Token budgets for fleets of AI agents

An old fluid queue in a new setting

Andres Ferragut, Universidad ORT Uruguay

joint work with Diego Goldsztajn and Fernando Paganini

TU/Eindhoven – November 2026

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

Agents run on tokens

- AI agents (coding, research, customer support) work by consuming tokens: every step of a task is paid in tokens.
- Providers sell usage as a quota:
  - it refills at a contracted rate;
  - it is capped at a maximum depth;
  - when it runs out, the user is throttled.
- An organisation runs a fleet of agents against one shared quota. Agents start and finish tasks at random times.

Two purchasing questions

What refill rate should we buy?

How large a quota should we hold?

Two purchasing questions

What refill rate should we buy? How large a quota should we hold?

- Buying too little: agents are throttled and tasks stall.
- Buying too much: the quota sits full and the refill is wasted.

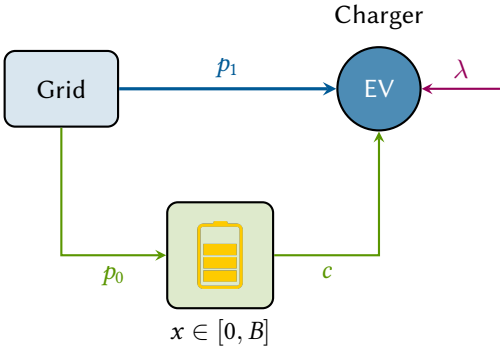
Two purchasing questions

What refill rate should we buy? How large a quota should we hold?

- Buying too little: agents are throttled and tasks stall.
- Buying too much: the quota sits full and the refill is wasted.
- How do both answers scale with the size of the fleet?

Where this came from: a battery-backed EV charger

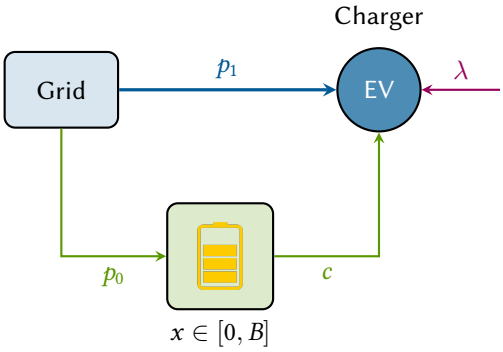
- A battery between grid and charger: recharged slowly, discharged fast.
- Fast charging at p_2 while $x > 0$; when the battery is empty, only the grid rate $p_1 < p_2$.



Joint work with S. Bibiloni (Universidad de la República).

Where this came from: a battery-backed EV charger

- A battery between grid and charger: recharged slowly, discharged fast.
- Fast charging at p_2 while $x > 0$; when the battery is empty, only the grid rate $p_1 < p_2$.
- Single-server loss system [Bibiloni and Ferragut, 2026]: closed form, with decoupling $f_1 = \frac{p_0}{c} f_0$, $f_0 \propto e^{\theta x}$.



Joint work with S. Bibiloni (Universidad de la República).

Strip away the batteries and the cars

- A reservoir, refilled at a constant rate, up to a cap.

Strip away the batteries and the cars

- A reservoir, refilled at a constant rate, up to a cap.
- Users come and go, and draw from it fast while it lasts.

Strip away the batteries and the cars

- A reservoir, refilled at a constant rate, up to a cap.
- Users come and go, and draw from it fast while it lasts.
- When it is empty, everyone is held to the refill rate.

Strip away the batteries and the cars

- A reservoir, refilled at a constant rate, up to a cap.
- Users come and go, and draw from it fast while it lasts.
- When it is empty, everyone is held to the refill rate.

Have we seen this before?

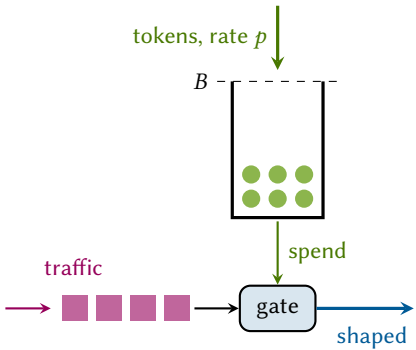
Yes!

It is a token bucket.

The classical tool of traffic shaping in networks.

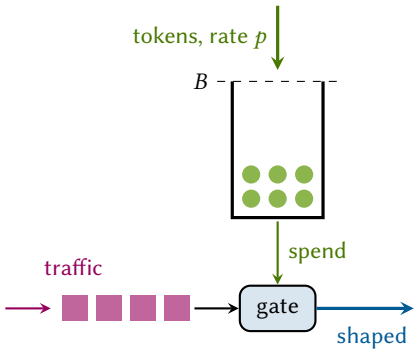
What is a token bucket?

- Tokens drip into a bucket at rate p, up to a depth B.



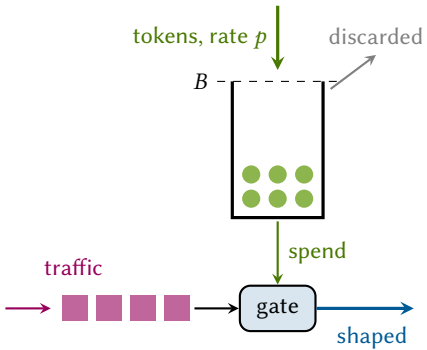
What is a token bucket?

- Tokens drip into a bucket at rate p, up to a depth B.
- Traffic leaves only by spending tokens: bursts up to B pass at once, the long-run rate is p.



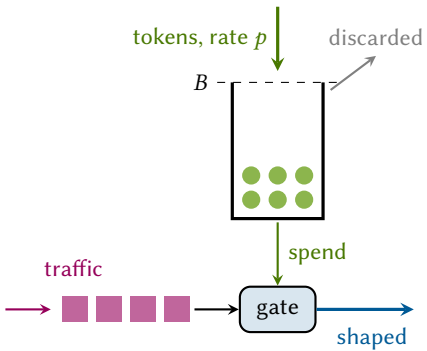
What is a token bucket?

- Tokens drip into a bucket at rate p, up to a depth B.
- Traffic leaves only by spending tokens: bursts up to B pass at once, the long-run rate is p.
- Bucket empty: traffic waits (shaping) or is dropped (policing).
- Bucket full: new tokens are discarded.



What is a token bucket?

- Tokens drip into a bucket at rate p, up to a depth B.
- Traffic leaves only by spending tokens: bursts up to B pass at once, the long-run rate is p.
- Bucket empty: traffic waits (shaping) or is dropped (policing).
- Bucket full: new tokens are discarded.



Introduced to police and shape traffic in ATM and the Internet [Turner, 1986]; today it is the standard rate limiter of APIs and cloud services.

Token buckets are fluid queues

- The token level is a fluid buffer: filled at rate p, drained by the traffic.

Token buckets are fluid queues

- The token level is a fluid buffer: filled at rate p, drained by the traffic.
- Token buckets fed by Markov-modulated on–off sources were analysed with spectral methods in the early 1990s, e.g. [Elwalid and Mitra, 1991].

Token buckets are fluid queues

- The token level is a fluid buffer: filled at rate p, drained by the traffic.
- Token buckets fed by Markov-modulated on–off sources were analysed with spectral methods in the early 1990s, e.g. [Elwalid and Mitra, 1991].

An old fluid queue, in a new setting: AI agents sharing a token quota.

The connection: a shared token bucket

token bucket	agent fleet
tokens in the bucket	remaining quota x
bucket depth	quota cap B
token rate	refill rate p
peak rate	burn rate c of an agent
flows arriving	agents activating
empty: held to the token rate	empty: throttled
full: tokens discarded	full: refill wasted

The connection: a shared token bucket

token bucket	agent fleet
tokens in the bucket	remaining quota x
bucket depth	quota cap B
token rate	refill rate p
peak rate	burn rate c of an agent
flows arriving	agents activating
empty: held to the token rate	empty: throttled
full: tokens discarded	full: refill wasted

The interesting state is the empty bucket.

