

What Ten Per Cent More Traffic Costs Depends on Where You Already Are

— Raising the utilisation from **0.50** to **0.55** multiplies the wait by **1.2222**; from **0.90** to **0.99** it multiplies it by **11.0000** — both are the same ten per cent — and at **0.99** the wait is **99.0000** minutes, ninety-nine times that at **0.50** — [Paper 469]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— *From thinking to syncing.* —

Abstract

Congestion is thought to worsen in proportion to the load. This paper shows that the same ten per cent costs **9** times more depending on where one starts—and the ratio has a closed form. No new theorem or law is claimed. First, at a utilisation of **0.50** the wait is exactly **1.000000000000** service times (Section 2). Second, at **0.90** it is exactly **9.000000000000** times that, and at **0.99** it is **99.0000** minutes (Section 2). Third, this is the core of the paper. The same ten per cent gives **1.2222** from **0.50** and **11.0000** from **0.90** (Section 3). Fourth, the ratio matches the closed form $a(1 - \rho)/(1 - a\rho)$ exactly (Section 3). Fifth, cutting the slack to a thousandth multiplies the wait by **1001.000000** (Section 3). Sixth, the separator is whether the effect follows the load or the reciprocal of the slack (Section 4). **Keywords:** queueing theory, M/M/1 queue, Little's law, utilisation, waiting time.

Scope of this paper (scope note): No new theorem or law is claimed—the M/M/1 queue and Little's law are entirely standard. No system is designed—number of servers, priorities, appointments and pre-emption are not the subject. Only M/M/1 is treated—Poisson arrivals, exponential service and a single server; real systems often violate these. Variability is excluded—only means are treated; variance and quantiles are not the subject. Human behaviour is excluded—balking, reneging and returning later are not the subject. Relation to earlier papers: Paper 308 showed that the gap between the buses one boards is twice the average, and its scope note stated that it would not build queueing theory—this paper takes that reserved seat. Paper 300 showed that common root and distinct root can be decided —“the load” and “the slack” are distinct roots by that criterion. Paper 468 showed that the same proportional increase buys less at the top—there too, the starting point set the effect; here the direction is reversed. What is added is tabulating the wait at five utilisations, giving the effect of the same ten per cent as **1.2222** and **11.0000**, matching the ratio to the closed form exactly, showing that a thousandth of the slack costs **1001.000000**, and placing the separator at load versus slack.

1. Notation and the Question

Setting: For a single server with arrival rate λ and service rate μ , let the utilisation be $\rho = \lambda/\mu$. For M/M/1 the mean number in system and the mean wait are

$$L = \frac{\rho}{1 - \rho}, \quad W_q = \frac{\rho}{\mu(1 - \rho)}$$

Below, $\mu = 1$: one item per minute.

Observation 1.1

Congestion is thought to worsen in proportion to the load. Ten per cent more arrivals, ten per cent more waiting—a linear picture.

Observation 1.2

But the denominator carries $1 - \rho$. What acts is not the load but the slack that remains—and as the slack goes to zero its reciprocal takes off.

Question: Where does the same ten per cent cost what.

2. At 0.99 the Wait Is 99 Minutes

Computation 2.1: Take the number in system and the wait at five utilisations.

Utilisation	Mean number	Mean wait [min]
0.50	1.0000	1.0000
0.80	4.0000	4.0000
0.90	9.0000	9.0000
0.95	19.0000	19.0000
0.99	99.0000	99.0000

Observation 2.2

At 0.50 the wait is exactly **1.000000000000** minutes. One service time—a check that the expression is set up correctly.

Observation 2.3

Going from 0.50 to 0.90 multiplies the wait by exactly **9.000000000000**. While the load rises only by 1.8—1.8 times the input, nine times the wait.

Observation 2.4

At 0.99 it is **99.0000** minutes. 1.98 times the load, ninety-nine times the wait—far from proportional.

Observation 2.5

At a utilisation of 0 the wait is exactly **0.0000000000000000**. Nobody there, nobody waiting—the expression behaves at both ends.

3. The Same Ten Per Cent Gives 1.2222 or 11.0000

Computation 3.1: Multiply the utilisation by 1.1 from two starting points.

Operation	Wait	Factor
0.50 → 0.55	1.0000 → 1.2222 min	1.2222
0.90 → 0.99	9.0000 → 99.0000 min	11.0000

Observation 3.2

The same ten per cent differs ninefold in effect. 1.2222 against 11.0000—“ten per cent more” has no meaning until one says where.

Observation 3.3

The ratio matches the closed form $a(1 - \rho)/(1 - a\rho)$ exactly. With $a = 1.1$: 1.2222 at $\rho = 0.50$ and 11.0000 at $\rho = 0.90$ —computation and formula land on the same value.

Observation 3.4

What acts is $1 - \rho$, the slack. 0.50 of slack at $\rho = 0.50$, 0.10 at $\rho = 0.90$ —the same ten per cent either takes a tenth of the slack or nine tenths of it.

Computation 3.5: Take the wait as the slack shrinks to 10^{-3} , 10^{-6} and 10^{-9} .

Slack $1 - \rho$	Wait [min]
10^{-3}	9.990000×10^2
10^{-6}	9.999990×10^5
10^{-9}	1.000000×10^9

Observation 3.6

A thousandth of the slack costs **1001.000000** times the wait. Nearly inverse—the product of wait and slack converges to 1 (1.000000027 at 10^{-9}).

Observation 3.7

So “we can squeeze in a little more” is the most expensive judgement. It cuts where little slack is left—the same cut costing almost nothing while there is room.

4. The Separator

The separator: does the effect follow the load, or the reciprocal of the slack.

