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

David A. Goldberg

Cornell

LNMB

Outline

Outline

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

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

- Central insight of queueing theory:
  - L (s.s num in q) scales as $\frac{1}{1-\rho}$ as $\rho \uparrow 1$
  - Many operational applications

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

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

- Central insight of queueing theory:
  - L (s.s num in q) scales as $\frac{1}{1-\rho}$ as $\rho \uparrow 1$
  - Many operational applications

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

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

- Central insight of queueing theory:
  - L (s.s num in q) scales as $\frac{1}{1-\rho}$ as $\rho \uparrow 1$
  - Many operational applications

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

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

- 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 ...
- The exception is Kingman's bound, but ...
- A major difficulty is that any such bound ...
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 . . .
- The exception is Kingman's bound, but . . .
- A major difficulty is that any such bound . . .
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 ...
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 . . .
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 ...
  - Only for single server
- A major difficulty is that any such bound ...
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 ...
  - Only for single server
- A major difficulty is that any such bound ...
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 . . .
  - 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 $\rho \to 1$
- Multi-server Kingman's bound open for 50 years!

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

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

- 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 ...
  - 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!

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

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

- 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 ...
  - 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!

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

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

- 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 ...
  - 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!

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

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

- 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 ...
  - 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!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

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

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

- 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 $\frac{1}{1-\rho}$ heuristic for multi-server queues
- Proof of concept , work to do!

Outline

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

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_{wait}: s.s. prob. all servers busy

FCFS GI/GI/n Queue

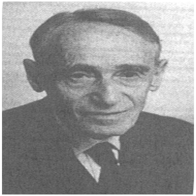
FCFS GI/GI/n Queue

- 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_{wait}: s.s. prob. all servers busy

Outline

Early 20th Century

(a) Erlang



(b) Pollaczek

(c) Khinchin

Early 20th Century

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

Early 20th Century

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

Early 20th Century

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

Early 20th Century

Early 20th Century

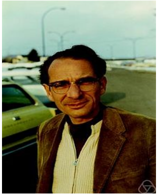
- Model “invented” to study early telephone networks
- Pioneering work by engineers such as Erlang
- Soon found many other applications
- Erlang solves the M/M/n case
- P-K formula for E[L] in M/G/1 case

Early 20th Century

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

Mid 20th Century



(d) Spitzer

(e) Lindley

(f) Kendall

Mid 20th Century

Mid 20th Century

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

Mid 20th Century

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

Mid 20th Century

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

Mid 20th Century

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

Mid 20th Century

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

Mid 20th Century

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

The 60's

(g) Sir John Kingman

The 60's

The 60's

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

The 60's

The 60's

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

The 60's

The 60's

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

The 60's

The 60's

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

The 60's

The 60's

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

The 60's

The 60's

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

The 60's

The 60's

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

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}(Var[A\mu_A] + Var[S\mu_S]) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

- Kingman's heavy-traffic analysis for G/G/1 queue
  - Consider a sequence of queues indexed by intensity ρ
  - Let L_ρ be the s.s. r.v. for system with intensity ρ
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - Certain technical conditions required
  - Shows Kingman's bound is tight as $\rho \uparrow 1$

The 60's (cont.)

The 60's (cont.)

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

The 60's (cont.)

The 60's (cont.)

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

The 60’s (cont.)

The 60’s (cont.)

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

The 70's

(h) Borovkov

(i) Whitt

(j) Iglehart

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - 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

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\left(Var[A\mu_A] + Var[S\mu_S]\right) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's

The 70's

- Kollerstrom solves the multi-server heavy-traffic analysis
  - FIXED n, $\rho \uparrow 1$
  - $\{(1-\rho)L_\rho, \rho \uparrow 1\} \Rightarrow \frac{1}{2}\big(Var[A\mu_A] + Var[S\mu_S]\big) \times Expo(1)$
  - 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
  - Complicated corrections to the heavy-traffic approximation
  - Kollerstrom, Nagaev, Kennedy
  - All require FIXED n, $\rho \uparrow 1$
  - No hope of general explicit bound

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
    - Especially when $n \to \infty, \rho \uparrow 1$ together
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
    - Especially when $n \to \infty$, $\rho \uparrow 1$ together
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
    - Especially when $n \to \infty, \rho \uparrow 1$ together
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
    - Especially when $n \to \infty, \rho \uparrow 1$ together
  - Can we get rid of it?

