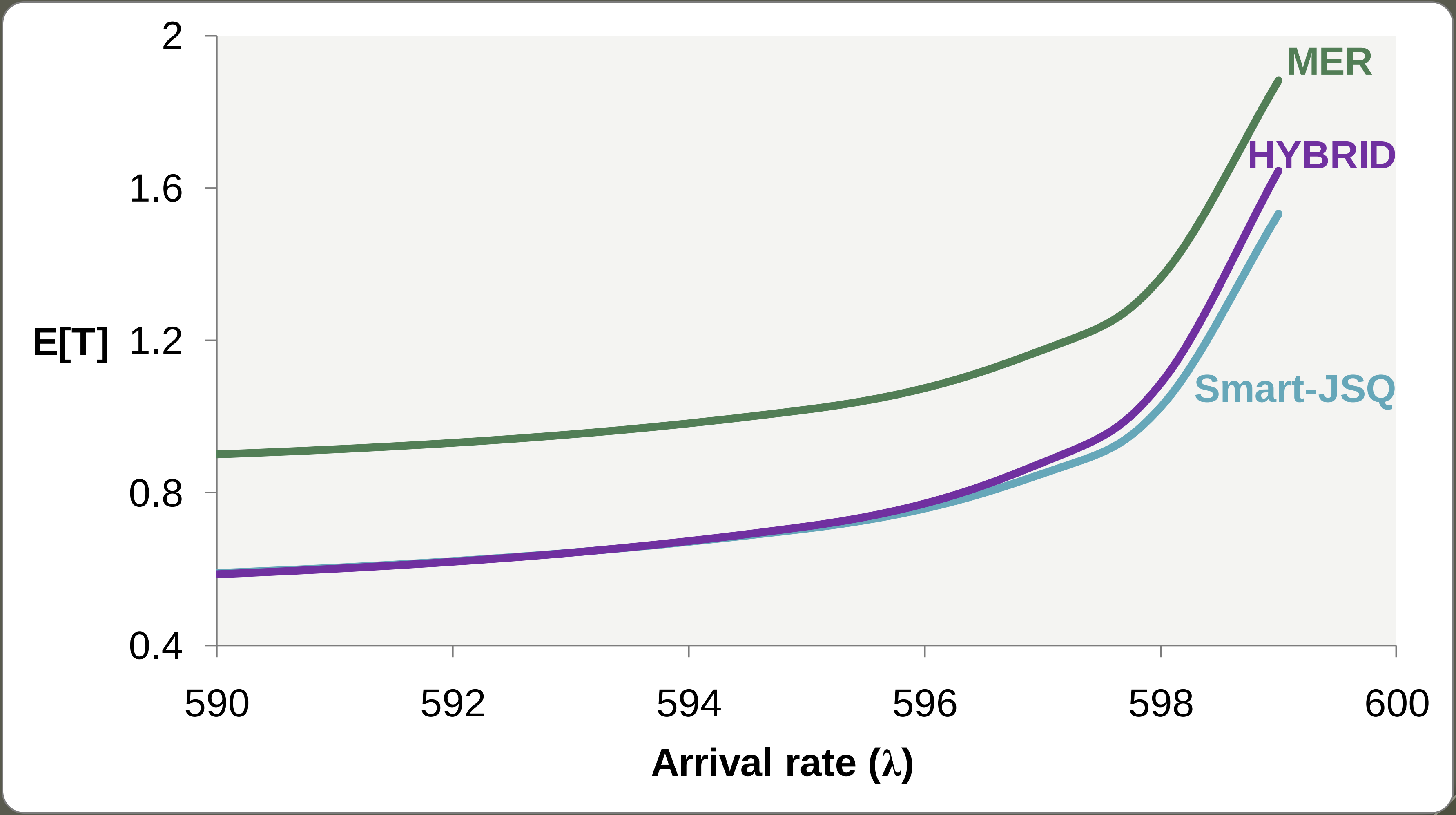
Many-servers light-traffic

$\mu_1=4$, $\mu_2=1$, $K_1=100$, $K_2=200$



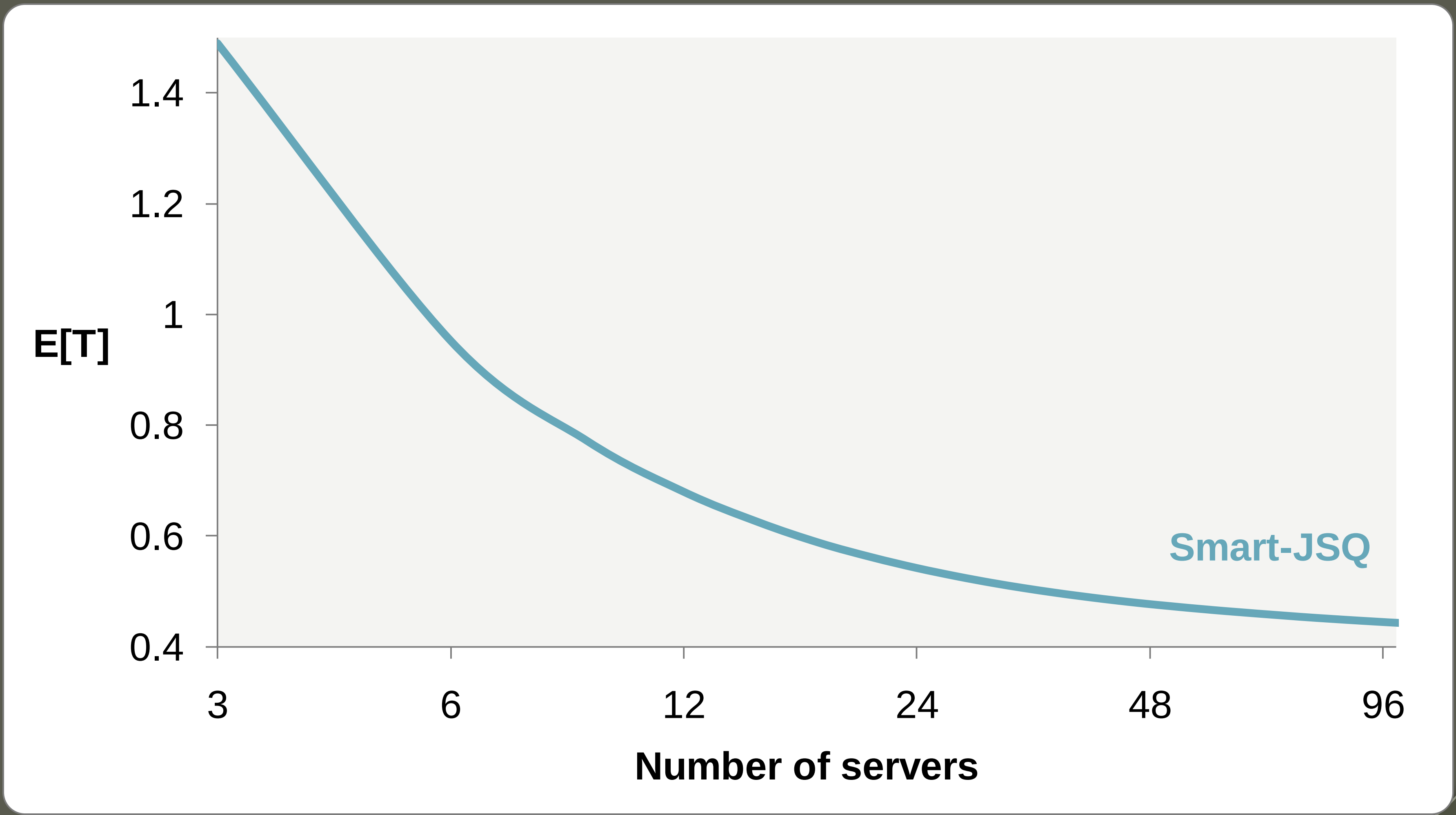
Many-servers heavy-traffic

