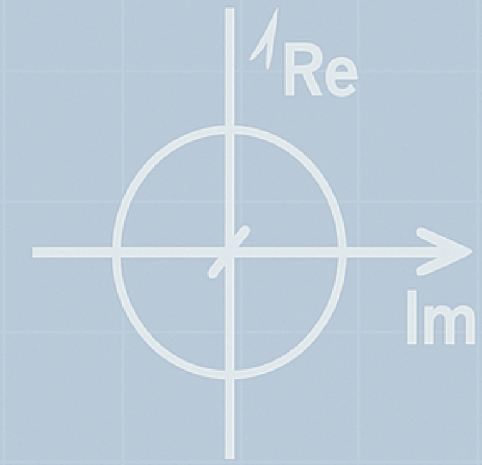
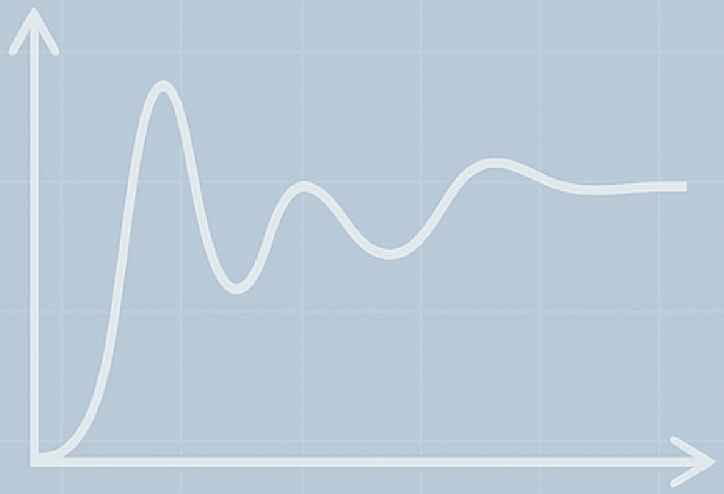
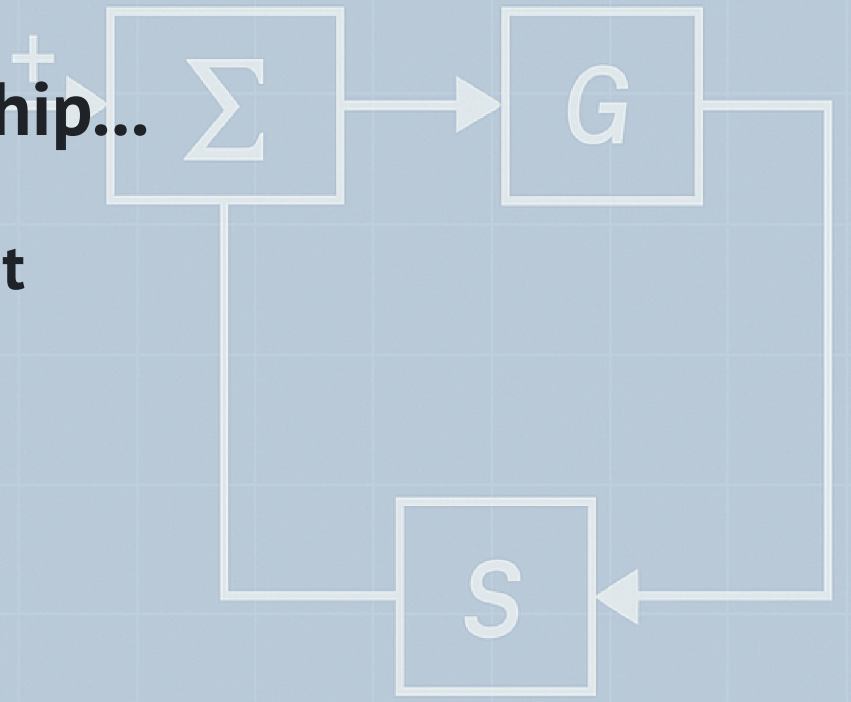


Queueing Theory meets Control

A fruitful friendship...

