# Simple and explicit bounds for multi-server queues with universal $\frac{1}{1-\rho}$ scaling

## David A. Goldberg

Cornell

LNMB

(日) (日) (日) (日) (日) (日) (日)

Punchline	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
Outline					





#### **3** History



## 5 Proof



▲□▶▲圖▶▲≣▶▲≣▶ ≣ のQ@

Punchline ●○○○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
Outline					



2 Model

#### 3 History

Main results

## 5 Proof



Punchline o●oo	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

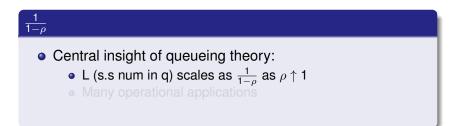
▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ - 三 - のへぐ

# $\frac{1}{1-\rho}$

#### • Central insight of queueing theory:

L (s.s num in q) scales as <sup>1</sup>/<sub>1-ρ</sub> as ρ ↑ 1
 Many operational applications

Punchline ○●○○	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					



< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

Punchline ○●○○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

# $\frac{1}{1-\rho}$

Central insight of queueing theory:

- L (s.s num in q) scales as <sup>1</sup>/<sub>1-ρ</sub> as ρ ↑ 1
   Many operational applications

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1- ho}$ ?					

#### Only rigorously justified for G/G/n in a few special cases!

- Single server
- Exponential or deterministic service times
- Special asymptotic regimes
- Far less known is known when it comes to ...
- The exception is Kingman's bound, but ...
- A major difficulty is that any such bound ....
- Multi-server Kingman's bound open for 50 years!

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					
$\frac{1}{1-2}$ ?					

#### • Only rigorously justified for G/G/n in a few special cases!

- Single server
- Exponential or deterministic service times
- Special asymptotic regimes
- Far less known is known when it comes to ...
- The exception is Kingman's bound, but . . .
- A major difficulty is that any such bound ....
- Multi-server Kingman's bound open for 50 years!

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to . .
- The exception is Kingman's bound, but . . .
- A major difficulty is that any such bound ....
- Multi-server Kingman's bound open for 50 years!

Punchline ○○●○	Model oo	History 0000000000000	Main results	Proof oo	Conclusion
$\frac{1}{1-\rho}$ ?					

- $\frac{1}{1-\rho}$ ?
  - Only rigorously justified for G/G/n in a few special cases!
    - Single server
    - Exponential or deterministic service times
    - Special asymptotic regimes
  - Far less known is known when it comes to ...

Simple, explicit, non-asymptotic bounds

- The exception is Kingman's bound, but ...
- A major difficulty is that any such bound ....

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
- A major difficulty is that any such bound ....
- Multi-server Kingman's bound open for 50 years!

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
- A major difficulty is that any such bound ....

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
  - Only for single server
- A major difficulty is that any such bound ....

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
  - Only for single server
- A major difficulty is that any such bound ...

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
  - Only for single server
- A major difficulty is that any such bound ...
  - Must scale as  $\frac{1}{1-\rho}$  even if  $n \to \infty$  as  $\rho \uparrow 1$
- Multi-server Kingman's bound open for 50 years!

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
  - Only for single server
- A major difficulty is that any such bound ...
  - Must scale as  $\frac{1}{1-\rho}$  even if  $n \to \infty$  as  $\rho \uparrow 1$

Punchline ○○●○	Model 00	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ ?					

- Only rigorously justified for G/G/n in a few special cases!
  - Single server
  - Exponential or deterministic service times
  - Special asymptotic regimes
- Far less known is known when it comes to ...
  - Simple, explicit, non-asymptotic bounds
- The exception is Kingman's bound, but ...
  - Only for single server
- A major difficulty is that any such bound ...
  - Must scale as  $\frac{1}{1-\rho}$  even if  $n \to \infty$  as  $\rho \uparrow 1$
- Multi-server Kingman's bound open for 50 years!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof oo	Conclusion
$\frac{1}{1-\rho}$					

#### Our main result resolves this open question

- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-c}$
- General G/G/n only requiring finite 2 +  $\epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat <sup>1</sup>/<sub>1-c</sub>
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof oo	Conclusion
$\frac{1}{1-\rho}$					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as <sup>1</sup>/<sub>1-a</sub>
- General G/G/n only requiring finite 2 +  $\epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat <sup>1</sup>/<sub>1-6</sub>
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ !					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-a}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat <sup>1</sup>/<sub>1-</sub>
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ !					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat <sup>1</sup>/<sub>1-a</sub>
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$ !					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite 2 + 
   e moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat  $\frac{1}{1-}$
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat  $\frac{1}{1-\rho}$
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-a</sub> heuristic for multi-server queues
- Proof of concept , work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat  $\frac{1}{1-\rho}$
- Implications for Halfin-Whitt regime
- Broadly justifies the <sup>1</sup>/<sub>1-ρ</sub> heuristic for multi-server queues
   Proof of concept, work to do!

Punchline ○○○●	Model oo	History 0000000000000	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat  $\frac{1}{1-a}$
- Implications for Halfin-Whitt regime
- Broadly justifies the  $\frac{1}{1-\rho}$  heuristic for multi-server queues
- Proof of concept, work to do!