The new setting

The quota is a shared bucket, drawn on by a random number of agents that come and go. When it runs dry, the agents finish more slowly.

Key points of this talk

- A flow-level model of a token quota shared by a fleet of on–off agents: a PDMP with state (active agents, tokens left).
- All eigenvalues of the stationary problem are real: a short proof from reversibility.
- The stationary law exists, is unique and positive: a 2×2 Perron–Frobenius argument.
- Fluid limit: the bucket ends full or empty; under-provisioning makes the fleet pile up.
- Square-root provisioning (work in progress): a refill margin of order $\sqrt{N}$ above the mean burn should need a quota of order $\sqrt{N}$ only.

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

One agent

- An agent is idle or active. An idle agent starts a task at rate λ.
- A task requires $Y \sim \mathrm{Exp}(\mu)$ tokens, drawn afresh each time.
- An active agent burns tokens at rate c (tokens per unit time).



One agent: rates

- The active period is $Y/c \sim \text{Exp}(\nu)$, with completion rate

$$\nu := \mu c.$$

- Memoryless demands: the state of an agent is just idle/active.

One agent: rates

- The active period is $Y/c \sim \mathrm{Exp}(\nu)$, with completion rate

$$\nu := \mu c.$$

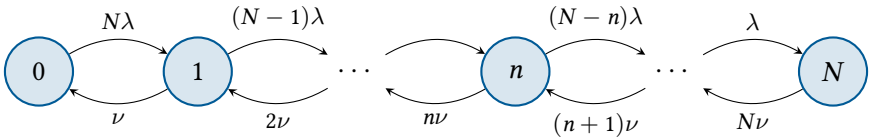
- Memoryless demands: the state of an agent is just idle/active.

Notation

$$\gamma := \lambda + \nu, \qquad \rho := \frac{\lambda}{\gamma} \quad \text{(duty cycle: fraction of time active).}$$

The fleet: an Engset population

N agents, independent as long as tokens remain. Let $n(t)$ = number of active agents.



The finite-source birth–death chain of [Engset, 1918], with generator Q.

The fleet: stationary law

- Stationary law: binomial,

$$\pi_n = \binom{N}{n} \rho^n (1-\rho)^{N-n}, \qquad N\rho \text{ agents active on average.}$$

- Mean burn rate of the fleet: $cN\rho$.

The fleet: stationary law

- Stationary law: binomial,

$$\pi_n = \binom{N}{n} \rho^n (1-\rho)^{N-n}, \qquad N\rho \text{ agents active on average.}$$

- Mean burn rate of the fleet: $cN\rho$.

Reversibility [Kelly, 1979]

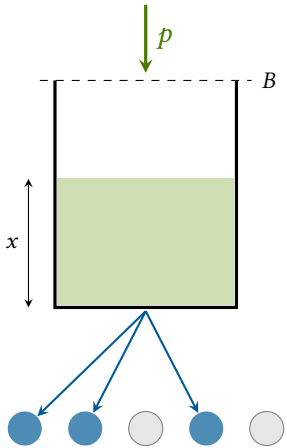
$$\pi_n \, (N-n)\lambda = \pi_{n+1} \, (n+1)\nu.$$

We will use it twice: for the spectrum and for positivity.

The quota

- Tokens left: $x(t) \in [0, B]$.
- The quota refills at rate p.
- With n agents active, the net drift is

$$\dot{x} = r_n := p - nc.$$



$n = 3$ active, each burns c

Filling and draining modes

- The drift $r_n = p - nc$ changes sign once, at $n^* := \lfloor p/c \rfloor$:
  - $n \leq n^*$: the quota fills;
  - $n > n^*$: the quota drains.

Filling and draining modes

- The drift $r_n = p - nc$ changes sign once, at $n^* := \lfloor p/c \rfloor$:
  - $n \leq n^*$: the quota fills;
  - $n > n^*$: the quota drains.

- Standing assumptions: $0 < p < Nc$, and $p/c \notin \mathbb{Z}$, so that no mode has zero drift.

Filling and draining modes

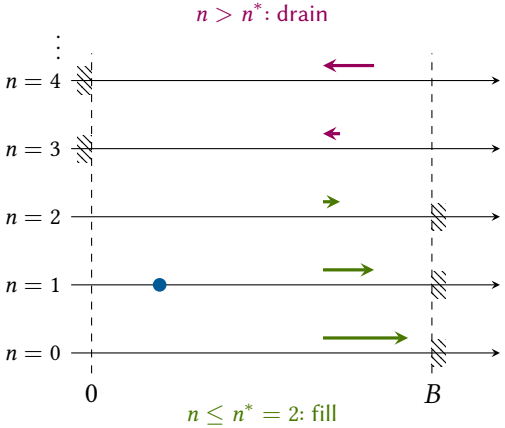
- The drift $r_n = p - nc$ changes sign once, at $n^* := \lfloor p/c \rfloor$:
  - $n \leq n^*$: the quota fills;
  - $n > n^*$: the quota drains.

- Standing assumptions: $0 < p < Nc$, and $p/c \notin \mathbb{Z}$, so that no mode has zero drift.

Mean drift $p - cN\rho$

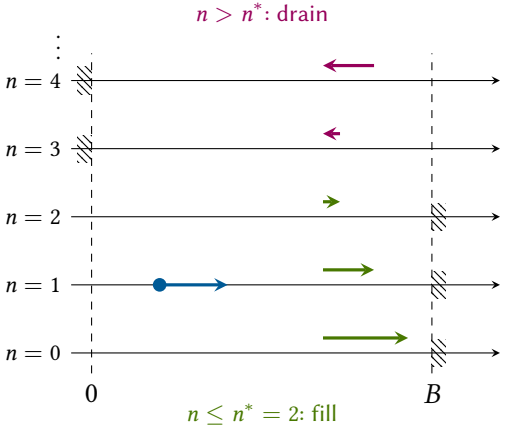
The plan is over-provisioned if $p > cN\rho$, under-provisioned if $p < cN\rho$.

State space and dynamics



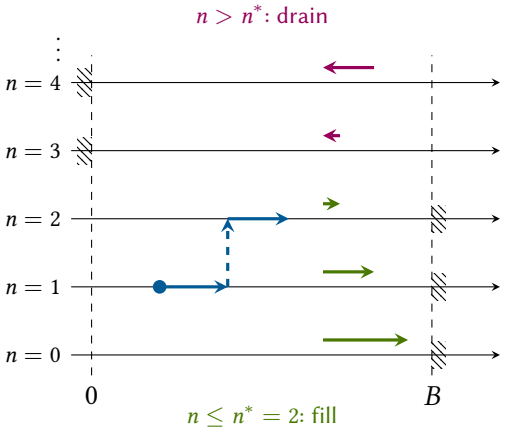
- State: $Z(t) = (n(t), x(t))$.
- Arrows: speed $|r_n|$, here $p/c = 2.5$.

State space and dynamics



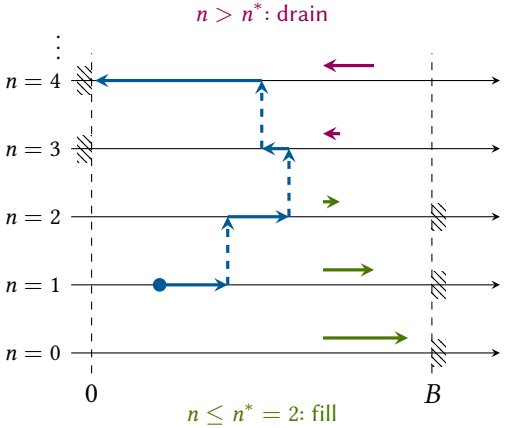
- State: $Z(t) = (n(t), x(t))$.
- Arrows: speed $|r_n|$, here $p/c = 2.5$.
- Between jumps: x moves deterministically; n is frozen.