Andres Ferragut

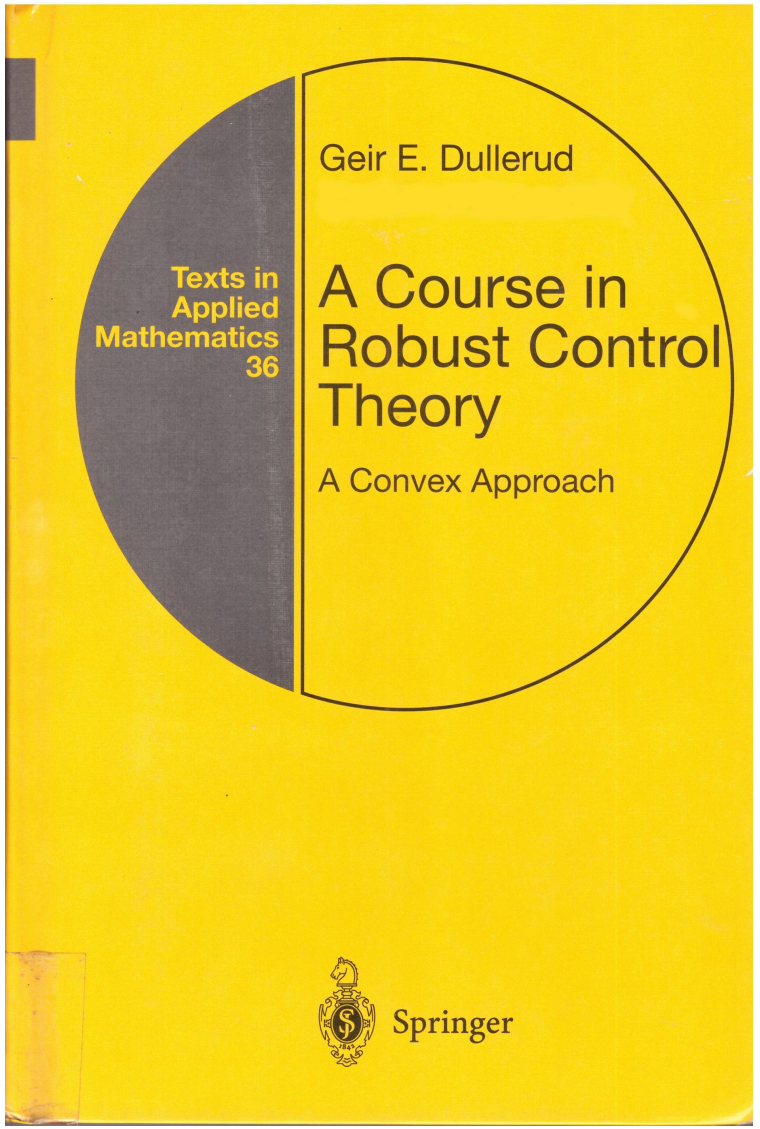




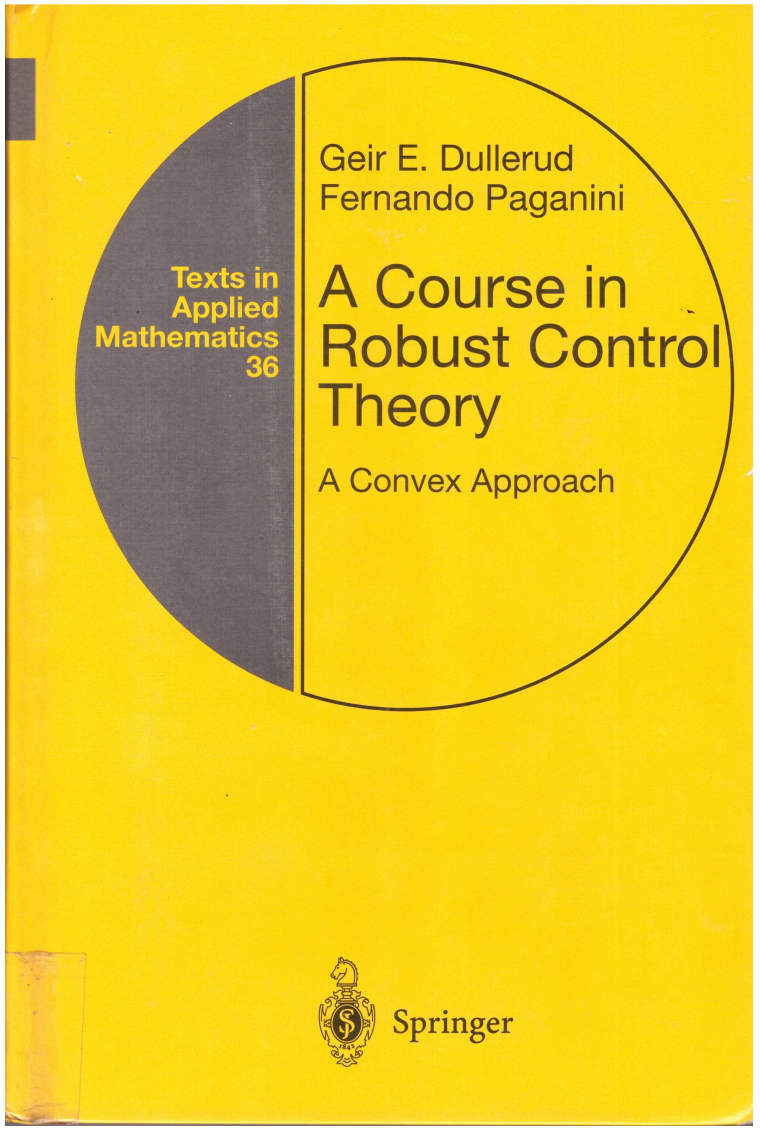
UNIVERSIDAD ORT
Uruguay

A bit of history...

A bit of history...



A bit of history...