$\mu_1=4$, $\mu_2=1$, $K_1=100$, $K_2=200$



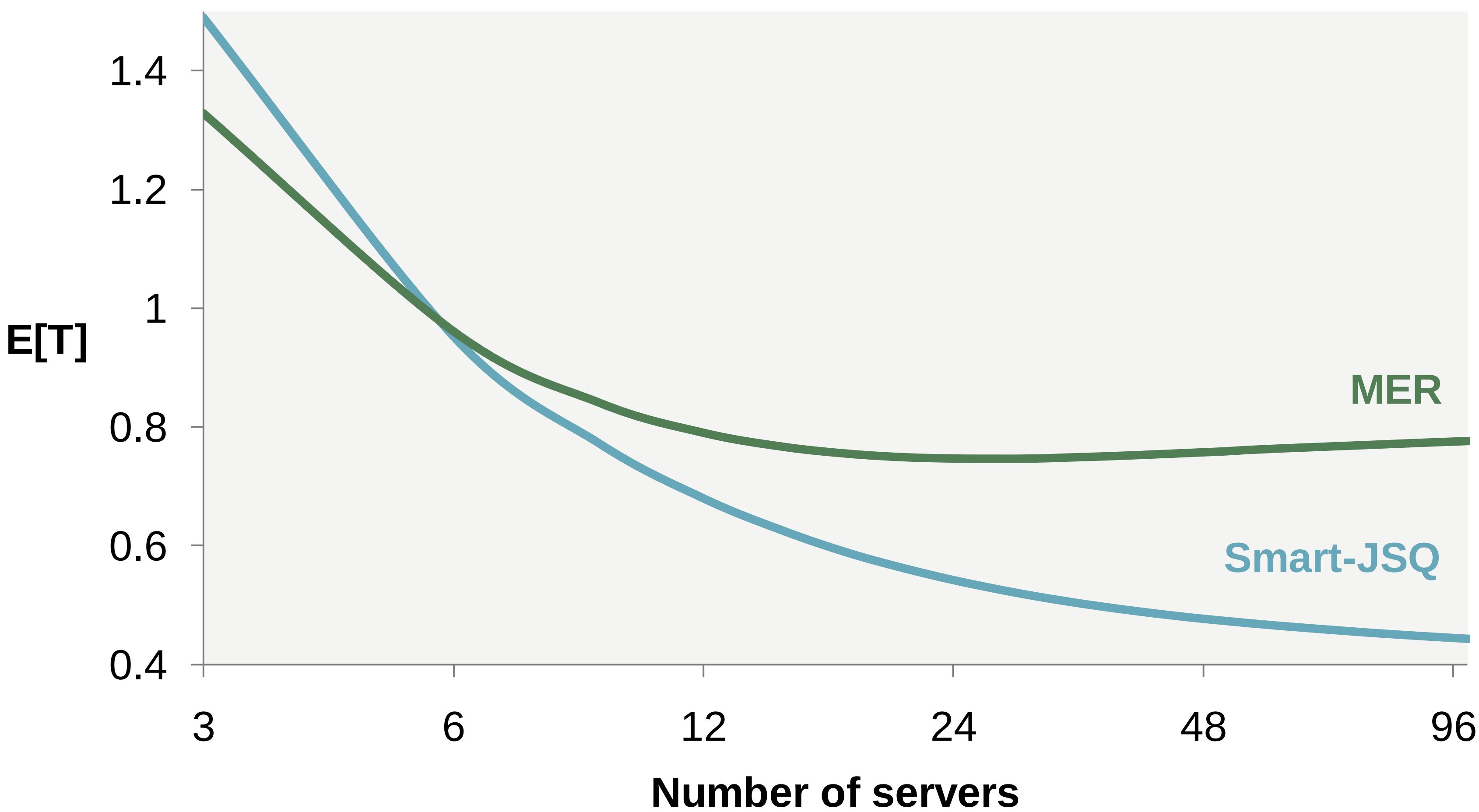
Effect of number of servers

$\mu_1=4,\ \mu_2=1,\ \alpha_1=1/3,\ \alpha_2=2/3$



