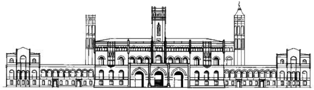
# Telescoping-Tightness Single-node Performance Bounds

Amr Rizk

Joint work with F. Ciucu and S. Mehri

Published as: *On Ultra-Sharp Queueing Bounds*: In Proc. of IEEE INFOCOM 2024



Institute of Communications Technology
Leibniz University Hannover

# Weakness of current Martingale bounds



- Martingale bounds usually consider the maximal value of a random walk, i.e.

$$W = \max_{n \geq 0} \{X_1 + \cdots + X_n\}$$

with increments

$$X_n = S_n - T_n$$

satisfying $\mathsf{E}[X_n] < 0$ for stability ($(S)_n$ being service times and $(T)_n$ being inter-arrival times)

- $W$ is the waiting time of an arbitrary job in steady state
- The technique used to construct tail bounds is based on the following observation

$$\mathsf{P}[W > \sigma] = \mathsf{P}[\mathsf{T} < \infty]$$

with the stopping time

$$\mathsf{T} := \min \{n : X_1 + \cdots + X_n > \sigma\}$$

# Weakness of current Martingale bounds



- Note that $\mathsf{P}[\mathsf{T} = \infty] > 0$ since $\mathsf{E}[X_n] < 0$
- Now, to bound $\mathsf{P}[W > \sigma]$ we can construct the Martingale

$$M_n = e^{\theta(X_1 + \cdots + X_n)}$$

  where $\theta > 0$ satisfies $\phi(\theta) := \mathsf{E}[e^{\theta X}] = 1$ for the Moment-generating function (MGF) $\phi(\theta)$ of the increment $X$.
- Important properties of the stochastic process $M_n$:
  - $\mathsf{E}[M_{n+1} - M_n | \mathcal{F}_n] = 0, \ \forall n$ with $\mathcal{F}_n = \sigma(M_1, \ldots, M_n)$
  - $\mathsf{E}[M_0] = \mathsf{E}[M_\mathsf{T}]$ if T is a finite stopping time; note that T is random (OST)

# Weakness of current Martingale bounds

- Now, we can construct the tail bound on $W$ for a GI/G/1 system as

$$
\begin{aligned}
1 &= \mathsf{E}[M_0] \\
&= \mathsf{E}[M_{\mathsf{T}} 1_{\mathsf{T}<\infty}] \\
&= \mathsf{E}[e^{\theta(X_1+\cdots+X_{\mathsf{T}})} 1_{\mathsf{T}<\infty}] \\
&\geq \mathsf{E}[e^{\theta\sigma} 1_{\mathsf{T}<\infty}] \\
&= e^{\theta\sigma} \mathsf{P}[\mathsf{T} < \infty] \\
&= e^{\theta\sigma} \mathsf{P}[W > \sigma]
\end{aligned}
$$

- We now get the Kingman bound $\mathsf{P}[W > \sigma] < e^{-\theta\sigma}$ with the specified $\theta$ that satisfies $\phi(\theta) = 1$

# Weakness of current Martingale bounds

- Now, we can construct the tail bound on $W$ for a GI/G/1 system as

$$\begin{aligned}1 &= \mathsf{E}[M_0] \\ &= \mathsf{E}[M_{\mathsf{T}} 1_{\mathsf{T}<\infty}] \\ &= \mathsf{E}[e^{\theta(X_1+\cdots+X_{\mathsf{T}})} 1_{\mathsf{T}<\infty}] \\ &\geq \mathsf{E}[e^{\theta\sigma} 1_{\mathsf{T}<\infty}] \\ &= e^{\theta\sigma} \mathsf{P}[\mathsf{T} < \infty] \\ &= e^{\theta\sigma} \mathsf{P}[W > \sigma]\end{aligned}$$

- We now get the Kingman bound $\mathsf{P}[W > \sigma] < e^{-\theta\sigma}$ with the specified $\theta$ that satisfies $\phi(\theta) = 1$
  - where does the error come from? (Observation from numerical evaluations: This error becomes smaller at high utilization)