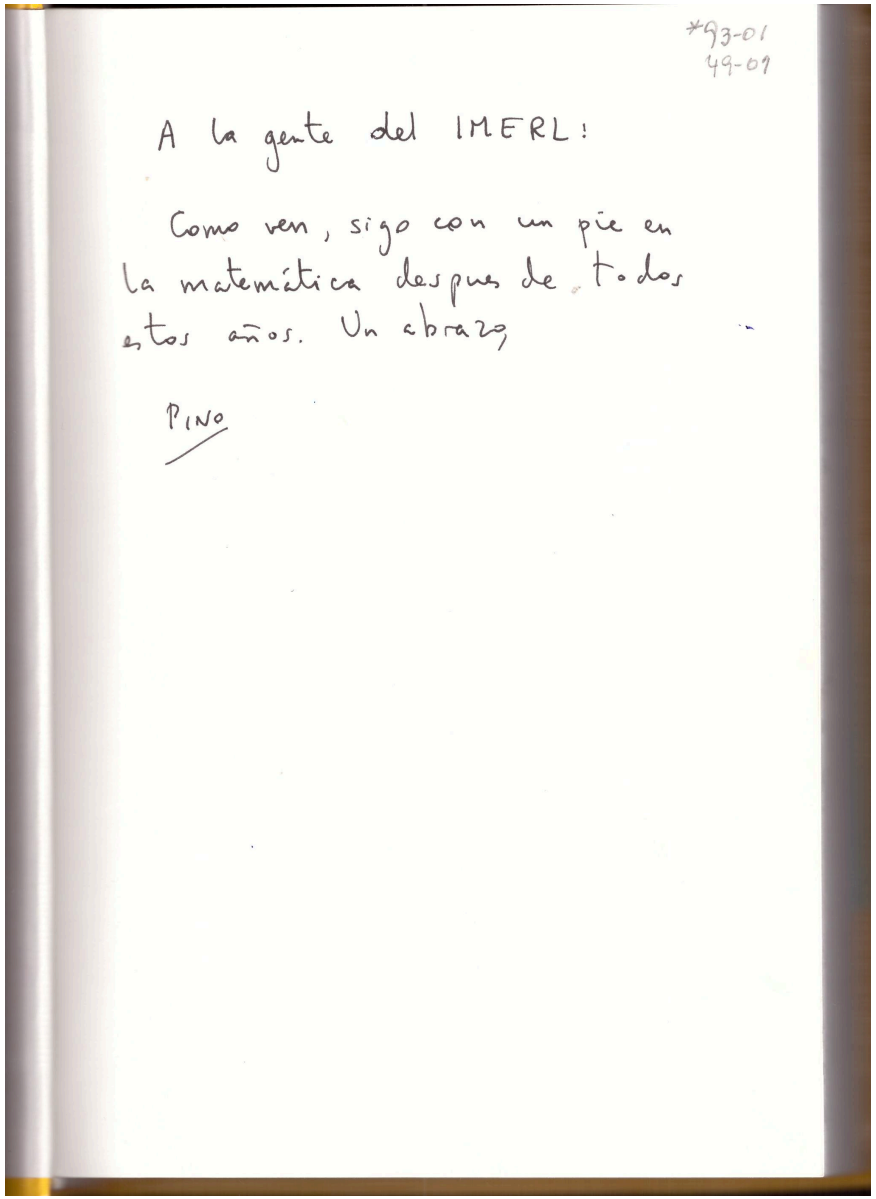


*93-01
49-09

A la gente del IMERL:

Como ven, sigo con un pie en la matemática despues de todos estos años. Un abrazo

PINO



A la gente del IMERL:

Como ven, sigo con un pie en la matemática después de todos estos años. Un abrazo

PINO

To the people of the Institute of Mathematics and Statistics:

As you can see, I still have a foot on mathematics after all these years.

With a hug,

Pino

chapter we begin our study of optimal synthesis and in particular controllers that optimize the H_2 performance criterion. We will by defining the synthesis problem to be solved, and will then provide number of motivating interpretations. Following this, we will develop new matrix tools for the task at hand, before proceeding to solve this control problem.

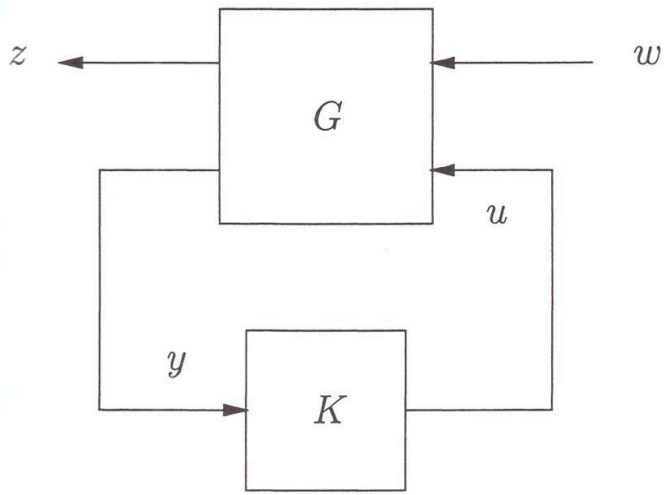


Figure 6.1. Synthesis arrangement

performance criterion of the chapter is defined from the space RH_2 matrix valued transfer functions. This space consists of rational, strictly transfer functions, which have all their poles in the half-plane $\mathbb{C}^-$.

equations. For each $z \in L_2[0,\infty)$ and $x_0 \in \mathbb{C}^n$

$$\begin{aligned}\langle \Psi_o^* z, x_0\rangle_{\mathbb{C}^n} &= \langle z, \Psi_o x_0\rangle_2 = \int_0^\infty z^*(\tau) C e^{A\tau} x_0 d\tau \\ &= \left(\int_0^\infty e^{A^*\tau} C^* z(\tau) d\tau\right)^* x_0 \\ &= \left\langle \int_0^\infty e^{A^*\tau} C^* z(\tau) d\tau,\, x_0 \right\rangle_{\mathbb{C}^n} .\end{aligned}$$

Thus we see that Ψ_o^* is given by

$$\Psi_o^* z = \int_0^\infty e^{A^*\tau} C^* z(\tau) d\tau ,$$

for $z \in L_2[0,\infty)$. Now the 2-norm of $y = \Psi_o x_0$ is given by $\langle x_0, \Psi_o^* \Psi_o x_0\rangle$ and we have