Punchline ○○○●	Model oo	History ococococococo	Main results	Proof 00	Conclusion
$\frac{1}{1-\rho}$					

- Our main result resolves this open question
- Simple and explicit bounds for E[L] scaling as  $\frac{1}{1-\rho}$
- General G/G/n only requiring finite  $2 + \epsilon$  moments
- Higher moments and tails
- Steady-state probability of delay
- In some cases we even beat  $\frac{1}{1-\rho}$
- Implications for Halfin-Whitt regime
- Broadly justifies the  $\frac{1}{1-a}$  heuristic for multi-server queues
- Proof of concept , work to do!

Punchline	Model ●○	History 0000000000000	Main results	Proof 00	Conclusion
Outline					





#### 3 History



## 5 Proof



Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

- $\bullet~$  Inter-arrival times i.i.d.  $\sim A$
- Service times i.i.d.  $\sim$  S
- $\mu_A = \frac{1}{E[A]}$  ,  $\mu_S = \frac{1}{E[S]}$
- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

- $\bullet\,$  Inter-arrival times i.i.d.  $\sim\,$  A
- $\bullet~$  Service times i.i.d.  $\sim$  S
- $\mu_A = \frac{1}{E[A]}$  ,  $\mu_S = \frac{1}{E[S]}$
- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

- $\bullet\,$  Inter-arrival times i.i.d.  $\sim\,$  A
- Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

- $\bullet\,$  Inter-arrival times i.i.d.  $\sim\,$  A
- Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

## FCFS GI/GI/n Queue

- Inter-arrival times i.i.d.  $\sim A$
- Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

・ロト・日本・日本・日本・日本

## FCFS GI/GI/n Queue

- $\bullet\,$  Inter-arrival times i.i.d.  $\sim\,$  A
- $\bullet~$  Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- *P<sub>wait</sub>*: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

## FCFS GI/GI/n Queue

- Inter-arrival times i.i.d.  $\sim A$
- $\bullet~$  Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- P<sub>wait</sub>: s.s. prob. all servers busy

Punchline	Model ○●	History 0000000000000	Main results	Proof 00	Conclusion

- Inter-arrival times i.i.d.  $\sim A$
- Service times i.i.d.  $\sim$  S

• 
$$\mu_A = \frac{1}{E[A]}$$
 ,  $\mu_S = \frac{1}{E[S]}$ 

- n servers
- Traffic intensity  $\rho = \frac{\mu_A}{n\mu_S}$
- Jobs served FCFS
- L: s.s. number waiting in queue
- *P<sub>wait</sub>*: s.s. prob. all servers busy

Punchline	Model 00	History •ooooooooooooooooo	Main results	Proof 00	Conclusion
Outline					









### 5 Proof





Punchline	Model oo	History o●oooooooooooo	Main results	Proof 00	Conclusion
Early 2	0th Cer	ntury			



(a) Erlang



(b) Pollaczek



(c) Khinchin

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000			

### Early 20th Century

- Model "invented" to study early telephone networks
- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

・ロト・西ト・山田・山田・山下

Punchline	Model	History ००●००००००००००	Main results	Proof 00	Conclusion

### Early 20th Century

Model "invented" to study early telephone networks

・ コット (雪) ( 小田) ( コット 日)

- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

Punchline	Model 00	History oo●ooooooooooo	Main results	Proof 00	Conclusion

### Early 20th Century

Model "invented" to study early telephone networks

- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

Punchline	Model 00	History ००●००००००००००	Main results	Proof 00	Conclusion

### Early 20th Century

Model "invented" to study early telephone networks

- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

Punchline	Model 00	History ००●००००००००००	Main results	Proof 00	Conclusion

### Early 20th Century

Model "invented" to study early telephone networks

- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

Punchline 0000	Model 00	History ocoecococococo	Main results	Proof 00	Conclusion
Mid 20t	h Cent	ury			



(d) Spitzer



(e) Lindley



Punchline	Model	History	Main results	Proof	Conclusion
0000	00	00000000000000	00000	00	000

(日) (日) (日) (日) (日) (日) (日)

### **Mid 20th Century**

### **Mid 20th Century**

#### Great progress on single-server queue

- Lindley's recursion
- Spitzer's identity

Little progress on multi-server queue

Punchline	Model 00	History ○○○○●○○○○○○○○	Main results	Proof 00	Conclusion

### **Mid 20th Century**

- Great progress on single-server queue
  - Lindley's recursion
  - Spitzer's identity
- Little progress on multi-server queue
  - Pollaczek: Extremely complicated transforms

・ コット (雪) ( 小田) ( コット 日)

Punchline	Model	History ০০০০●০০০০০০০০	Main results	Proof 00	Conclusion

### **Mid 20th Century**

- Great progress on single-server queue
  - Lindley's recursion
  - Spitzer's identity

Little progress on multi-server queue

Kendall solves G/M/n case

Pollaczek: Extremely complicated transforms

・ コット (雪) ( 小田) ( コット 日)