State space and dynamics



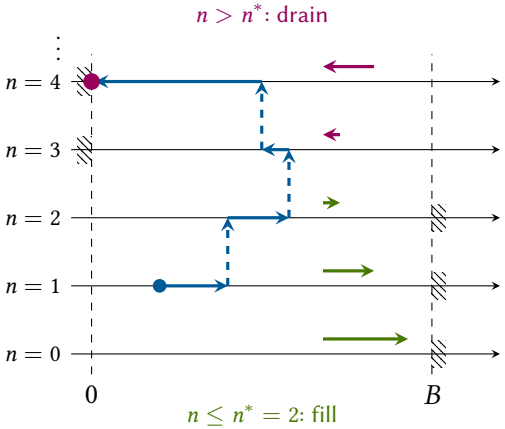
- State: $Z(t) = (n(t), x(t))$.
- Arrows: speed $|r_n|$, here $p/c = 2.5$.
- Between jumps: x moves deterministically; n is frozen.
- Jumps change n by ± 1; x is unchanged.

State space and dynamics



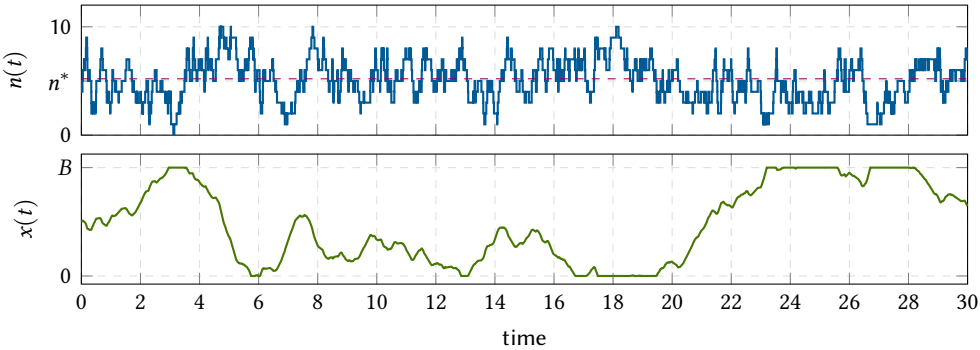
- State: $Z(t) = (n(t), x(t))$.
- Arrows: speed $|r_n|$, here $p/c = 2.5$.
- Between jumps: x moves deterministically; n is frozen.
- Jumps change n by ± 1; x is unchanged.

State space and dynamics



- State: $Z(t) = (n(t), x(t))$.
- Arrows: speed $|r_n|$, here $p/c = 2.5$.
- Between jumps: x moves deterministically; n is frozen.
- Jumps change n by ± 1; x is unchanged.
- The flow reaches the boundary in finite time, and sticks there.

A sample path



$N = 20$, $\rho = 0.25$ ($N\rho = 5$), $c = 1$, refill $p = 5.2$, $B = 4$. Long run: quota empty 9% of the time, full 15%, in between 76%.

The full bucket

- If $x = B$ and $n \leq n^*$, the refill would push the quota above its cap: the excess is discarded.
- The quota sticks at B; the agents are unaffected.
- It leaves B when the fleet crosses to $n = n^* + 1$.

The full bucket

- If $x = B$ and $n \leq n^*$, the refill would push the quota above its cap: the excess is discarded.
- The quota sticks at B; the agents are unaffected.
- It leaves B when the fleet crosses to $n = n^* + 1$.
- Atoms of the stationary law at (n, B), $n \leq n^*$.
- Wasted refill rate: $W = \sum_{n \leq n^*} (p - nc)\, P(n = n,\ x = B)$.

The empty bucket

- If $x = 0$ and $n > n^*$, the agents want nc tokens per unit time, but only p arrive.

The empty bucket

- If $x = 0$ and $n > n^*$, the agents want nc tokens per unit time, but only p arrive.
- Tokens are the unit of work: a task is done when its Y tokens have been burnt.

The empty bucket

- If $x = 0$ and $n > n^*$, the agents want nc tokens per unit time, but only p arrive.
- Tokens are the unit of work: a task is done when its Y tokens have been burnt.
- A starved agent does not slow down a stream: it stops progressing.

The empty bucket: throttling

The p incoming tokens are shared among the n active agents: each burns at rate p/n, and completes at rate $\mu p/n$.

The empty bucket: throttling

The p incoming tokens are shared among the n active agents: each burns at rate p/n, and completes at rate $\mu p/n$.

Completion rate of the fleet

$$m(n,x) = \begin{cases} n\nu & x > 0 \text{ or } n \le n^*, \\ \mu p & x = 0,\ n > n^* \quad \text{(throttled)}. \end{cases}$$

The empty bucket: throttling

The p incoming tokens are shared among the n active agents: each burns at rate p/n, and completes at rate $\mu p/n$.

Completion rate of the fleet

$$m(n,x) = \begin{cases} n\nu & x > 0 \text{ or } n \le n^*, \\ \mu p & x = 0,\ n > n^* \quad \text{(throttled)}. \end{cases}$$

- Activations continue at $(N-n)\lambda$; the quota stays at 0 until n drops to n^*.
- Work is conserved: no task is lost, the fleet just progresses at the refill rate.
- At $x = 0$ the fleet's rates depend on the quota: the chain $n(t)$ is not autonomous.

A piecewise-deterministic Markov process

- $Z(t) = (n(t), x(t))$ is a PDMP [Davis, 1984, 1993]:
  - deterministic flow $\dot{x} = r_n$ between jumps;
  - state-dependent jump rates $(N - n)\lambda$ and $m(n, x)$;
  - jumps that leave x unchanged.

A piecewise-deterministic Markov process

- $Z(t) = (n(t), x(t))$ is a PDMP [Davis, 1984, 1993]:
  - deterministic flow $\dot{x} = r_n$ between jumps;
  - state-dependent jump rates $(N - n)\lambda$ and $m(n, x)$;
  - jumps that leave x unchanged.

- It sits slightly outside Davis's framework: compact state space, and a flow that reaches the boundary.

- The construction is explicit, and the strong Markov property holds.

Extended generator

For $\psi(n,\cdot)$ absolutely continuous:

$$\mathcal{A}\psi(n,x) = r_n\,\partial_x\psi(n,x) + (N-n)\lambda\,[\psi(n+1,x)-\psi(n,x)] + m(n,x)\,[\psi(n-1,x)-\psi(n,x)],$$

with the transport term switched off at the sticky boundaries.

Extended generator

For $\psi(n,\cdot)$ absolutely continuous:

$$\mathcal{A}\psi(n,x) = r_n\,\partial_x\psi(n,x) + (N-n)\lambda\,[\psi(n+1,x)-\psi(n,x)] + m(n,x)\,[\psi(n-1,x)-\psi(n,x)],$$

with the transport term switched off at the sticky boundaries.

A stationary law π satisfies $\int \mathcal{A}\psi\, d\pi = 0$ for all ψ in the domain.

The stationary measure

We seek a mixed measure:

- densities $f_n(x)$ on $(0, B)$;
- atoms π_n^B at $x = B$ for the filling modes $n \leq n^*$;
- atoms π_n^0 at $x = 0$ for the draining modes $n > n^*$.

The stationary measure

We seek a mixed measure:

- densities $f_n(x)$ on $(0, B)$;
- atoms π_n^B at $x = B$ for the filling modes $n \leq n^*$;
- atoms π_n^0 at $x = 0$ for the draining modes $n > n^*$.