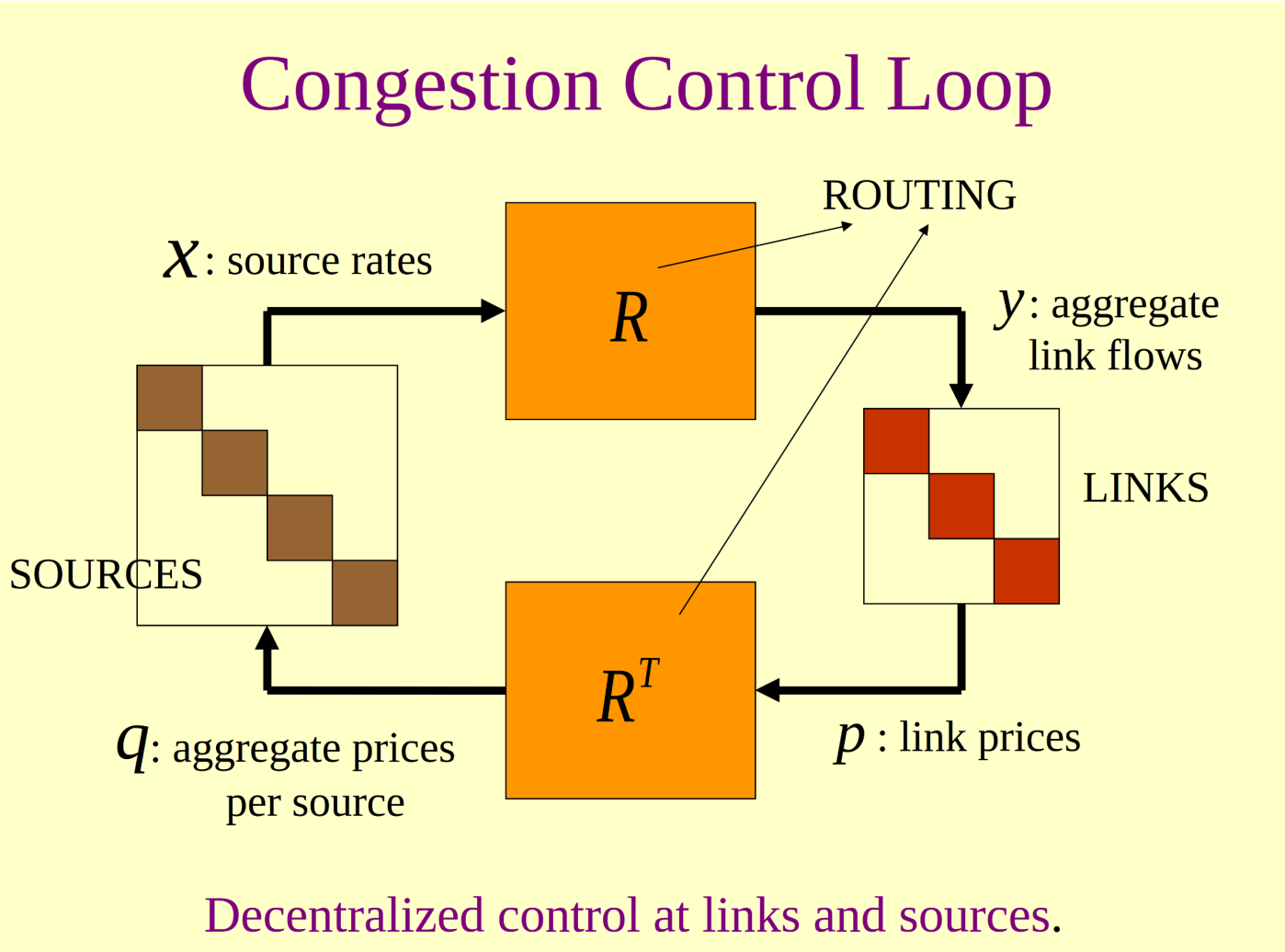
$$\begin{aligned}\Psi_o^*(\Psi_o x_0) &= \int_0^\infty e^{A^*\tau} C^* C e^{A\tau} x_0 d\tau \\ &= \left(\int_0^\infty e^{A^*\tau} C^* C e^{A\tau} d\tau\right) x_0 = (\Psi_o^* \Psi_o) x_0 .\end{aligned}$$

Therefore the operator $\Psi_o^* \Psi_o$ is given by the matrix

$$Y_o := \Psi_o^* \Psi_o = \int_0^\infty e^{A^*\tau} C^* C e^{A\tau} d\tau ,$$

Is this networking?