Punchline	Model oo	History ○○○○●○○○○○○○○	Main results	Proof 00	Conclusion

#### Mid 20th Century

#### Great progress on single-server queue

- Lindley's recursion
- Spitzer's identity

#### Little progress on multi-server queue

- Kendall solves G/M/n case
- Pollaczek: Extremely complicated transforms

・ロト ・聞ト ・ヨト ・ヨト 三日

Punchline	Model oo	History ○○○○●○○○○○○○○	Main results	Proof 00	Conclusion

#### **Mid 20th Century**

- Great progress on single-server queue
  - Lindley's recursion
  - Spitzer's identity

#### Little progress on multi-server queue

- Kendall solves G/M/n case
- Pollaczek: Extremely complicated transforms

・ロト ・個 ト ・ ヨト ・ ヨト … ヨ

Punchline	Model	History 0000●000000000	Main results	Proof	Conclusion

#### Mid 20th Century

- Great progress on single-server queue
  - Lindley's recursion
  - Spitzer's identity
- Little progress on multi-server queue
  - Kendall solves G/M/n case
  - Pollaczek: Extremely complicated transforms

・ コット (雪) ( 小田) ( コット 日)

Punchline	Model 00	History ○○○○○●○○○○○○○○	Main results	Proof 00	Conclusion
The 60's					



#### (g) Sir John Kingman

(ロ)、(型)、(E)、(E)、 E、のQの

Punchline	Model 00	History oooooo●ooooooo	Main results	Proof 00	Conclusion
The 60'	S				

#### Kingman's Bound for general G/G/1 queue

- $E[L] \leq \frac{1}{2} \left( Var[A\mu_A] + \rho^2 Var[S\mu_S] \right) \times \frac{1}{1-\rho}$
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Simple, explicit, general, scalable
- Useful in theory and practice
- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free
- Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

Punchline	Model oo	History ooooco●ooooooo	Main results	Proof 00	Conclusion
The 60	's				

Kingman's Bound for general G/G/1 queue

- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Simple, explicit, general, scalable
- Useful in theory and practice
- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free
- Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

< □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □

Punchline	Model oo	History ooooco●ooooooo	Main results	Proof 00	Conclusion
The 60	's				

Kingman's Bound for general G/G/1 queue

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$$
 is scale-free

• Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

Punchline	Model 00	History ○○○○○●○○○○○○○	Main results	Proof 00	Conclusion
The 60 <sup>3</sup>	's				

Kingman's Bound for general G/G/1 queue

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• Simple, explicit, general, scalable

Useful in theory and practice

- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free
- Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

Punchline	Model 00	History ○○○○○●○○○○○○○	Main results	Proof 00	Conclusion
The 60	's				

Kingman's Bound for general G/G/1 queue

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$$

- Simple, explicit, general, scalable
- Useful in theory and practice
- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free
- Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

Punchline 0000	Model 00	History ○○○○○●○○○○○○○	Main results	Proof 00	Conclusion
The 60 <sup>3</sup>	s				

Kingman's Bound for general G/G/1 queue

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$$

- Simple, explicit, general, scalable
- Useful in theory and practice
- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free

• Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

・ロト・日本・日本・日本・日本

Punchline	Model oo	History ○○○○○●○○○○○○○	Main results	Proof 00	Conclusion
The 60's	2				

Kingman's Bound for general G/G/1 queue

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + \rho^2 Var[S\mu_S]) \times \frac{1}{1-\rho}$$

• 
$$E[L] \leq \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times \frac{1}{1-\rho}$$

- Simple, explicit, general, scalable
- Useful in theory and practice
- $\frac{1}{2}(Var[A\mu_A] + Var[S\mu_S])$  is scale-free
- Scales as  $\frac{1}{1-\rho}$  as  $\rho \uparrow 1$  in a broad sense

・ロト・日本・日本・日本・日本

Punchline	Model 00	History ooooooo●oooooo	Main results	Proof 00	Conclusion

### The 60's (cont.)

#### Kingman's heavy-traffic analysis for G/G/1 queue

- Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$

◆□ ▶ ◆□ ▶ ◆三 ▶ ◆三 ▶ ● ○ ○ ○ ○

- Certain technical conditions required
- Shows Kingman's bound is tight as  $\rho \uparrow 1$

Punchline	Model oo	History ○○○○○○●○○○○○○	Main results	Proof 00	Conclusion

#### The 60's (cont.)

Kingman's heavy-traffic analysis for G/G/1 queue

- $\bullet\,$  Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2}(Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$

- Certain technical conditions required
- Shows Kingman's bound is tight as  $\rho \uparrow 1$

Punchline	Model 00	History ○○○○○○●○○○○○○	Main results	Proof 00	Conclusion

#### The 60's (cont.)

• Kingman's heavy-traffic analysis for G/G/1 queue

- $\bullet\,$  Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$

・ロト・日本・日本・日本・日本

- Certain technical conditions required
- Shows Kingman's bound is tight as  $\rho \uparrow 1$

Punchline	Model oo	History ○○○○○○●○○○○○○	Main results	Proof 00	Conclusion
The 60's	(cont.)				