Interior, with $\mathbf{f} = (f_0, \ldots, f_N)$ a row vector and $R = \operatorname{diag}(r_n)$:

$$\mathbf{f}'(x)\, R = \mathbf{f}(x)\, Q, \qquad x \in (0, B).$$

Boundary conditions

At $x = B$:

- draining modes enter only through the gateway $n^* \to n^* + 1$:

$$|r_{n^*+1}|\, f_{n^*+1}(B^-) = (N - n^*)\lambda\, \pi^B_{n^*},$$

  and $f_n(B^-) = 0$ for $n > n^* + 1$;
- atoms: Engset balance on $\{0, \dots, n^*\}$, fed by the interior.

At $x = 0$:

- filling modes enter only through the gateway $n^* + 1 \to n^*$:

$$r_{n^*}\, f_{n^*}(0^+) = \mu p\; \pi^0_{n^*+1},$$

  and $f_n(0^+) = 0$ for $n < n^*$;
- atoms: throttled balance on $\{n^* + 1, \dots, N\}$, deaths at rate μp.

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

Related models

- Fluid buffers modulated by on–off sources have a long history: [Kosten, 1974], [Anick et al., 1982], the finite buffer of [Mitra, 1988], and token buffers [Elwalid and Mitra, 1991].

Related models

- Fluid buffers modulated by on–off sources have a long history: [Kosten, 1974], [Anick et al., 1982], the finite buffer of [Mitra, 1988], and token buffers [Elwalid and Mitra, 1991].

- Sources whose rates depend on the buffer: feedback fluid queues [Mandjes et al., 2003].

Related models

- Fluid buffers modulated by on–off sources have a long history: [Kosten, 1974], [Anick et al., 1982], the finite buffer of [Mitra, 1988], and token buffers [Elwalid and Mitra, 1991].
- Sources whose rates depend on the buffer: feedback fluid queues [Mandjes et al., 2003].

We use the same objects (eigenvalues, boundary conditions), with self-contained proofs.

Exponential solutions

Try $\mathbf{f}(x) = \varphi\, e^{\theta x}$ in $\mathbf{f}' R = \mathbf{f} Q$:

$$\theta\, \varphi R = \varphi Q.$$

Exponential solutions

Try $\mathbf{f}(x) = \varphi\, e^{\theta x}$ in $\mathbf{f}' R = \mathbf{f} Q$:

$$\theta\, \varphi R = \varphi Q.$$

- R is invertible ($p/c \notin \mathbb{Z}$): $N+1$ eigenpairs (θ_k, φ_k), and

$$\mathbf{f}(x) = \sum_{k=0}^{N} a_k\, \varphi_k\, e^{\theta_k x}.$$

- $\theta = 0$ is always an eigenvalue, with $\varphi = \pi$ the binomial law.

Two questions about the solution

- The coefficients a_k and the atoms are fixed by the boundary conditions.
- There are as many conditions as unknowns.

Two questions about the solution

- The coefficients a_k and the atoms are fixed by the boundary conditions.
- There are as many conditions as unknowns.

Are the θ_k real?

Is the resulting measure positive, and unique?

All the eigenvalues are real

Theorem

Every eigenvalue θ of $\theta\,\varphi R = \varphi Q$ is real.

All the eigenvalues are real

Theorem

Every eigenvalue θ of $\theta\,\varphi R = \varphi Q$ is real.

- φ is a row vector, possibly complex; $R = \operatorname{diag}(r_n)$ is real and invertible.

All the eigenvalues are real

Theorem

Every eigenvalue θ of $\theta\,\varphi R = \varphi Q$ is real.

- φ is a row vector, possibly complex; $R = \operatorname{diag}(r_n)$ is real and invertible.
- Divide the eigenvector by the stationary law:

$$h_n := \frac{\varphi_n}{\pi_n}, \qquad \text{i.e.} \qquad \varphi = h^\top \Pi, \quad \Pi = \operatorname{diag}(\pi_n),\ \pi_n > 0.$$

All the eigenvalues are real

Theorem

Every eigenvalue θ of $\theta\,\varphi R = \varphi Q$ is real.

- φ is a row vector, possibly complex; $R = \operatorname{diag}(r_n)$ is real and invertible.
- Divide the eigenvector by the stationary law:

$$h_n := \frac{\varphi_n}{\pi_n}, \qquad \text{i.e.} \qquad \varphi = h^\top \Pi, \quad \Pi = \operatorname{diag}(\pi_n),\ \pi_n > 0.$$

- For $\theta = 0$: $\varphi = \pi$ and $h \equiv 1$.

Step 1: from φ to h

Two facts:

- $\Pi R = R\Pi$, since both are diagonal;
- $\Pi Q = Q^\top \Pi$, i.e. $\pi_n Q_{nm} = \pi_m Q_{mn}$: this is detailed balance.

Step 1: from φ to h

Two facts:

- $\Pi R = R\Pi$, since both are diagonal;
- $\Pi Q = Q^\top \Pi$, i.e. $\pi_n Q_{nm} = \pi_m Q_{mn}$: this is detailed balance.

Substitute $\varphi = h^\top \Pi$ in $\theta\, \varphi R = \varphi Q$:

$$\theta\, h^\top \Pi R = h^\top \Pi Q$$

Step 1: from φ to h

Two facts:

- $\Pi R = R\Pi$, since both are diagonal;
- $\Pi Q = Q^\top \Pi$, i.e. $\pi_n Q_{nm} = \pi_m Q_{mn}$: this is detailed balance.

Substitute $\varphi = h^\top \Pi$ in $\theta\, \varphi R = \varphi Q$:

$$\theta\, h^\top \Pi R = h^\top \Pi Q$$
$$\theta\, h^\top R \Pi = h^\top Q^\top \Pi \qquad \text{(the two facts)}$$

Step 1: from φ to h

Two facts:

- $\Pi R = R\Pi$, since both are diagonal;
- $\Pi Q = Q^\top \Pi$, i.e. $\pi_n Q_{nm} = \pi_m Q_{mn}$: this is detailed balance.

Substitute $\varphi = h^\top \Pi$ in $\theta\, \varphi R = \varphi Q$:

$$\begin{aligned} \theta\, h^\top \Pi R &= h^\top \Pi Q \\ \theta\, h^\top R \Pi &= h^\top Q^\top \Pi && \text{(the two facts)} \\ \theta\, (Rh)^\top \Pi &= (Qh)^\top \Pi && (R \text{ symmetric}). \end{aligned}$$

Step 1: from φ to h

Two facts:

- $\Pi R = R\Pi$, since both are diagonal;
- $\Pi Q = Q^\top \Pi$, i.e. $\pi_n Q_{nm} = \pi_m Q_{mn}$: this is detailed balance.

Substitute $\varphi = h^\top \Pi$ in $\theta\, \varphi R = \varphi Q$:

$$\begin{aligned} \theta\, h^\top \Pi R &= h^\top \Pi Q \\ \theta\, h^\top R \Pi &= h^\top Q^\top \Pi && \text{(the two facts)} \\ \theta\, (Rh)^\top \Pi &= (Qh)^\top \Pi && (R \text{ symmetric}). \end{aligned}$$

Cancel the invertible Π and transpose:

$$Q\, h = \theta\, R\, h.$$

Self-adjointness

Inner product weighted by the stationary law:

$$\langle u, v \rangle_\pi := \sum_n \pi_n \, u_n \, \bar{v}_n.$$

Self-adjointness

Inner product weighted by the stationary law:

$$\langle u, v \rangle_\pi := \sum_n \pi_n \, u_n \, \bar{v}_n .$$

- Q is self-adjoint, by detailed balance:

$$\langle Qu, v \rangle_\pi = \sum_{n,m} \pi_n Q_{nm} \, u_m \bar{v}_n = \sum_{n,m} \pi_m Q_{mn} \, u_m \bar{v}_n = \langle u, Qv \rangle_\pi ,$$

  using that Q is real.