Is this networking?



Network utility maximization

Network utility maximization

$$\max_{x_i \geqslant 0} \sum_i U_i(x_i)$$

$$\text{subject to:} \quad \sum_i R_{il} x_i \leqslant c_l \quad \forall l$$

- U_i concave, and represents connection utilities derived from bandwidth.
- $R = (R_{il})$ is the routing matrix (1 if route i goes through link l).
- Typical choices:

$$U_i(x) = \frac{w_i x^{1-\alpha}}{1-\alpha} \quad \text{or} \quad U_i(x) = w_i \log(x)$$

Network utility maximization

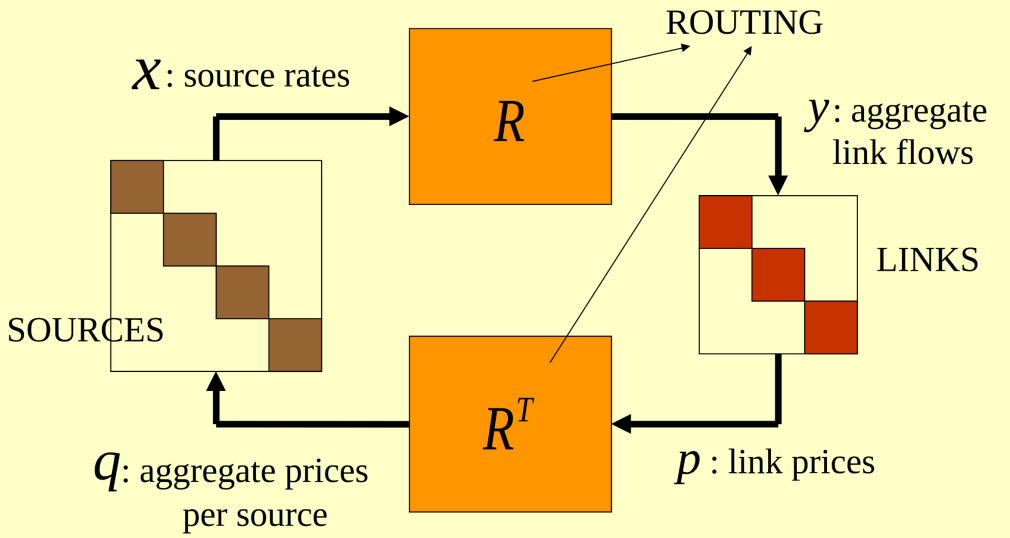
Theorem: The dynamics

$$\dot{x}_i = U_i'(x_i) - q_i \quad \dot{p}_l = [y_l - c_l]_{p_l}^+$$

$$y_l = \sum_i R_{il} x_i \quad q_i = \sum_l R_{il} p_l$$

are globally asymptotically stable and its equilibrium solves the optimization problem.

Congestion Control Loop



Decentralized control at links and sources.

Is the Internet stable?

Is the Internet stable?

- The above result proves that congestion control is stable for a **fixed number** of connections.
- But the Internet is *stochastic* in nature, with **random** request arrivals and service times.
- In the longer timescale, it behaves more like a **queueing system**.
- **Will it be stable?**