• Kingman's heavy-traffic analysis for G/G/1 queue

- $\bullet\,$  Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$

• 
$$\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$$

・ロト・日本・日本・日本・日本

Certain technical conditions required

• Shows Kingman's bound is tight as  $\rho \uparrow 1$ 

Punchline	Model 00	History ⊙⊙⊙⊙⊙⊙●⊙⊙⊙⊙⊙⊙	Main results	Proof 00	Conclusion

#### The 60's (cont.)

Kingman's heavy-traffic analysis for G/G/1 queue

- Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$

- Certain technical conditions required
- Shows Kingman's bound is tight as  $\rho \uparrow 1$

Punchline	Model 00	History ○○○○○○●○○○○○○	Main results	Proof 00	Conclusion

#### The 60's (cont.)

Kingman's heavy-traffic analysis for G/G/1 queue

- Consider a sequence of queues indexed by intensity  $\rho$
- Let  $L_{\rho}$  be the s.s. r.v. for system with intensity  $\rho$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$

- Certain technical conditions required
- Shows Kingman's bound is tight as  $\rho \uparrow 1$

Punchline	Model 00	History ○○○○○○○●○○○○○	Main results	Proof 00	Conclusion
The 60'	's (cont	· )			

### Kingman poses some open problems

- Multi-server analogue of Kingman's bound?
- Multi-server analogue of heavy-traffic analysis?

(日) (日) (日) (日) (日) (日) (日)

Punchline	Model 00	History ○○○○○○○●○○○○○	Main results	Proof 00	Conclusion
The 60	's (cont	.)			

### Kingman poses some open problems

- Multi-server analogue of Kingman's bound?
- Multi-server analogue of heavy-traffic analysis?

(日) (日) (日) (日) (日) (日) (日)

Punchline	Model oo	History ○○○○○○○●○○○○○	Main results	Proof 00	Conclusion
The 60'	s (cont	.)			

- Kingman poses some open problems
  - Multi-server analogue of Kingman's bound?
  - Multi-server analogue of heavy-traffic analysis?

Punchline	Model 00	History ○○○○○○○○●○○○○	Main results	Proof 00	Conclusion
The 70	's				



(h) Borovkov



(i) Whitt



(j) Iglehart

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

Punchline	Model 00	History ooooooooooooooo	Main results	Proof 00	Conclusion
The 70'	S				

Kollerstrom solves the multi-server heavy-traffic analysis

#### • FIXED n, $\rho \uparrow 1$

- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as ρ ↑ 1
- Same  $\frac{1}{1-a}$  scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou

Attempts to use for a multi-server Kingman's bound

Punchline	Model oo	History ○○○○○○○○○●○○○	Main results	Proof 00	Conclusion
The 70 <sup>3</sup>	S				

Kollerstrom solves the multi-server heavy-traffic analysis

#### • FIXED n, $\rho \uparrow 1$

- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as  $ho \uparrow 1$
- Same <sup>1</sup>/<sub>1-a</sub> scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound

Punchline	Model 00	History ○○○○○○○○○●○○○	Main results	Proof 00	Conclusion
The 70's					

Kollerstrom solves the multi-server heavy-traffic analysis

- FIXED n,  $\rho \uparrow 1$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as  $ho \uparrow 1$
- Same  $\frac{1}{1-a}$  scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou

Attempts to use for a multi-server Kingman's bound

Punchline	Model 00	History ooooooooooooooo	Main results	Proof 00	Conclusion
The 70's					

Kollerstrom solves the multi-server heavy-traffic analysis

- FIXED n, *ρ* ↑ 1
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as  $\rho \uparrow 1$
- Same <sup>1</sup>/<sub>1-a</sub> scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound

Punchline 0000	Model 00	<b>History</b> ○○○○○○○○○●○○○	Main results	Proof 00	Conclusion
The 70					

ine /us

Kollerstrom solves the multi-server heavy-traffic analysis

- FIXED n,  $\rho \uparrow 1$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as  $\rho \uparrow 1$
- Same  $\frac{1}{1-\rho}$  scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound

All require FIXED n, p † 1
 No hope of general explicit bound

Punchline	Model 00	History ooooooooooooooo	Main results	Proof 00	Conclusion
The 70's	S				

Kollerstrom solves the multi-server heavy-traffic analysis

- FIXED n,  $\rho \uparrow 1$
- $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
- System behaves like a single sped-up server as  $\rho \uparrow 1$
- Same  $\frac{1}{1-a}$  scaling as single-server case
- Related results by Whitt, Borovkov, Iglehart, Loulou

Attempts to use for a multi-server Kingman's bound

Punchline	Model oo	History oooooooooooooooo	Main results	Proof 00	Conclusion
The 70	's				

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n,  $\rho \uparrow 1$
  - $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - System behaves like a single sped-up server as  $\rho \uparrow 1$
  - Same  $\frac{1}{1-a}$  scaling as single-server case
  - Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound
  - Complicated corrections to the heavy-traffic approximation
  - All require FIXED n,  $\rho \uparrow 1$
  - No hope of general explicit bound

Punchline	Model	<b>History</b> ०००००००००●०००	Main results	Proof 00	Conclusion

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n,  $\rho \uparrow 1$
  - $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - System behaves like a single sped-up server as ρ ↑ 1
  - Same  $\frac{1}{1-a}$  scaling as single-server case
  - Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n,  $\rho \uparrow 1$
  - No hope of general explicit bound

Punchline	Model 00	History ०००००००००●०००	Main results	Proof 00	Conclusion

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n,  $\rho \uparrow 1$
  - $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - System behaves like a single sped-up server as ρ ↑ 1
  - Same  $\frac{1}{1-a}$  scaling as single-server case
  - Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n,  $\rho \uparrow \uparrow$
  - No hope of general explicit bound