Observation 4.1

Following the load, adding anywhere costs the same. Ten per cent in, ten per cent worse—easy to forecast.

Observation 4.2

Following the reciprocal of the slack, it takes off as the room runs out. Because $1 - \rho$ sits in the denominator—the same operation differing ninefold with the starting point.

Observation 4.3

These are distinct roots by the criterion of Paper 300. Both are called “how busy it is”—one a quantity of load, one of remaining slack.

Observation 4.4

This stands where Paper 468 stands, with the direction reversed. There, the same proportional operation bought 1.48 times less at the top—here it costs nine times more. In both the starting point decides, but one shrinks and one grows.

Observation 4.5

It fills the seat reserved by Paper 308. There, the gap between the buses one boards was shown to be twice the average, with the scope note stating that queueing theory would not be built—this paper takes that seat.

Observation 4.6

So state utilisation targets as slack. Not “ninety per cent utilised” but “ten per cent of slack”—the same statement, but now going to ninety-five is visibly halving the slack.

Observation 4.7

To be honest—only $M/M/1$ is treated. Poisson arrivals and exponential service are often violated in practice—the 99 minutes depends on the model; that the effect follows the reciprocal of the slack holds across a wide family of queues.

5. Roll Call and Closing

1. **1.000000000000 min at 0.50** (Observation 2.2): **a check on the setup.**
2. **Exactly 9.000000000000 at 0.90** (Observation 2.3): **1.8 times the load.**
3. **99.0000 min at 0.99** (Computation 2.1): **far from proportional.**
4. **0.50 $\rightarrow$ 0.55 gives 1.2222** (Computation 3.1): **the core.**
5. **0.90 $\rightarrow$ 0.99 gives 11.0000** (Computation 3.1): **the same ten per cent.**
6. **Exactly the closed form** (Observation 3.3): **$a(1 - \rho)/(1 - a\rho)$.**
7. **What acts is the slack** (Observation 3.4): **$1 - \rho$ in the denominator.**
8. **A thousandth costs 1001.000000** (Observation 3.6): **nearly inverse.**
9. **The separator is load or slack** (Observation 4.3): **distinct roots by Paper 300.**
10. **Paper 308's reserved seat** (Observation 4.5): **it excluded queueing theory.**

Closing: Congestion is thought to worsen in proportion to the load—ten per cent more arrivals, ten per cent more waiting. But the denominator carries $1 - \rho$: what acts is not the load but the slack that remains, and as the slack goes to zero its reciprocal takes off. The wait itself is already far from proportional—at 0.50 it is exactly **1.000000000000** minutes, one service time. Going to 0.90 multiplies it by exactly **9.000000000000** while the load rises only by 1.8. At 0.99 it is **99.0000** minutes: 1.98 times the load, ninety-nine times the wait. And at a utilisation of 0 the wait is exactly zero, so the expression behaves at both ends. The same ten per cent differs ninefold with the starting point—0.50 $\rightarrow$ 0.55 takes the wait from 1.0000 to 1.2222 minutes, a factor of **1.2222**; 0.90 $\rightarrow$ 0.99 takes it from 9.0000 to 99.0000, a factor of **11.0000**. The ratio matches the closed form $a(1 - \rho)/(1 - a\rho)$ exactly. What acts is $1 - \rho$: with 0.50 of slack the same ten per cent takes a tenth of it, with 0.10 of slack it takes nine tenths. Shrinking the slack to 10^{-3} , 10^{-6} and 10^{-9} gives waits of 9.990000×10^2 , 9.999990×10^5 and 1.000000×10^9 minutes: a thousandth of the slack costs **1001.000000** times, and the product of wait and slack converges to 1. So “we can squeeze in a little more” is the most expensive judgement, because it cuts where little slack is left—the same cut costing almost nothing while there is room. One thing separates them—does the effect follow the load, or the reciprocal of the slack. Following the load, adding anywhere costs the same and forecasts are easy. Following the reciprocal of the slack, it takes off as the room runs out, and the same operation differs ninefold with the starting

point. They are distinct roots by the criterion of Paper 300, one a quantity of load and one of remaining slack. This stands where Paper 468 stands with the direction reversed: there the same proportional operation bought 1.48 times less at the top, here it costs nine times more. And it fills the seat reserved by Paper 308, which showed the gap between the buses one boards to be twice the average while stating that it would not build queueing theory. So state utilisation targets as slack—not “ninety per cent utilised” but “ten per cent of slack”, which makes going to ninety-five visibly a halving. To be honest—only $M/M/1$ is treated, and Poisson arrivals with exponential service are often violated in practice. The 99 minutes depends on the model; that the effect follows the reciprocal of the slack holds across a wide family of queues.

References

1. Erlang, A. K. (1909). Sandsynlighedsregning og Telefonsamtaler. *Nyt Tidsskrift for Matematik B*.
2. Little, J. D. C. (1961). A Proof for the Queueing Formula $L = \lambda W$. *Operations Research*.
3. Kleinrock, L. (1975). *Queueing Systems, Volume 1: Theory*. Wiley.
4. Yamagishi, Y. (2026). *The Three Hundredth Census: Whether Two Things Share a Root Is Decidable* [Paper 300]. Zenodo. 10.5281/zenodo.22171914
5. Yamagishi, Y. (2026). *The Gap between the Buses You Board Is Twice the Average* [Paper 308]. Zenodo. 10.5281/zenodo.22172305
6. Yamagishi, Y. (2026). *The Same Proportional Increase Buys Less at the Top* [Paper 468]. Zenodo. 10.5281/zenodo.22733384

— *From thinking to syncing.* —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)