Self-adjointness

Inner product weighted by the stationary law:

$$\langle u, v \rangle_\pi := \sum_n \pi_n \, u_n \, \bar{v}_n .$$

- Q is self-adjoint, by detailed balance:

$$\langle Qu, v \rangle_\pi = \sum_{n,m} \pi_n Q_{nm} \, u_m \bar{v}_n = \sum_{n,m} \pi_m Q_{mn} \, u_m \bar{v}_n = \langle u, Qv \rangle_\pi ,$$

  using that Q is real.

- R is real and diagonal, so it is self-adjoint too.

Self-adjointness

Inner product weighted by the stationary law:

$$\langle u, v\rangle_\pi := \sum_n \pi_n\, u_n\, \bar{v}_n.$$

- Q is self-adjoint, by detailed balance:

$$\langle Qu, v\rangle_\pi = \sum_{n,m} \pi_n Q_{nm}\, u_m \bar{v}_n = \sum_{n,m} \pi_m Q_{mn}\, u_m \bar{v}_n = \langle u, Qv\rangle_\pi ,$$

  using that Q is real.

- R is real and diagonal, so it is self-adjoint too.

- Hence $\langle Qh, h\rangle_\pi$ and $\langle Rh, h\rangle_\pi$ are real, for every complex h.

All the eigenvalues are real: proof

Let $Qh = \theta Rh$, $h \neq 0$.

- Step 2. Take the inner product with h:

$$\langle Qh, h\rangle_\pi = \theta\ \langle Rh, h\rangle_\pi\,, \qquad \text{both brackets real.}$$

All the eigenvalues are real: proof

Let $Qh = \theta Rh$, $h \neq 0$.

- Step 2. Take the inner product with h:

$$\langle Qh, h\rangle_\pi = \theta\ \langle Rh, h\rangle_\pi\,, \qquad \text{both brackets real.}$$

- Step 3. Q is a generator (rows sum to 0) and $\pi_n Q_{nm}$ is symmetric, so

$$\langle Qh, h\rangle_\pi = -\frac{1}{2}\sum_{m\neq n} \pi_n Q_{nm}\,|h_n - h_m|^2 \leq 0,$$

with equality iff h is constant (Q irreducible).

All the eigenvalues are real: end of proof

- Step 4. If $\langle Rh, h\rangle_\pi \neq 0$:

$$\theta = \frac{\langle Qh, h\rangle_\pi}{\langle Rh, h\rangle_\pi} \in \mathbb{R}.$$

All the eigenvalues are real: end of proof

- Step 4. If $\langle Rh, h\rangle_\pi \neq 0$:
$$\theta = \frac{\langle Qh, h\rangle_\pi}{\langle Rh, h\rangle_\pi} \in \mathbb{R}.$$

- If $\langle Rh, h\rangle_\pi = 0$: then $\langle Qh, h\rangle_\pi = 0$, so h is constant and $Qh = 0$.
Thus $\theta\, Rh = 0$ with $Rh \neq 0$ (R invertible), so $\theta = 0$. □

All the eigenvalues are real: end of proof

- Step 4. If $\langle Rh, h\rangle_\pi \neq 0$:

$$\theta = \frac{\langle Qh, h\rangle_\pi}{\langle Rh, h\rangle_\pi} \in \mathbb{R}.$$

- If $\langle Rh, h\rangle_\pi = 0$: then $\langle Qh, h\rangle_\pi = 0$, so h is constant and $Qh = 0$.
  Thus $\theta\, Rh = 0$ with $Rh \neq 0$ (R invertible), so $\theta = 0$. □

Step 3 is needed because R is indefinite: self-adjointness alone allows $\langle Rh, h\rangle_\pi = 0$.

Reality is proved by other routes in [Anick et al., 1982] (an explicit discriminant) and [Mitra, 1988] (a Schur complement and Sylvester's law of inertia).

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

What the spectral solution does not tell us

- Equal numbers of conditions and unknowns do not make the linear system non-singular.
- Nor do they make the resulting measure non-negative: exponents of both signs, coefficients of both signs.
- For the classical finite buffer, [Mitra, 1988] does not prove either.

What the spectral solution does not tell us

- Equal numbers of conditions and unknowns do not make the linear system non-singular.
- Nor do they make the resulting measure non-negative: exponents of both signs, coefficients of both signs.
- For the classical finite buffer, [Mitra, 1988] does not prove either.

We use a probabilistic route instead of the linear algebra.

Existence and uniqueness

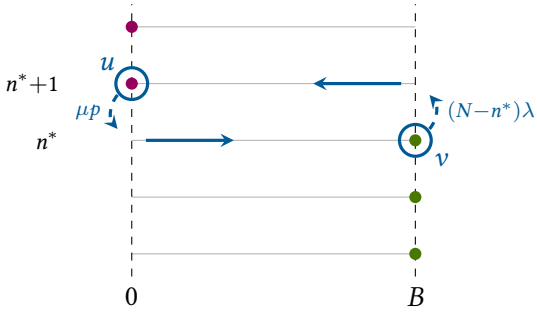
Theorem

For $0 < p < Nc$, $p/c \notin \mathbb{Z}$, the process has a unique stationary distribution.
Its densities are positive on $(0, B)$ and its atoms are positive.

The gateways

- Only two atoms leak into the interior:

$$u = \pi^{0}_{n^*+1}, \qquad v = \pi^{B}_{n^*}.$$

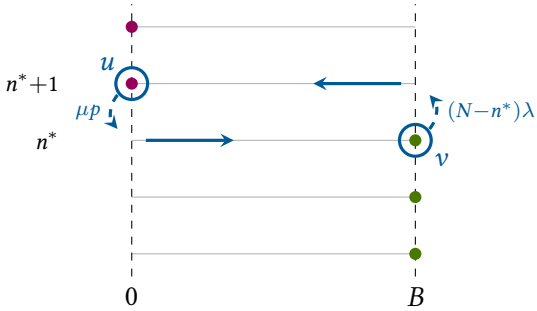


The gateways

- Only two atoms leak into the interior:

$$u = \pi^{0}_{n^*+1}, \qquad v = \pi^{B}_{n^*}.$$

- Given (u, v), the interior and all other atoms are determined.



Proof, part 1: positive boundary data, positive interior

- Fix gateway data $(u, v) \geq 0$, not both zero.

Proof, part 1: positive boundary data, positive interior

- Fix gateway data $(u, v) \geq 0$, not both zero.
- Solve the interior system $\mathbf{f}'R = \mathbf{f}Q$, with each mode's condition at its inflow end.

Proof, part 1: positive boundary data, positive interior

- Fix gateway data $(u, v) \geq 0$, not both zero.
- Solve the interior system $\mathbf{f}'R = \mathbf{f}Q$, with each mode's condition at its inflow end.
- The solution is positive on $(0, B)$: a maximum principle, which again rests on reversibility.

Proof, part 1: positive boundary data, positive interior

- Fix gateway data $(u, v) \geq 0$, not both zero.
- Solve the interior system $\mathbf{f}'R = \mathbf{f}Q$, with each mode's condition at its inflow end.
- The solution is positive on $(0, B)$: a maximum principle, which again rests on reversibility.

Positive boundary data $\Rightarrow$ positive interior.
It remains to find the boundary data.

Proof, part 2: the gateways

- Feeding the interior back to the boundaries: $(u', v') = T(u, v)$, with T a 2×2 matrix, $T > 0$ by part 1.
- A stationary law needs a fixed point: $T(u, v) = (u, v)$.