Punchline	Model 00	History ०००००००००●०००	Main results	Proof 00	Conclusion

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n,  $\rho \uparrow 1$
  - $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - System behaves like a single sped-up server as  $\rho \uparrow 1$
  - Same  $\frac{1}{1-a}$  scaling as single-server case
  - Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n,  $\rho \uparrow 1$
  - No hope of general explicit bound

Punchline	Model 00	History ०००००००००●०००	Main results	Proof 00	Conclusion

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, *ρ* ↑ 1
  - $\{(1-\rho)L_{\rho}, \rho \uparrow 1\} \Rightarrow \frac{1}{2} (Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - System behaves like a single sped-up server as ρ ↑ 1
  - Same  $\frac{1}{1-a}$  scaling as single-server case
  - Related results by Whitt, Borovkov, Iglehart, Loulou
- Attempts to use for a multi-server Kingman's bound
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, ρ ↑ 1
  - No hope of general explicit bound

Punchline	Model 00	History ○○○○○○○○○●○○	Main results	Proof 00	Conclusion
The 70's	s (cont	.)			

#### Kingman derives a simple bound for G/G/n queues

By analyzing the cyclic routing bound

- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
- The *n* in front of *Var*[*S*µ<sub>*S*</sub>] renders it ineffective!

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

Punchline	Model 00	History ○○○○○○○○○○●○○	Main results	Proof 00	Conclusion
The 70	's (cont	.)			

#### • Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- *E*[*L*] ≤ ½ (*Var*[*Aµ<sub>A</sub>*] + *n* × *Var*[*Sµ<sub>S</sub>*]) × ¼/(1-ρ)
   Later made rigorous by Wolff, Brumelle, Mori
   The *n* in front of *Var*[*Sµ<sub>S</sub>*] renders it ineffective

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

Punchline	Model 00	History ooooooooooooooooo	Main results	Proof 00	Conclusion
<b>TI T</b> 01					

#### The 70's (cont.)

#### Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
   The *n* in front of Var[Suc] renders it ineffective!

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

Punchline	Model oo	History oooooooooooooooo	Main results	Proof 00	Conclusion
The 70's	(cont.)				

### Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
- The n in front of Var[SµS] renders it ineffective!

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

• Especially when  $n \to \infty, \rho \uparrow 1$  together

Punchline	Model	History ooooooooooooooo	Main results	Proof 00	Conclusion

#### The 70's (cont.)

#### Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
- The *n* in front of *Var*[*S*µ<sub>*S*</sub>] renders it ineffective!

- Especially when  $n \to \infty, \rho \uparrow 1$  together
- Can we get rid of it?

Punchline	Model	History ooooooooooooooo	Main results	Proof 00	Conclusion

#### The 70's (cont.)

Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
- The *n* in front of *Var*[*S*µ<sub>*S*</sub>] renders it ineffective!

◆□▶ ◆□▶ ▲□▶ ▲□▶ ■ ののの

• Especially when  $n \to \infty, \rho \uparrow 1$  together

Punchline	Model 00	History ○○○○○○○○○●○○	Main results	Proof 00	Conclusion

#### The 70's (cont.)

Kingman derives a simple bound for G/G/n queues

- By analyzing the cyclic routing bound
- $E[L] \leq \frac{1}{2} (Var[A\mu_A] + n \times Var[S\mu_S]) \times \frac{1}{1-\rho}$
- Later made rigorous by Wolff, Brumelle, Mori
- The n in front of Var[SµS] renders it ineffective!

- Especially when  $n \to \infty, \rho \uparrow 1$  together
- Can we get rid of it?

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

## Halfin-Whitt regime

### • Halfin-Whitt scaling regime:

- Used to study quality-efficiency trade-off in service systems
- Many servers, regular service times, sped-up arrivals
- $ho \sim 1 Bn^{-\frac{1}{2}}$
- $P_{wait}$  has a non-trivial limiting value as  $n o \infty$
- Introduced in 1981 by Halfin and Whitt
- Intensely studied in 90's and 2000's
- Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^{rac{1}{2}}$
- Complicated non-explicit measure-valued weak limits
- Heavy-traffic corrections and bounds
  - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
- No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $ho \sim$  1  $Bn^{-1}$
  - $P_{\mathit{wait}}$  has a non-trivial limiting value as  $n o \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^{rac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramana
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $ho \sim$  1  $Bn^{-rac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n o \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^{rac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, .
     No general, scalable, simple and explicit bound

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n o \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^{rac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan,
     No general scalable simple and explicit bound

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^{rac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
     No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $rac{1}{1-a} \sim n^rac{1}{2}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramanal
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-a} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
  - Heavy-traffic corrections and bounds
     Dai, Brav., Gurvich, Leeuw., Zwart, Ramanar
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-\rho} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-a} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-\rho} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		000000000000000000000000000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-a} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		0000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-\rho} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History	Main results	Proof	Conclusion
		00000000000000000			