# Weakness of current Martingale bounds - refined result

- We can construct a refined tail bound on $W$ as

$$\mathsf{E}[e^{\theta(X_1+\cdots+X_{\mathsf{T}})}1_{\mathsf{T}<\infty}] \geq \inf_{x\geq 0} K(x)e^{\theta\sigma}\mathsf{P}[\mathsf{T}<\infty]$$

  with $K(x) := \mathsf{E}[e^{\theta(X_1-x)}|X_1 \geq x]$

- We now get the Ross bound $\mathsf{P}[W>\sigma] < \frac{1}{\inf_{x\geq 0} K(x)} e^{-\theta\sigma}$ which is sharper than the previous bound as $K(x) \geq 1 \; \forall x \geq 0$
- the “error” in the construction persists

# Coarse treatment of the overshoot



- The weakness of the Martingale bound (Kingman version shown in the following) lies in the coarse treatment of the dependency here

$$\mathsf{E}[e^{\theta(X_1+\cdots+X_\mathsf{T})}1_{\mathsf{T}<\infty}] \geq \mathsf{E}[e^{\theta\sigma}1_{\mathsf{T}<\infty}]$$

  - At the stopping time T the random walk *overshoots* $\sigma$, i.e., $X_1+\cdots+X_\mathsf{T} > \sigma$
  - Make use of the overshoot, i.e., *by how much* does the random walk exceed $\sigma$

# Coarse treatment of the overshoot



- The weakness of the Martingale bound (Kingman version shown in the following) lies in the coarse treatment of the dependency here

$$\mathsf{E}[e^{\theta(X_1+\cdots+X_\mathsf{T})}1_{\mathsf{T}<\infty}] \geq \mathsf{E}[e^{\theta\sigma}1_{\mathsf{T}<\infty}]$$

  - At the stopping time T the random walk *overshoots* $\sigma$, i.e., $X_1+\cdots+X_\mathsf{T} > \sigma$
  - Make use of the overshoot, i.e., *by how much* does the random walk exceed $\sigma$
- We aim for a new approach to obtain

$$\mathsf{P}[W > \sigma] = \mathsf{P}[\mathsf{T} < \infty] \leq e^{-\theta\sigma} f(\sigma)$$

# A new hope and a short detour



- Wald's Fundamental Identity: For a stopping time T and a non-negative random variable $Y$ (under some mild technical conditions) the following holds

$$\mathsf{E}_\theta[Y 1_{\mathsf{T}<\infty}] = \mathsf{E}[Y e^{\theta(X_1+\cdots+X_\mathsf{T})}\phi(\theta)^{-\mathsf{T}} 1_{\mathsf{T}<\infty}]$$

  where $\mathsf{E}_\theta[\cdot]$ is the expectation corresponding to $\mathsf{P}_\theta$ which is a probability measure defined from

$$\mathsf{P}_{n,\theta}(A) = \mathsf{E}\left[\frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} 1_A\right] = \int_A \frac{e^{\theta(X_1+\cdots+X_\mathsf{T})}}{\phi(\theta)^n} d\mathsf{P}_n$$

  defined for every $n \in \mathbf{N}$ and to $A \in \mathcal{F}_n$ (change-of-measure).

- Note that if $\mathsf{T} = n$ and $Y$ is measurable $\mathcal{F}_n$

$$\mathsf{E}_\theta[Y] = \mathsf{E}[Y e^{\theta(X_1+\cdots+X_n)}\phi(\theta)^{-n}]$$

  and

$$\mathsf{E}[Y] = \mathsf{E}_\theta[Y e^{-\theta(X_1+\cdots+X_n)}\phi(\theta)^{n}]$$

# Breaking the dependency



- Now we can consider the dependency inside $\mathsf{E}[e^{\theta(X_1+\cdots+X_\mathsf{T})}1_{\mathsf{T}<\infty}]$ by utilizing WFI with $Y = 1$

$$\mathsf{E}[1_{\mathsf{T}<\infty}] = \mathsf{E}_\theta[e^{-\theta(X_1+\cdots+X_\mathsf{T})}1_{\mathsf{T}<\infty}]$$