Queueing model of the Internet

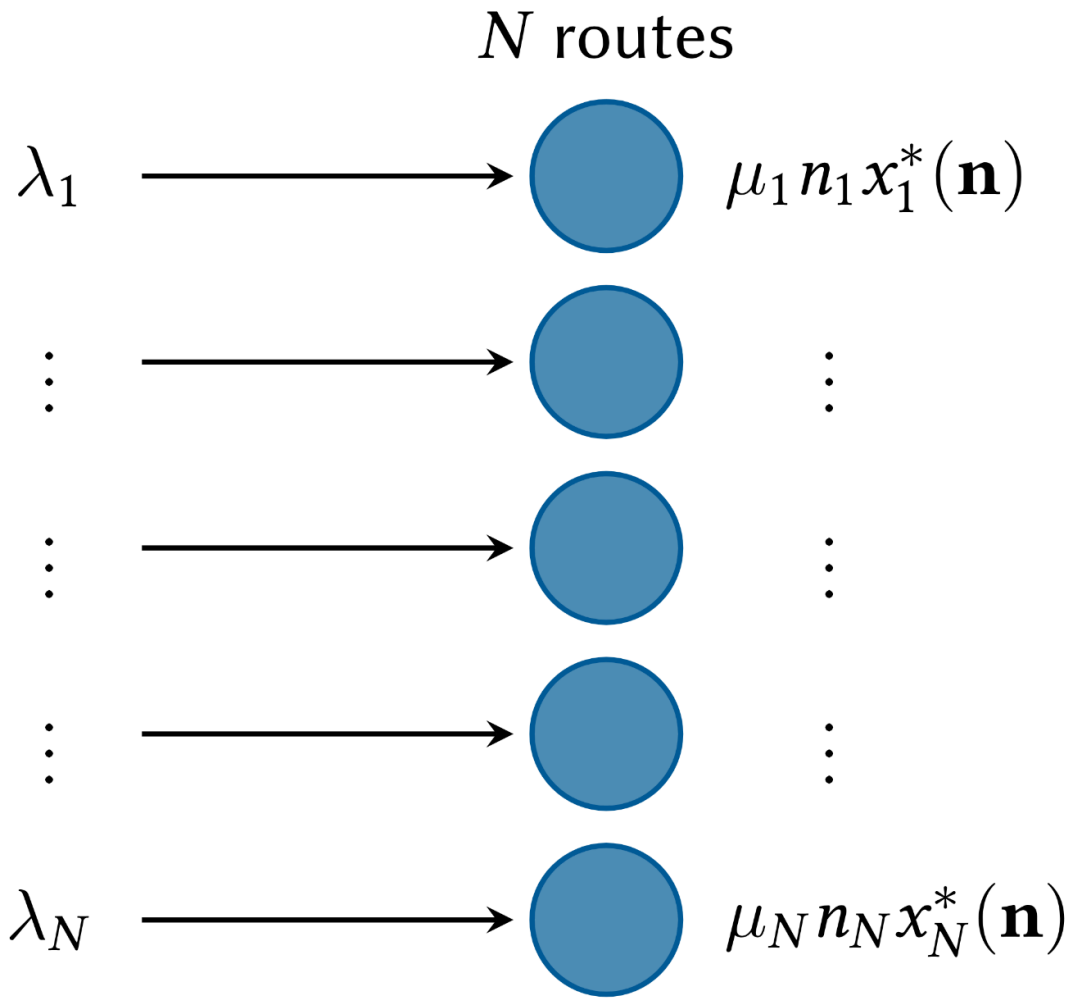
- Let us include the *number of connections* $\mathbf{n} = (n_i)$ in the picture.
- Define:

$$\mathbf{x}^*(\mathbf{n}) = \arg\max_{x_i \geqslant 0} \sum_i n_i U_i(x_i)$$

$$\text{subject to:} \quad \sum_i R_{il} n_i x_i \leqslant c_l \quad \forall l$$

- Connections arrive at each route at rate λ_i and have avg. job size $1/\mu_i$.
- Timescale separation: congestion control solves the map $\mathbf{n} \mapsto \mathbf{x}^*$ quickly...

Queueing model of the Internet



- Each route acts like a queue…
- with rates coupled through $\mathbf{x}^*(\mathbf{n})$
- Avg. request rate on each route

$$\rho_i := \lambda_i / \mu_i$$

- Natural stability condition:

$$R\rho < c$$

Stability and Lyapunov

- For *exponential* file sizes, the state of the system is a Markov Chain:

$$\begin{cases} \mathbf{n} \mapsto \mathbf{n} + e_i & \text{at rate } \lambda_i \\ \mathbf{n} \mapsto \mathbf{n} - e_i & \text{at rate } \mu_i n_i x_i^*(\mathbf{n}) \end{cases}$$

- There is an analog of Lyapunov stability for Markov chains, take:

$$V(\mathbf{n}) = \sum_i \frac{1}{\mu_i} n_i^{\alpha+1}$$

- Under the natural stability condition, the *average drift* satisfies:

$$\Delta V(\mathbf{n}) = \lim_{h \to 0} \frac{1}{h} E\left[V(\mathbf{n}(t+h)) - V(\mathbf{n}(t)) \mid \mathbf{n}(t)\right] \leqslant -\varepsilon < 0,$$