- Halfin-Whitt scaling regime:
  - Used to study quality-efficiency trade-off in service systems
  - Many servers, regular service times, sped-up arrivals
  - $\rho \sim 1 Bn^{-\frac{1}{2}}$
  - $P_{wait}$  has a non-trivial limiting value as  $n \to \infty$
  - Introduced in 1981 by Halfin and Whitt
  - Intensely studied in 90's and 2000's
  - Proven that L scales (roughly) as  $\frac{1}{1-\rho} \sim n^{\frac{1}{2}}$
  - Complicated non-explicit measure-valued weak limits
    - HW,GM,PR,Reed,GG,DDG,AR,...
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Punchline	Model	History ○○○○○○○○○○○○	Main results	Proof	Conclusion

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-a}$  bounds
    - No multi-server Kingman's bound
- Daley has lamented / conjectured on this in 70's,80's,90's
  Such a bound may not even exist

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000	00000	00	000

#### Until the present

#### In spite of a century of work on multi-server queues ....

- No universal explicit  $\frac{1}{1-\rho}$  bounds
  - No multi-server Kingman's bound
  - Normalized moments  $\times \frac{1}{1-\rho}$

Daley has lamented / conjectured on this in 70's,80's,90's

Such a bound may not even exist

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000	00000	00	000

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-a}$  bounds
    - No multi-server Kingman's bound
    - Normalized moments  $\times \frac{1}{1-1}$

Daley has lamented / conjectured on this in 70's,80's,90's
Such a bound may not even exist

・ロト ・聞ト ・ヨト ・ヨト 三日

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000	00000	00	000

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-\rho}$  bounds
    - No multi-server Kingman's bound
    - Normalized moments  $\times \frac{1}{1-\rho}$

Daley has lamented / conjectured on this in 70's,80's,90's
Such a bound may not even exist

・ロト ・聞ト ・ヨト ・ヨト 三日

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000	00000	00	000

#### Until the present

In spite of a century of work on multi-server queues ....

- No universal explicit  $\frac{1}{1-\rho}$  bounds
  - No multi-server Kingman's bound
  - Normalized moments  $\times \frac{1}{1-\rho}$

Daley has lamented / conjectured on this in 70's,80's,90's

Such a bound may not even exist

Punchline	Model oo	History ○○○○○○○○○○○○	Main results	Proof 00	Conclusion

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-\rho}$  bounds
    - No multi-server Kingman's bound
    - Normalized moments  $\times \frac{1}{1-\rho}$
- Daley has lamented / conjectured on this in 70's,80's,90's

- Such a bound may not even exist
  - Negative results of Gupta et al.
  - Complexity of HW limits

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000	00000	00	000

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-\rho}$  bounds
    - No multi-server Kingman's bound
    - Normalized moments  $\times \frac{1}{1-\rho}$
- Daley has lamented / conjectured on this in 70's,80's,90's

- Such a bound may not even exist
  - Negative results of Gupta et al.
  - Complexity of HW limits

Punchline	Model oo	History ○○○○○○○○○○○○	Main results	Proof 00	Conclusion

#### Until the present

- In spite of a century of work on multi-server queues ....
  - No universal explicit  $\frac{1}{1-\rho}$  bounds
    - No multi-server Kingman's bound
    - Normalized moments  $\times \frac{1}{1-\rho}$
- Daley has lamented / conjectured on this in 70's,80's,90's

- Such a bound may not even exist
  - Negative results of Gupta et al.
  - Complexity of HW limits

Punchline	Model oo	History 0000000000000	Main results ●○○○○	Proof 00	Conclusion
Outline					





## 3 History



## 5 Proof





Punchline	Model 00	History 0000000000000	Main results o●ooo	Proof 00	Conclusion

## Multi-server Kingman's Bound

#### Corollary

## For any G/G/n queue s.t. $E[A^3], E[S^3] < \infty$ ,

E[L] is at most

$$10^{500} \left( E[(S\mu_S)^3] E[(A\mu_A)^3] \right)^3 \times \frac{1}{1-\rho}$$

Punchline	Model 00	History 0000000000000	Main results ○●○○○	Proof 00	Conclusion

## Multi-server Kingman's Bound

#### Corollary

For any G/G/n queue s.t.  $E[A^3], E[S^3] < \infty$ ,

E[L] is at most

$$10^{500} \left( \boldsymbol{E} \left[ (\boldsymbol{S} \mu_{\boldsymbol{S}})^3 \right] \boldsymbol{E} \left[ (\boldsymbol{A} \mu_{\boldsymbol{A}})^3 \right] \right)^3 \times \frac{1}{1 - \rho}$$

▲□▶▲圖▶▲≣▶▲≣▶ ■ のみの

Punchline	Model	History	Main results	Proof	Conclusion
			00000		

# Theorem For any G/G/n queue s.t. $E[A^3], E[S^3] < \infty$ , $P_{wait}$ is at most $10^{500} \left( E[(S\mu_S)^3] E[(A\mu_A)^3] \right)^3 \left( n(1-\rho)^2 \right)^{-\frac{3}{2}}$ .