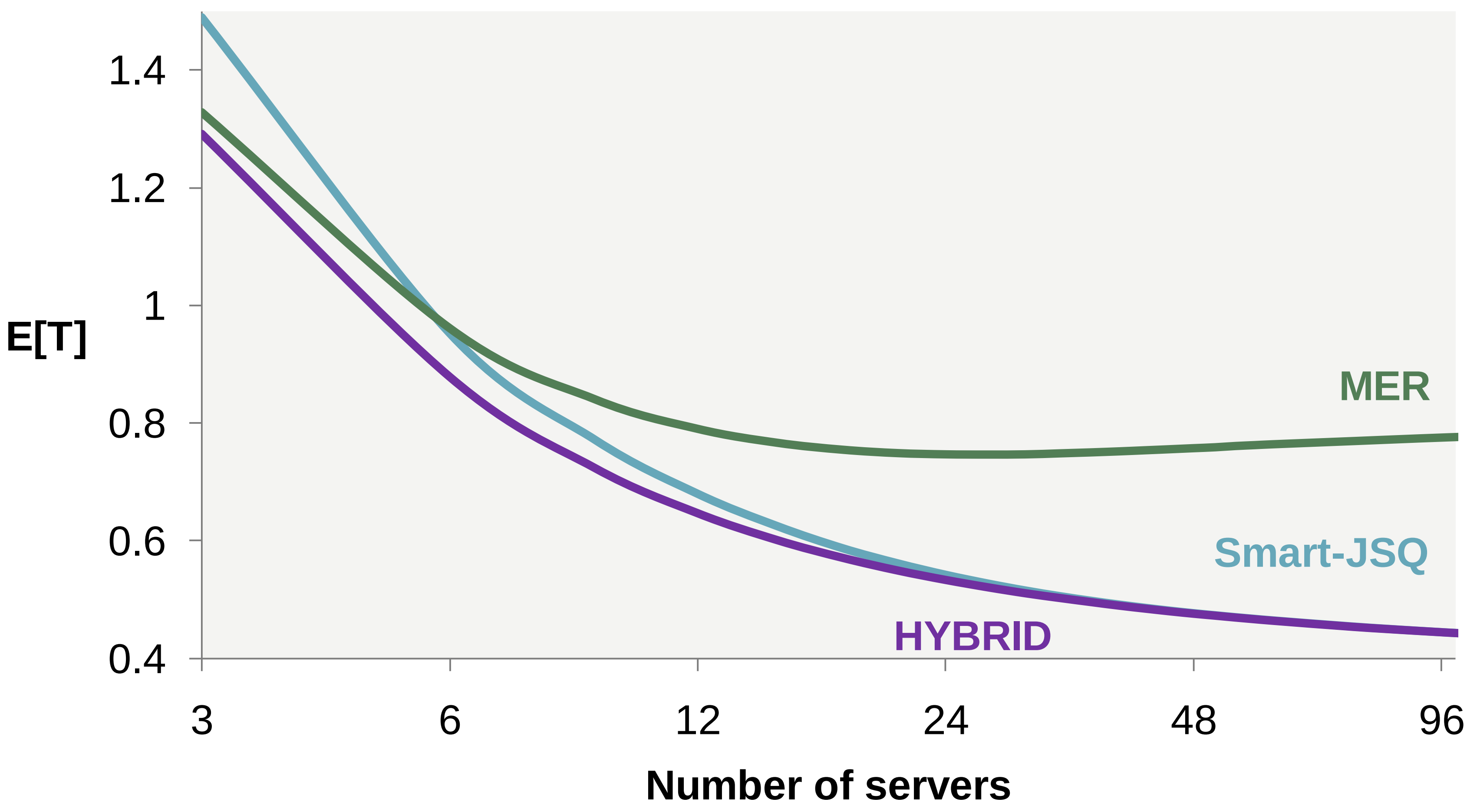
Effect of number of servers

$\mu_1=4,\ \mu_2=1,\ \alpha_1=1/3,\ \alpha_2=2/3$



Effect of number of servers

$\mu_1=4, \mu_2=1, \alpha_1=1/3, \alpha_2=2/3$