Proof, part 2: the gateways

- Feeding the interior back to the boundaries: $(u', v') = T(u, v)$, with T a 2×2 matrix, $T > 0$ by part 1.
- A stationary law needs a fixed point: $T(u, v) = (u, v)$.
- Flow conservation: what leaves the boundaries through the gateways comes back. Testing the balance equations with $\psi \equiv 1$:

$$w^\top T = w^\top, \qquad w = \big(\mu p,\ (N - n^*)\lambda\big) > 0.$$

Proof, part 2: the gateways

- Feeding the interior back to the boundaries: $(u', v') = T(u, v)$, with T a 2×2 matrix, $T > 0$ by part 1.
- A stationary law needs a fixed point: $T(u, v) = (u, v)$.
- Flow conservation: what leaves the boundaries through the gateways comes back. Testing the balance equations with $\psi \equiv 1$:

$$w^\top T = w^\top, \qquad w = \big(\mu p,\ (N - n^*)\lambda\big) > 0.$$

- Perron–Frobenius: a positive left eigenvector with eigenvalue 1 makes 1 the spectral radius of T, simple, with a positive right eigenvector (u, v), unique up to scale.

Proof, part 2: the other atoms

- With the gateways known, the remaining atoms solve a birth–death recursion on each side:
  - at $x = B$, on $\{0, \dots, n^*\}$, starting from v;
  - at $x = 0$, on $\{n^* + 1, \dots, N\}$, starting from u.

Proof, part 2: the other atoms

- With the gateways known, the remaining atoms solve a birth–death recursion on each side:
  - at $x = B$, on $\{0, \dots, n^*\}$, starting from v;
  - at $x = 0$, on $\{n^* + 1, \dots, N\}$, starting from u.
- The inputs from the interior are positive, so every atom is positive.

Proof, part 2: the other atoms

- With the gateways known, the remaining atoms solve a birth–death recursion on each side:
  - at $x = B$, on $\{0, \dots, n^*\}$, starting from v;
  - at $x = 0$, on $\{n^* + 1, \dots, N\}$, starting from u.
- The inputs from the interior are positive, so every atom is positive.
- Normalise. Uniqueness comes from the simplicity of the Perron eigenvalue. $\square$

Work in progress: feedback fluid queues

- We are exploring the connection with feedback fluid queues [Mandjes et al., 2003], where the sources' rates depend on the buffer.

Work in progress: feedback fluid queues

- We are exploring the connection with feedback fluid queues [Mandjes et al., 2003], where the sources' rates depend on the buffer.
- There, excursions away from the empty buffer are unaffected by the feedback.

Work in progress: feedback fluid queues

- We are exploring the connection with feedback fluid queues [Mandjes et al., 2003], where the sources' rates depend on the buffer.
- There, excursions away from the empty buffer are unaffected by the feedback.
- If this carries over, the whole effect of the throttle reduces to one number: the mass at the empty bucket.

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

The two costs

Throttling

$$P_{\text{thr}} = \sum_{n > n^*} \pi_n^0$$

fraction of time the fleet is held to p.

Waste

$$W = \sum_{n \le n^*} (p - nc)\, \pi_n^B$$

refill discarded at the cap.

Fluid scaling

Scale the plan with the fleet:

$$p = N\bar{p}, \qquad B = N\bar{b}.$$

Fluid scaling

Scale the plan with the fleet:

$$p = N\bar{p}, \qquad B = N\bar{b}.$$

Fluid variables:

$$\bar{n}(t) = \frac{n(t)}{N} \text{ (fraction of agents active)}, \qquad \bar{x}(t) = \frac{x(t)}{N} \text{ (tokens left, per agent)}.$$

The fluid equations

- Interior, $0 < \bar{x} < \bar{b}$:

$$\dot{\bar{n}} = \lambda(1 - \bar{n}) - \nu\,\bar{n}, \qquad \dot{\bar{x}} = \bar{p} - c\,\bar{n}.$$

The fluid equations

- Interior, $0 < \bar{x} < \bar{b}$:

$$\dot{\bar{n}} = \lambda(1 - \bar{n}) - \nu\,\bar{n}, \qquad \dot{\bar{x}} = \bar{p} - c\,\bar{n}.$$

- Full bucket, $\bar{x} = \bar{b}$ and $c\bar{n} < \bar{p}$: $\dot{\bar{x}} = 0$, and $\bar{n}$ as in the interior.

The fluid equations

- Interior, $0 < \bar{x} < \bar{b}$:

$$\dot{\bar{n}} = \lambda(1 - \bar{n}) - \nu\,\bar{n}, \qquad \dot{\bar{x}} = \bar{p} - c\,\bar{n}.$$

- Full bucket, $\bar{x} = \bar{b}$ and $c\bar{n} < \bar{p}$: $\dot{\bar{x}} = 0$, and $\bar{n}$ as in the interior.

- Empty bucket, $\bar{x} = 0$ and $c\bar{n} > \bar{p}$: $\dot{\bar{x}} = 0$, and the fleet is throttled:

$$\dot{\bar{n}} = \lambda(1 - \bar{n}) - \mu\,\bar{p}.$$

The fleet burns exactly the refill $\bar{p}$, and a task needs $1/\mu$ tokens on average.

Convergence

Work in progress

- Expected: $\big(n(t)/N,\ x(t)/N\big) \to \big(\bar{n}(t), \bar{x}(t)\big)$ on compact time intervals, as $N \to \infty$.

Convergence

Work in progress

- Expected: $\big(n(t)/N,\ x(t)/N\big) \to \big(\bar{n}(t), \bar{x}(t)\big)$ on compact time intervals, as $N \to \infty$.
- In the interior, a density-dependent jump process [Kurtz, 1970; Ethier and Kurtz, 1986].

Convergence

Work in progress

- Expected: $\big(n(t)/N,\ x(t)/N\big) \to \big(\bar{n}(t), \bar{x}(t)\big)$ on compact time intervals, as $N \to \infty$.
- In the interior, a density-dependent jump process [Kurtz, 1970; Ethier and Kurtz, 1986].
- At the boundaries the vector field is discontinuous: the limit slides along them.

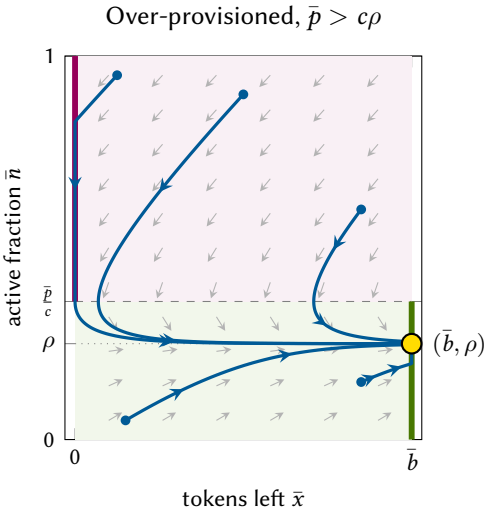
Convergence

Work in progress

- Expected: $\big(n(t)/N,\ x(t)/N\big) \to \big(\bar{n}(t), \bar{x}(t)\big)$ on compact time intervals, as $N \to \infty$.
- In the interior, a density-dependent jump process [Kurtz, 1970; Ethier and Kurtz, 1986].
- At the boundaries the vector field is discontinuous: the limit slides along them.

No formal proof yet; it should follow from standard arguments for density-dependent processes in the Kurtz setting.