(日) (日) (日) (日) (日) (日) (日)

- "Kicks in" exactly in HW regime
  - $(n(1-\rho)^2)^{-2} = B^{-3}$

Punchline	Model	History	Main results	Proof	Conclusion
			00000		

#### Theorem

For any G/G/n queue s.t.  $E[A^3], E[S^3] < \infty$ ,

P<sub>wait</sub> is at most

$$10^{500} \Big( E[(S\mu_S)^3] E[(A\mu_A)^3] \Big)^3 \Big( n(1-
ho)^2 \Big)^{-\frac{5}{2}}$$

"Kicks in" exactly in HW regime
 (n(1 − ρ)<sup>2</sup>)<sup>-3</sup> = B<sup>-3</sup>

Punchline	Model	History	Main results	Proof	Conclusion
			00000		

#### Theorem

For any G/G/n queue s.t.  $E[A^3], E[S^3] < \infty$ ,

P<sub>wait</sub> is at most

$$10^{500} \Big( E[(S\mu_S)^3] E[(A\mu_A)^3] \Big)^3 \Big( n(1-\rho)^2 \Big)^{-\frac{1}{2}}$$

(日) (日) (日) (日) (日) (日) (日)

"Kicks in" exactly in HW regime

Punchline	Model	History	Main results	Proof	Conclusion
			00000		

#### Theorem

For any G/G/n queue s.t.  $E[A^3], E[S^3] < \infty$ ,

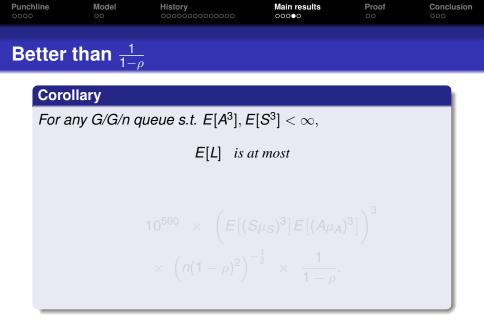
P<sub>wait</sub> is at most

$$10^{500} \Big( E[(S\mu_S)^3] E[(A\mu_A)^3] \Big)^3 \Big( n(1-\rho)^2 \Big)^{-\frac{1}{2}}$$

▲□▶ ▲□▶ ▲三▶ ▲三▶ - 三 - のへで

"Kicks in" exactly in HW regime

• 
$$(n(1-\rho)^2)^{-\frac{3}{2}} = B^{-3}$$



(日) (日) (日) (日) (日) (日) (日)

• Vast generalization of M/M/n . .

Punchline	Model oo	History oooooooooooooo	Main results ○○○●○	Proof 00	Conclusion
Better	than $\frac{1}{1-}$	$\overline{ ho}$			
Coro	llary				
For a	any G/G/n d	queue s.t. E[A³], E	$[S^3] < \infty,$		
		E[L] is a	t most		
		$10^{500} \times (E[(S \times (n(1-\rho)^2)^{-1})])$	$[S\mu_S)^3]E[(A\mu_A)^{-\frac{1}{2}}]$	$)^{3}]\Big)^{3}$	
		$\times (n(1-\rho))$	$\times \frac{1-\rho}{1-\rho}$		

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへぐ

Vast generalization of M/M/n . .
 E[L] = Pwat × (Ep)

Punchline	Model oo	History 00000000000000	Main results ○○○●○	Proof 00	Conclusion
Bett	er than $\frac{1}{1-}$	- <u>_</u>			
С	orollary				
F	or any G/G/n	queue s.t. E[A <sup>3</sup> ], E	$[S^3] < \infty,$		
		E[L] is a	at most		
		$10^{500} \times (E[(3$		,	
		$\times \left(n(1-\rho)^2\right)^2$	$-\frac{1}{2}$ $\times$ $\frac{1}{1-\rho}$ .		

Vast generalization of M/M/n ...
 E[L] = P<sub>wait</sub> × <sup>P</sup>/<sub>1−ρ</sub>

Punchlin 0000	e Model	History ०००००००००००००	Main results ○○○●○	Proof 00	Conclusion	
Bet	ter than $\frac{1}{1}$	<u>Ι</u>				
	Corollary					
ŀ	For any G/G/n	queue s.t. E[A <sup>3</sup> ], E	$\mathbf{S}^{3}]<\infty,$			
	E[L] is at most					
		$10^{500} \times (E[($	$S\mu_S)^3]E[(A\mu_A)^3]E[(A\mu_B)^3]B]$	$\left[\right)^{3}\right]^{3}$		
		$\times \left(n(1-\rho)^2\right)$	$-$ × $\overline{1-\rho}$ .			

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● のへぐ

• Vast generalization of M/M/n ...

• 
$$E[L] = P_{wait} \times \frac{\rho}{1-\rho}$$

Punchline	Model 00	History	Main results ○○○○●	Proof 00	Conclusion
Other I	results i	in paper			

▲□▶ ▲□▶ ▲ 三▶ ▲ 三▶ - 三 - のへぐ

### • More moments $\rightarrow$ better bounds

- Need at least  $2 + \epsilon$
- Explicit tail bounds
- Implications for H-W regime

Punchline	Model 00	History	Main results ○○○○●	Proof 00	Conclusion
Other I	results i	in paper			

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

#### $\bullet \ \ \text{More moments} \to \text{better bounds}$

- Need at least  $\mathbf{2} + \epsilon$
- Explicit tail bounds
- Implications for H-W regime

Punchline	Model oo	History occocococococo	Main results ○○○○●	Proof 00	Conclusion
Other r	esults i	in paper			

▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

#### • More moments $\rightarrow$ better bounds

- Need at least  $\mathbf{2} + \epsilon$
- Explicit tail bounds

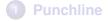
Implications for H-W regime

Punchline	Model oo	History	Main results ○○○○●	Proof 00	Conclusion
Other r	aeulte i	in naner			

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● ● ● ●

- $\bullet \ \ \text{More moments} \to \text{better bounds}$ 
  - Need at least  $\mathbf{2} + \epsilon$
- Explicit tail bounds
- Implications for H-W regime

Punchline	Model 00	History 0000000000000	Main results	Proof ●○	Conclusion
Outline					





### 3 History









Punchline	Model	History	Main results	Proof	Conclusion
				00	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^n N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ....