- The **key** is to observe that $T < \infty$ a.s. on the probability measure $\mathsf{P}_\theta$
  - Using the convexity of $\phi$ and $\phi(0) = \phi(\theta) = 1$ we find that $\mathsf{E}_\theta[X] > 0$
    - through $\mathsf{E}_\theta[X] = \mathsf{E}[Xe^{\theta X}] = \phi'(\theta) > 0$
  - The change of measure **reverses the sign of the drift** $\mathsf{E}_\theta[X]$ of the underlying random walk
    - thus $\mathsf{P}_\theta[T < \infty] = 1$
  - Now we can write $\mathsf{E}_\theta[e^{-\theta(X_1+\cdots+X_\mathsf{T})}1_{\mathsf{T}<\infty}] = \mathsf{E}_\theta[e^{-\theta(X_1+\cdots+X_\mathsf{T})}]$

# Exact expression of $P[W > \sigma]$



- The previous key observation allows us to express

$$\mathsf{P}[W > \sigma] = \mathsf{E}_\theta[e^{-\theta(X_1+\cdots+X_\mathsf{T})}] = e^{-\theta\sigma}\mathsf{E}_\theta[e^{-\theta R_\sigma}]$$

  using the definition of the overshoot

$$R_\sigma = X_1 + \cdots + X_\mathsf{T} - \sigma$$

- Given finite $\sigma$, we express the overshoots's tail in terms of a union of disjoint events as

$$\{R_\sigma > x\} = \bigcup_{n\geq 1}\left\{\sum_{i=1}^{n} X_i > \sigma + x, \max_{1\leq k\leq n-1}\sum_{i=1}^{k} X_i \leq \sigma\right\}$$

  for $x > 0$.

- Now we only need to compute $\mathsf{E}_\theta[e^{-\theta R_\sigma}]$ !

# Exact expression of $P[W > \sigma]$



Theorem

*The waiting time distribution satisfies*

$$\mathsf{P}[W > \sigma] = e^{-\theta\sigma}\left(1 - \sum_{n=1}^{\infty} g_n(\sigma)\right)$$

*for all* $\sigma > 0$, *where* $g_n(\sigma) \geq 0$ *are*

$$\begin{aligned} g_n(\sigma) &:= \mathsf{E}\left[\left(e^{\theta\sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{\mathsf{T}=n}\right] \;\forall n \geq 1 \\ &= \mathsf{E}_\theta\left[\left(1 - e^{\theta(\sigma - \sum_{i=1}^{n} X_i)}\right) 1_{\mathsf{T}=n}\right] \;\forall n \geq 1 \end{aligned}$$

- Upper bounds on $\mathsf{P}[W > \sigma]$ follow by taking any number of terms $g_n(\sigma)$.

# Proof sketch

Fix $\sigma \geq 0$

$$
\begin{aligned}
\mathsf{E}_\theta[e^{-\theta R_\sigma}] &= \int_0^1 \mathsf{P}_\theta[e^{-\theta R_\sigma} > y] dy \\
&= 1 - \int_0^1 \mathsf{P}_\theta[R_\sigma > z] \theta e^{-\theta z} dz \\
&= 1 - \mathsf{P}_\theta[R_\sigma > Z] \\
&= 1 - \sum_{n=1}^{\infty} g_n(\sigma)
\end{aligned}
$$

- The first step follows by rearrangement and the second follows from observing that $Z$ is an exponential random variable with parameter $\theta$.
- The last step follows from the elementary expansion of the overshoot tail in terms of a union of disjoint events

# How do we *obtain* $g_n(\sigma)$? Per expansion of $R_\sigma$



$$g_n(\sigma) = \mathsf{P}_\theta\left(\underbrace{\sum_{i=1}^{n} X_i}_{:=U} > \sigma + Z, \underbrace{\max_{1\leq k\leq n-1}\sum_{i=1}^{k} X_i}_{:=V} \leq \sigma\right)$$

$$= \mathsf{E}\left[e^{\theta U} 1_{U>\sigma+Z, V\leq\sigma}\right] \quad \text{(rewriting } \mathsf{E}_\theta \text{ in terms of } \mathsf{E}\text{)}$$

$$= \mathsf{E}\left[1_{V\leq\sigma}\mathsf{E}\left[e^{\theta U} 1_{U>\sigma+Z} | \mathcal{F}_n\right]\right] \quad (1_{V\leq\sigma} \text{ is measurable wrt. } \mathcal{F}_n)$$