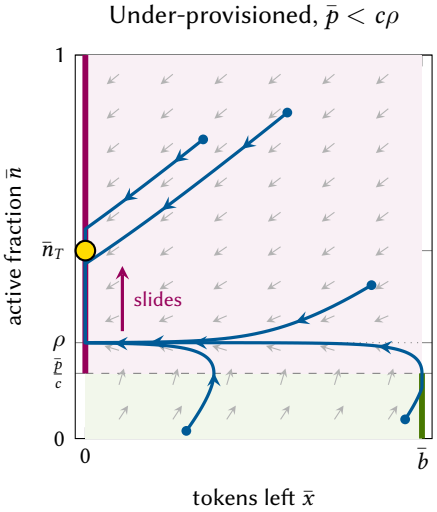
Phase portrait: over-provisioned



- The fleet relaxes to ρ; the bucket fills.
- Equilibrium: bucket full, $\bar{n} = \rho$. No throttling.
- Refill wasted at rate $N(\bar{p} - c\rho)$: of order N.

Green edge: full bucket. Red edge: empty bucket, fleet throttled. $\lambda = 1$, $\nu = 3$, $c = 1$.

Phase portrait: under-provisioned



- The bucket drains and the fleet is throttled.
- The flow slides along the empty boundary…
- …to an equilibrium $\bar{n}_T$ above ρ, with the bucket empty.

Green edge: full bucket. Red edge: empty bucket, fleet throttled. $\lambda = 1$, $\nu = 3$, $c = 1$.

Under-provisioning: the fleet piles up

The flow slides along the empty boundary to an equilibrium that sits above ρ:

$$\bar{n}_T = 1 - \frac{\mu\,\bar{p}}{\lambda} \;>\; \rho.$$

Under-provisioning: the fleet piles up

The flow slides along the empty boundary to an equilibrium that sits above ρ:

$$\bar{n}_T = 1 - \frac{\mu\,\bar{p}}{\lambda} \;>\; \rho.$$

- It stays throttled, $c\bar{n}_T > \bar{p}$, exactly when $\bar{p} < c\rho$.

Under-provisioning: the fleet piles up

The flow slides along the empty boundary to an equilibrium that sits above ρ:

$$\bar{n}_T = 1 - \frac{\mu\,\bar{p}}{\lambda} \;>\; \rho.$$

- It stays throttled, $c\bar{n}_T > \bar{p}$, exactly when $\bar{p} < c\rho$.
- All the refill is burnt, none is wasted.

Under-provisioning: the fleet piles up

The flow slides along the empty boundary to an equilibrium that sits above ρ:

$$\bar{n}_T = 1 - \frac{\mu\,\bar{p}}{\lambda} \;>\; \rho.$$

- It stays throttled, $c\bar{n}_T > \bar{p}$, exactly when $\bar{p} < c\rho$.
- All the refill is burnt, none is wasted.
- Stalled tasks keep more agents busy at once: tasks last $c\bar{n}_T/\bar{p} > 1$ times longer.

Under-provisioning: the fleet piles up

The flow slides along the empty boundary to an equilibrium that sits above ρ:

$$\bar{n}_T = 1 - \frac{\mu\,\bar{p}}{\lambda} \;>\; \rho.$$

- It stays throttled, $c\bar{n}_T > \bar{p}$, exactly when $\bar{p} < c\rho$.
- All the refill is burnt, none is wasted.
- Stalled tasks keep more agents busy at once: tasks last $c\bar{n}_T/\bar{p} > 1$ times longer.
- The fleet relaxes at rate λ instead of γ.

What the fluid limit tells us

- Over-provisioned: the bucket ends full, waste of order N.
- Under-provisioned: the bucket ends empty, and the fleet piles up.

What the fluid limit tells us

- Over-provisioned: the bucket ends full, waste of order N.
- Under-provisioned: the bucket ends empty, and the fleet piles up.
- Balanced, $\bar{p} = c\rho$: at $\bar{n} = \rho$, $\dot{\bar{x}} = 0$, so every level $\bar{x} \in [0, \bar{b}]$ is an equilibrium.

What the fluid limit tells us

- Over-provisioned: the bucket ends full, waste of order N.
- Under-provisioned: the bucket ends empty, and the fleet piles up.
- Balanced, $\bar{p} = c\rho$: at $\bar{n} = \rho$, $\dot{\bar{x}} = 0$, so every level $\bar{x} \in [0, \bar{b}]$ is an equilibrium.

At the fluid scale the quota does not matter, and the balanced plan is degenerate.
The fluctuations decide: zoom in by $\sqrt{N}$.

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

How should the plan scale with the fleet?

- The mean burn $cN\rho$ grows like N.
- The fleet fluctuates around $N\rho$ by $\sigma_N := \sqrt{N\rho(1-\rho)}$, which grows like $\sqrt{N}$.

How should the plan scale with the fleet?

- The mean burn $cN\rho$ grows like N.
- The fleet fluctuates around $N\rho$ by $\sigma_N := \sqrt{N\rho(1-\rho)}$, which grows like $\sqrt{N}$.

Is a refill margin of order $\sqrt{N}$ above the mean enough,
with a quota that is also of order $\sqrt{N}$?

Diffusion scaling

Work in progress

Set the refill and the quota on the scale of the fluctuations:

$$p = cN\rho + \beta\, c\, \sigma_N, \qquad B = b\, c\, \sigma_N.$$

Diffusion scaling

Work in progress

Set the refill and the quota on the scale of the fluctuations:

$$p = cN\rho + \beta\, c\, \sigma_N, \qquad B = b\, c\, \sigma_N.$$

Scale the fleet and the tokens left:

$$\xi = \frac{n - N\rho}{\sigma_N}, \qquad \hat{x} = \frac{x}{c\,\sigma_N} \in [0, b].$$

Diffusion scaling

Work in progress

Set the refill and the quota on the scale of the fluctuations:

$$p = cN\rho + \beta\, c\, \sigma_N, \qquad B = b\, c\, \sigma_N.$$

Scale the fleet and the tokens left:

$$\xi = \frac{n - N\rho}{\sigma_N}, \qquad \hat{x} = \frac{x}{c\,\sigma_N} \in [0, b].$$

- Inside $(0, B)$, exactly: $\dfrac{d\hat{x}}{dt} = \beta - \xi.$
- The fleet alone obeys a CLT: $\xi \Rightarrow$ Ornstein–Uhlenbeck, $d\xi = -\gamma\, \xi\, dt + \sqrt{2\gamma}\, dW.$

The candidate limit

Work in progress

A two-dimensional process $(\hat{x}, \xi)$ on $[0, b] \times \mathbb{R}$:

- Interior: $d\hat{x} = (\beta - \xi)\, dt, \qquad d\xi = -\gamma\, \xi\, dt + \sqrt{2\gamma}\, dW.$

The candidate limit

Work in progress

A two-dimensional process $(\hat{x}, \xi)$ on $[0, b] \times \mathbb{R}$:

- Interior: $d\hat{x} = (\beta - \xi)\, dt, \qquad d\xi = -\gamma\, \xi\, dt + \sqrt{2\gamma}\, dW.$
- Full bucket, $\hat{x} = b$: held there while $\xi < \beta$; the fleet is unaffected.

The candidate limit

Work in progress

A two-dimensional process $(\hat{x}, \xi)$ on $[0, b] \times \mathbb{R}$:

- Interior: $d\hat{x} = (\beta - \xi)\, dt, \qquad d\xi = -\gamma\, \xi\, dt + \sqrt{2\gamma}\, dW.$
- Full bucket, $\hat{x} = b$: held there while $\xi < \beta$; the fleet is unaffected.
- Empty bucket, $\hat{x} = 0$: held there while $\xi > \beta$, and the fleet is throttled:

$$d\xi = -(\lambda\, \xi + \nu\, \beta)\, dt + \sqrt{2\gamma}\, dW.$$

The candidate limit

Work in progress

A two-dimensional process $(\hat{x}, \xi)$ on $[0, b] \times \mathbb{R}$:

- Interior: $d\hat{x} = (\beta - \xi)\, dt, \qquad d\xi = -\gamma\, \xi\, dt + \sqrt{2\gamma}\, dW.$
- Full bucket, $\hat{x} = b$: held there while $\xi < \beta$; the fleet is unaffected.
- Empty bucket, $\hat{x} = 0$: held there while $\xi > \beta$, and the fleet is throttled:

$$d\xi = -(\lambda\, \xi + \nu\, \beta)\, dt + \sqrt{2\gamma}\, dW.$$

Derived formally from the generator; convergence is not yet proved.