Punchline	Model	History	Main results	Proof	Conclusion
				00	

- GG.13 bounds L by 1-D random walk •  $\sup_{t>0} (A(t) - \sum_{i=1}^{n} N_i(t))$
- G.16 similarly bounds P<sub>wain</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ....

Punchline	Model	History	Main results	Proof	Conclusion
				$\circ \bullet$	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds . . .

Punchline	Model	History	Main results	Proof	Conclusion
				$\circ \bullet$	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds . .

Punchline	Model	History	Main results	Proof	Conclusion
				$\circ \bullet$	

#### **Proof Overview**

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically

Many novel explicit bounds . . .

Punchline	Model	History	Main results	Proof	Conclusion
				00	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ...
  - Pooled renewal processes
  - Negative drift r.w. with stat. inc.
  - Maximal inequalities

Punchline	Model	History	Main results	Proof	Conclusion
				0•	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ...
  - Pooled renewal processes
  - Negative drift r.w. with stat. inc.
  - Maximal inequalities

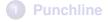
Punchline	Model	History	Main results	Proof	Conclusion
				0•	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ...
  - Pooled renewal processes
  - Negative drift r.w. with stat. inc.
  - Maximal inequalities

Punchline	Model	History	Main results	Proof	Conclusion
				0•	

- GG.13 bounds L by 1-D random walk
  - $\sup_{t\geq 0} \left( A(t) \sum_{i=1}^{n} N_i(t) \right)$
- G.16 similarly bounds P<sub>wait</sub>
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ...
  - Pooled renewal processes
  - Negative drift r.w. with stat. inc.
  - Maximal inequalities

Punchline	Model 00	History 0000000000000	Main results	Proof 00	Conclusion ●○○
Outline					





### 3 History



### 5 Proof



Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000		00	○●○
Summa	arv				

・ロト・日本・日本・日本・日本

- First multi-server analogue of Kingman's bound
- Explicit bounds with universal <sup>1</sup>/<sub>1-0</sub> scaling
- Higher moments and P<sub>wait</sub>
- Applications to HW regime
- Broad theoretical foundation for  $\frac{1}{1-\rho}$

Punchline	Model	History	Main results	Proof	Conclusion
0000	00	0000000000000		00	○●○
Summa	arv				

(日) (日) (日) (日) (日) (日) (日)

- First multi-server analogue of Kingman's bound
- Explicit bounds with universal  $\frac{1}{1-\rho}$  scaling
- Higher moments and P<sub>wait</sub>
- Applications to HW regime
- Broad theoretical foundation for  $\frac{1}{1-\rho}$

Punchline	Model oo	History 0000000000000	Main results	Proof 00	Conclusion ○●○
Summa	ary				

・ロト・日本・日本・日本・日本

- First multi-server analogue of Kingman's bound
- Explicit bounds with universal  $\frac{1}{1-\rho}$  scaling
- Higher moments and P<sub>wait</sub>
- Applications to HW regime
- Broad theoretical foundation for  $\frac{1}{1-\rho}$

Punchline	Model oo	History 0000000000000	Main results	Proof 00	Conclusion ○●○
Summa	ary				

・ロト・日本・日本・日本・日本

- First multi-server analogue of Kingman's bound
- Explicit bounds with universal  $\frac{1}{1-\rho}$  scaling
- Higher moments and P<sub>wait</sub>
- Applications to HW regime
- Broad theoretical foundation for 
   <sup>1</sup>/<sub>1-a</sub>

Punchline	Model oo	History 0000000000000	Main results	Proof 00	Conclusion ○●○
Summa	ary				

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● ● ● ●

- First multi-server analogue of Kingman's bound
- Explicit bounds with universal  $\frac{1}{1-\rho}$  scaling
- Higher moments and P<sub>wait</sub>
- Applications to HW regime
- Broad theoretical foundation for  $\frac{1}{1-\rho}$

Punchline	Model	History	Main results	Proof	Conclusion
					000

#### **Future research**

#### Get that constant down!

- Tighter analysis
- Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model 00	History 0000000000000	Main results	Proof 00	Conclusion ○○●

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?

Punchline	Model	History	Main results	Proof	Conclusion
					000

- Get that constant down!
  - Tighter analysis
  - Fundamentally different analysis
- Bridge to known asymptotic results e.g. in HW
- Bridge to moment results of Scheller-Wolf
- Heavy tails
- Other queueing models
- How to trade off simplicity and accuracy in bounds?
- What do we want from our analyses?