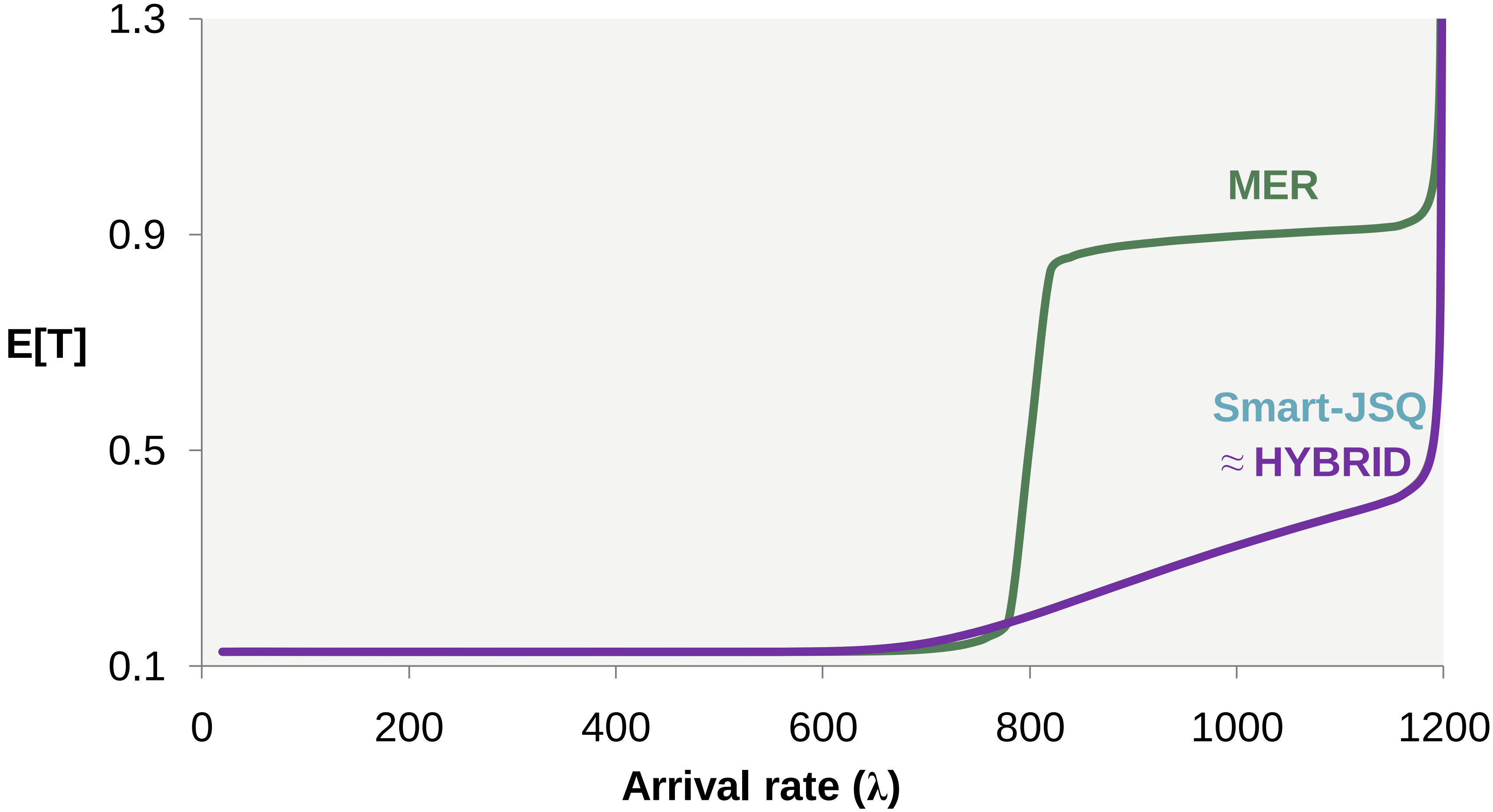


Conclusions

- A new many-servers heavy-traffic scaling to analyze load balancing policies
- First closed-form approx of load balancing heuristics
- Choosing the right load balancer
  - Few servers, Small load, High heterogeneity $\Rightarrow$ HYBRID
  - Many servers, High load, Low heterogeneity $\Rightarrow$ smart-JSQ