The 70's (cont.)

The 70's (cont.)

- Kingman derives a simple bound for G/G/n queues
  - By analyzing the cyclic routing bound
  - $E[L] \leq \frac{1}{2}\left(Var[A\mu_A] + n \times Var[S\mu_S]\right) \times \frac{1}{1-\rho}$
  - Later made rigorous by Wolff, Brumelle, Mori
  - The n in front of $Var[S\mu_S]$ renders it ineffective!
    - Especially when $n \to \infty, \rho \uparrow 1$ together
  - Can we get rid of it?

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

- Heavy-traffic corrections and bounds
  - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
- No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, . . .
  - No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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
  - Heavy-traffic corrections and bounds
    - Dai, Brav., Gurvich, Leeuw., Zwart, Ramanan, ...
  - No general, scalable, simple and explicit bounds

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Digression: When $n \to \infty, \rho \uparrow 1$ together

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
  - $\rho \sim 1 - Bn^{-\frac{1}{2}}$
  - 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

Until the present

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

Until the present

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

Until the present

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

Until the present

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

Until the present

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

Until the present

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

Until the present

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

Until the present

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

Outline

Multi-server Kingman's Bound

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

Multi-server Kingman's Bound

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

Universal and explicit P_{wait} bounds

Theorem

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

$$P_{wait} \quad \textit{is at most}$$

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

- "Kicks in" exactly in HW regime

Universal and explicit P_{wait} bounds

Theorem

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

$$P_{wait} \quad \textit{is at most}$$

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

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

Universal and explicit P_{wait} bounds

Theorem

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

$$P_{wait} \quad \textit{is at most}$$

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

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

Universal and explicit P_{wait} bounds

Theorem

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

$$P_{wait} \quad \textit{is at most}$$

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

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

Better than $\frac{1}{1-\rho}$

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

$$\times \left(n(1-\rho)^2 \right)^{-\frac{1}{2}} \times \frac{1}{1-\rho}.$$

- Vast generalization of M/M/n . . .

Better than $\frac{1}{1-\rho}$

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

- Vast generalization of M/M/n . . .
  - $E[L] = P_{wait} \times \frac{\rho}{1-\rho}$

Better than $\frac{1}{1-\rho}$

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

$$\times \left(n(1-\rho)^2 \right)^{-\frac{1}{2}} \times \frac{1}{1-\rho}.$$

- Vast generalization of M/M/n . . .
- $E[L] = P_{wait} \times \frac{\rho}{1-\rho}$

Better than $\frac{1}{1-\rho}$

Corollary

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

$$E[L] \quad \textit{is at most}$$

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

- Vast generalization of M/M/n . . .
  - $E[L] = P_{wait} \times \frac{\rho}{1-\rho}$

Other results in paper

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

Other results in paper

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

Other results in paper

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

Other results in paper

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

Outline

Proof Overview

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_{wait}
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds ...

Proof Overview

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_{wait}
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds . . .

Proof Overview

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_{wait}
- Previously analyzed asymptotically in HW regime
- We analyze universally and non-asymptotically
- Many novel explicit bounds . . .

Proof Overview

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_{wait}
- 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

Proof Overview

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_{wait}
- 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

Proof Overview

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_{wait}
- 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

Proof Overview

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_{wait}
- 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

Proof Overview

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_{wait}
- 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

Proof Overview

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_{wait}
- 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

Outline

Summary

Summary

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

Summary

Summary

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

Summary

Summary

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

Summary

Summary

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

Summary

Summary

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

Future research

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?

Future research

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?

Future research

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?

Future research

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?

Future research

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?

Future research

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?

Future research

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?

Future research

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?

Future research

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?