Reading the limit

- The quota integrates an Ornstein–Uhlenbeck process: it is not itself a reflected diffusion.
- At the empty bucket the throttle switches the drift of the fleet.

Reading the limit

- The quota integrates an Ornstein–Uhlenbeck process: it is not itself a reflected diffusion.
- At the empty bucket the throttle switches the drift of the fleet.
- The fleet size disappears. What remains:
  - the refill margin β;
  - the quota b, in units of $c\,\sigma_N$;
  - the agents' own rates λ, ν.

If the limit is right

Work in progress

Let $P(\beta, b)$ be the fraction of time the limit process spends at the empty bucket $\hat{x} = 0$. Then

$$P_{\text{thr}} \longrightarrow P(\beta, b) \in (0, 1), \qquad N \to \infty.$$

If the limit is right

Work in progress

Let $P(\beta, b)$ be the fraction of time the limit process spends at the empty bucket $\hat{x} = 0$. Then

$$P_{\text{thr}} \longrightarrow P(\beta, b) \in (0, 1), \qquad N \to \infty.$$

- Throttling stays at a positive level, set by the margin β and the quota b, and not by N.

If the limit is right

Work in progress

Let $P(\beta, b)$ be the fraction of time the limit process spends at the empty bucket $\hat{x} = 0$. Then

$$P_{\text{thr}} \longrightarrow P(\beta, b) \in (0, 1), \qquad N \to \infty.$$

- Throttling stays at a positive level, set by the margin β and the quota b, and not by N.
- The quota is always of order $\sqrt{N}$: $B = b\, c\, \sigma_N$.

If the limit is right

Work in progress

Let $P(\beta, b)$ be the fraction of time the limit process spends at the empty bucket $\hat{x} = 0$. Then

$$P_{\text{thr}} \longrightarrow P(\beta, b) \in (0, 1), \qquad N \to \infty.$$

- Throttling stays at a positive level, set by the margin β and the quota b, and not by N.
- The quota is always of order $\sqrt{N}$: $B = b\, c\, \sigma_N$.
- Relative to the mean burn, both the margin and the quota vanish like $1/\sqrt{N}$.

Square-root staffing, for tokens

	many servers [Halfin and Whitt, 1981]	agent fleet
mean demand	offered load R	mean burn $cN\rho$
fluctuations	$\sqrt{R}$	$c\,\sigma_N \propto \sqrt{N}$
provisioning	$s = R + \beta\sqrt{R}$ servers	$p = cN\rho + \beta\, c\,\sigma_N$ refill
buffer	—	quota $B = b\, c\,\sigma_N$
limit	$P(\text{wait}) \to \alpha(\beta) \in (0,1)$	$P_{\text{thr}} \to P(\beta, b) \in (0,1)$

Square-root staffing, for tokens

	many servers [Halfin and Whitt, 1981]	agent fleet
mean demand	offered load R	mean burn $cN\rho$
fluctuations	$\sqrt{R}$	$c\,\sigma_N \propto \sqrt{N}$
provisioning	$s = R + \beta\sqrt{R}$ servers	$p = cN\rho + \beta\, c\,\sigma_N$ refill
buffer	—	quota $B = b\, c\,\sigma_N$
limit	$P(\text{wait}) \to \alpha(\beta) \in (0,1)$	$P_{\text{thr}} \to P(\beta, b) \in (0,1)$

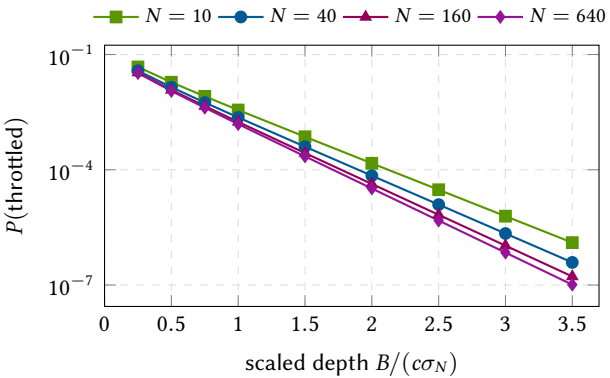
- The same quality-and-efficiency-driven regime: congestion neither vanishes nor dominates.

Square-root staffing, for tokens

	many servers [Halfin and Whitt, 1981]	agent fleet
mean demand	offered load R	mean burn $cN\rho$
fluctuations	$\sqrt{R}$	$c\,\sigma_N \propto \sqrt{N}$
provisioning	$s = R + \beta\sqrt{R}$ servers	$p = cN\rho + \beta\, c\,\sigma_N$ refill
buffer	—	quota $B = b\, c\,\sigma_N$
limit	$P(\text{wait}) \to \alpha(\beta) \in (0,1)$	$P_{\text{thr}} \to P(\beta, b) \in (0,1)$

- The same quality-and-efficiency-driven regime: congestion neither vanishes nor dominates.
- New here: a second knob, the quota, also on the $\sqrt{N}$ scale.

The scaling, numerically



- Margin fixed at $\beta = 1$; quota in units of $c\,\sigma_N$.
- As N grows from 10 to 640, the curves settle: throttling depends on b, not on N.

Balance equations on a grid in x, checked against simulation. $\lambda = 1$, $\nu = 3$, $c = 1$.

The message

Overprovision the refill by order $\sqrt{N}$ above the mean burn,
and a quota of order $\sqrt{N}$ suffices.

Outline

Motivation

The model

Spectral structure

Existence, uniqueness and positivity

Fluid limit

Square-root provisioning

Conclusions

Messages from the talk

- A token quota shared by a fleet of agents is a flow-level token bucket: a PDMP with state (active agents, tokens left).
- The empty bucket throttles the fleet: a feedback fluid queue.
- Real spectrum from reversibility.
- Existence, uniqueness and positivity from a 2×2 Perron–Frobenius argument.
- Fluid limit: full or empty; under-provisioning piles up the fleet above ρ.
- Square-root provisioning (work in progress): a refill margin of order $\sqrt{N}$ above the mean burn, and a quota of order $\sqrt{N}$ suffices.

Future work

- Heavy-tailed token demands: tasks are far from exponential.
- Heterogeneous agents, with different burn rates.
- Congestion: crowded fleets may become less efficient, and burn more tokens per task.
- A proof of the diffusion limit, and of the throttle decomposition.

Thank you!

Andres Ferragut
ferragut@ort.edu.uy
https://aferragu.github.io

References I

D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61(8):1871–1894, 1982.

S. Bibiloni and A. Ferragut. Queueing analysis of a battery-backed electric vehicle fast charging system. In *IFIP Performance*, 2026.

M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B*, 46(3):353–388, 1984.

M. H. A. Davis. *Markov Models and Optimization*. Chapman & Hall, London, 1993.

A. I. Elwalid and D. Mitra. Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Systems*, 9(1):29–63, 1991.

T. Engset. Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern. *Elektrotechnische Zeitschrift*, 39:304–306, 1918.

S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

References II

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

L. Kosten. Stochastic theory of a multi-entry buffer (I). *Delft Progress Report, Series F*, 1:10–18, 1974.

T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

M. Mandjes, D. Mitra, and W. Scheinhardt. Models of network access using feedback fluid queues. *Queueing Systems*, 44:365–398, 2003.

D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20(3):646–676, 1988.

J. S. Turner. New directions in communications (or which way to the information age?). *IEEE Communications Magazine*, 24(10):8–15, 1986.