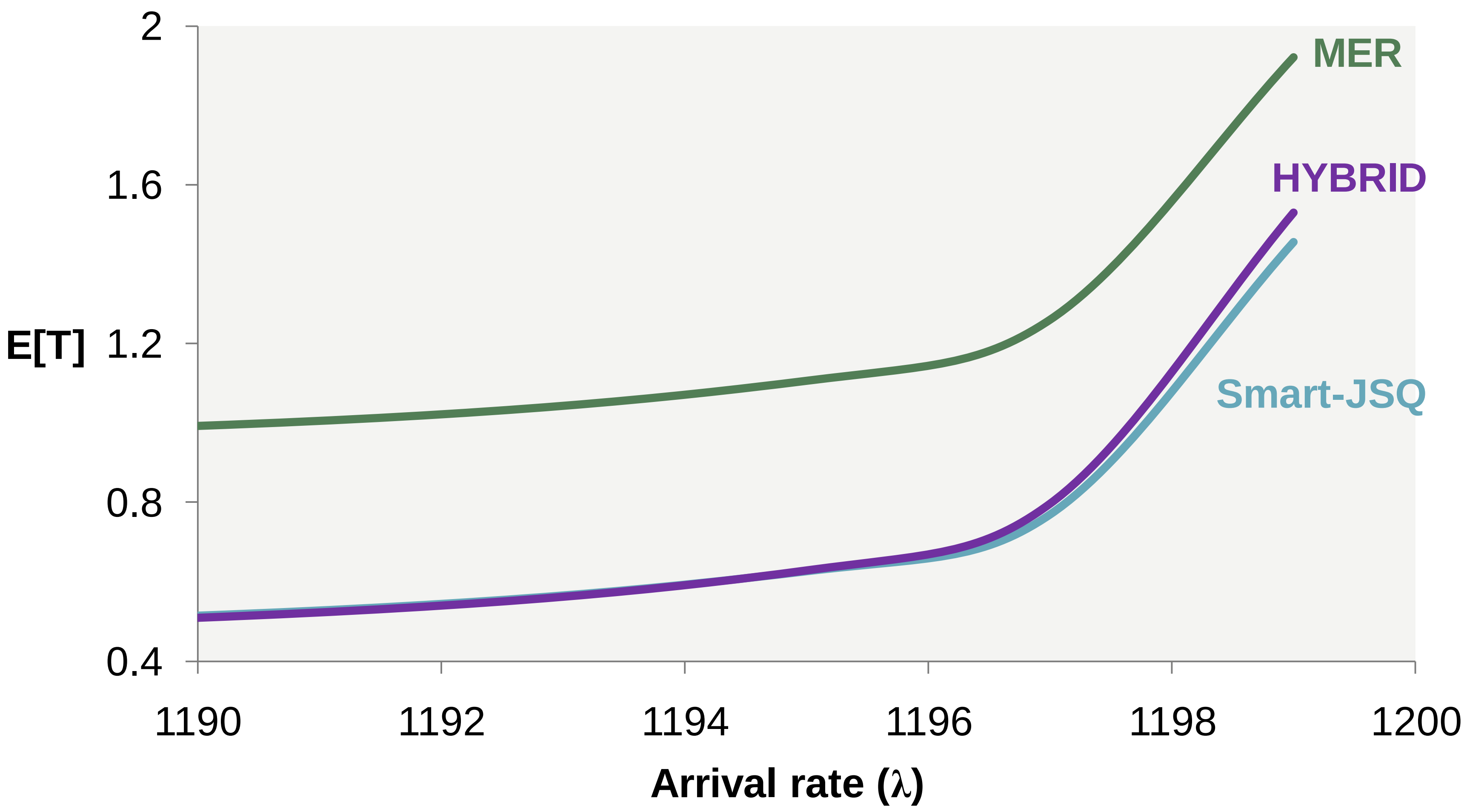
Many-servers light-traffic

$\mu_1=8,\ \mu_2=1,\ K_1=100,\ K_2=400$



Many-servers heavy-traffic

$\mu_1=8,\ \mu_2=1,\ K_1=100,\ K_2=400$



Effect of number of servers

$\mu_1=8, \mu_2=1, \alpha_1=1/5, \alpha_2=4/5$