- Term manipulation and similar arguments finds the conditional expectation

$$\mathsf{E}\left[e^{\theta U} 1_{U>\sigma+Z} | \mathcal{F}_n\right] = 1_{U>\sigma}\left(e^{\theta U} - e^{\theta\sigma}\right)$$

leading to

$$g_n(\sigma) = \mathsf{E}\left[\left(e^{\theta\sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{\{U>\sigma, V<\sigma\}}\right] = \mathsf{E}\left[\left(e^{\theta\sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{\mathsf{T}=n}\right]$$

# How do we *compute* $g_n(\sigma)$?

Given
$$g_n(\sigma) = \mathsf{E}\left[\left(e^{\theta \sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{\mathsf{T}=n}\right]$$

- Observe the repetitive structure when ascending in $n$: First, one can compute
$$g_1(\sigma) = \mathsf{E}\left[\left(e^{\theta X_1} - e^{\theta\sigma}\right) 1_{X_1>\sigma}\right]$$

- Now one can write $g_n(\sigma)$ recursively by conditioning in terms of $g_{n-1}$ as
$$\begin{aligned} g_n(\sigma) &= \mathsf{E}\left[e^{X_1} 1_{X_1\le\sigma} \mathsf{E}[(e^{\theta(X_2+\cdots+X_n)} - e^{\theta(\sigma-X_1)})1_A|X_1]\right] \\ &= \mathsf{E}\left[e^{X_1} 1_{X_1\le\sigma} g_{n-1}(\sigma - X_1)\right] \end{aligned}$$

- The key to this recursion is computing the conditional expectation on $X_1$ utilizing the event $A$
$$\{\sum_{i=2}^{n} X_i > \sigma - X_1, \max_{2\le k\le n-1} \sum_{i=2}^{k} X_i \le \sigma - X_1\}$$

# Numerical Example: M/D/1 System

Consider an M/D/1 queue with $T_n \sim \exp(\lambda)$ and deterministic service time $S$.

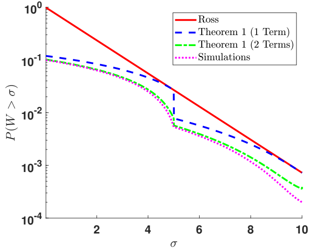
For $\sigma < S$, we compute $g_1(\sigma), g_2(\sigma)$ to obtain the following bounds

- Using 1 term:

$$\mathrm{P}[W > \sigma] \leq 1 - \frac{\theta}{\lambda + \theta} e^{-\lambda(S-\sigma)}$$

- Using 2 terms:

$$\mathrm{P}[W > \sigma] \leq 1 - (1 + \theta S e^{-\theta S} - e^{-2\theta S}) e^{-\lambda(2S-\sigma)}$$




- We observe that the gradual improvements appear to decay exponentially
- We conjecture that the dominant $g_n$ terms are the first ones due to the positive drift

# Conclusion



- We reformulate single node queueing models, first, for GI/G/1 then for AR/G/1 and Markov fluid queues (in the paper)
- **Telescoping-Tightness** of the computed bounds through a cutoff $\sum_{n=1}^{K} g_n(\sigma)$
- Find suitable change-of-measure to **reverse the sign** of the expected increment $\mathsf{E}[X]$ (slightly different construction for GI/$\cdot$/1 and Markov-modulated ones)
- Some models (e.g. M/D/1) have closed form solutions (rare & num. unstable)
- Numerical results show
  - that the first few $g_n(\sigma)$ terms are sufficient
  - non-monotonic behavior in $\sigma$ (construction of $g_n(\sigma)$ for given $\sigma$)
- **Open question:** Given any specific queueing model, is there a $K$ such that one can analytically bound $\sum_{n>K} g_n(\sigma)$?