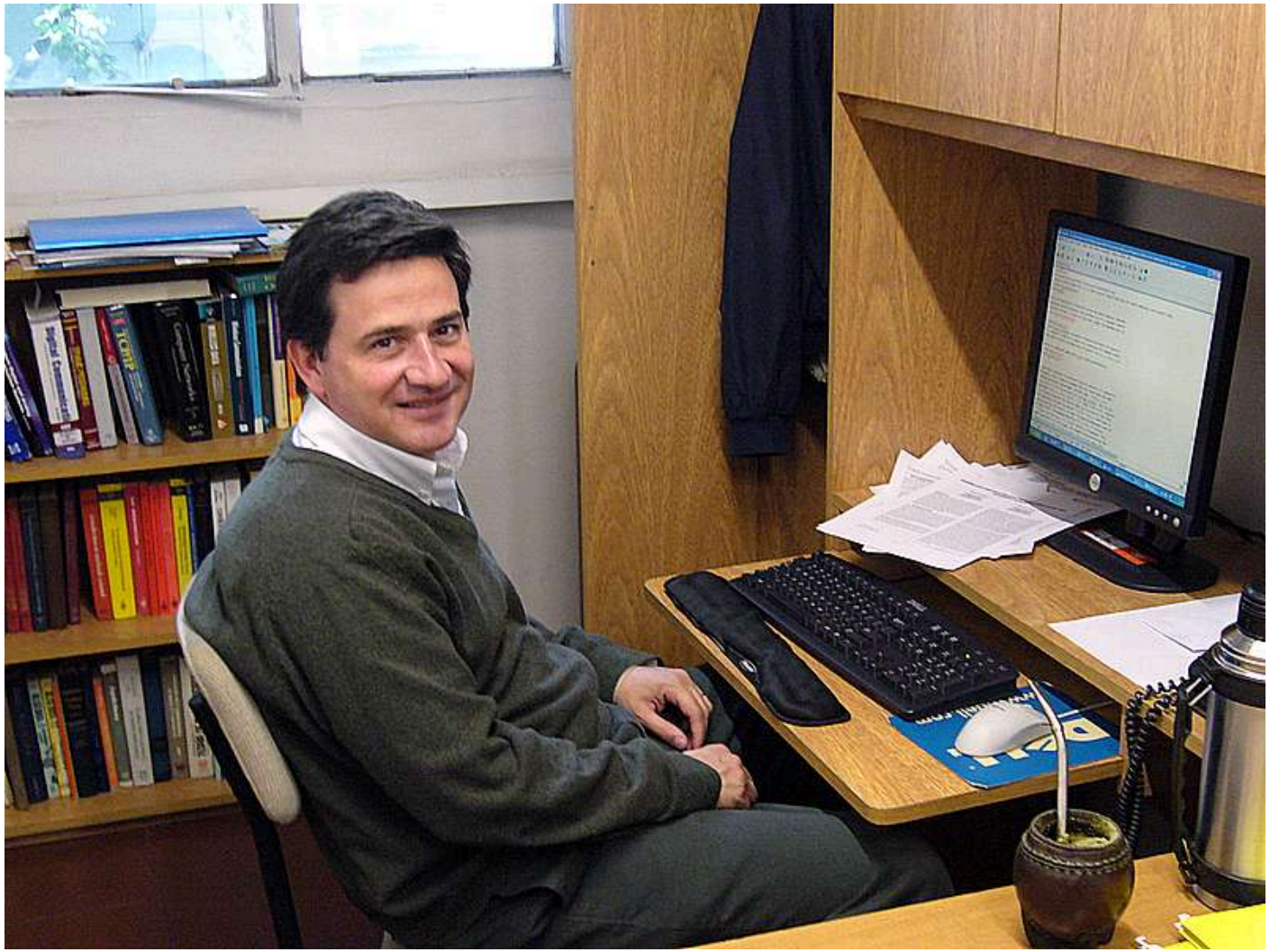
  and then the chain is stable (ergodic).

- **Problem:** file sizes are **not** exponential in real world scenarios.

Grupo MATE

Grupo MATE

Grupo MATE

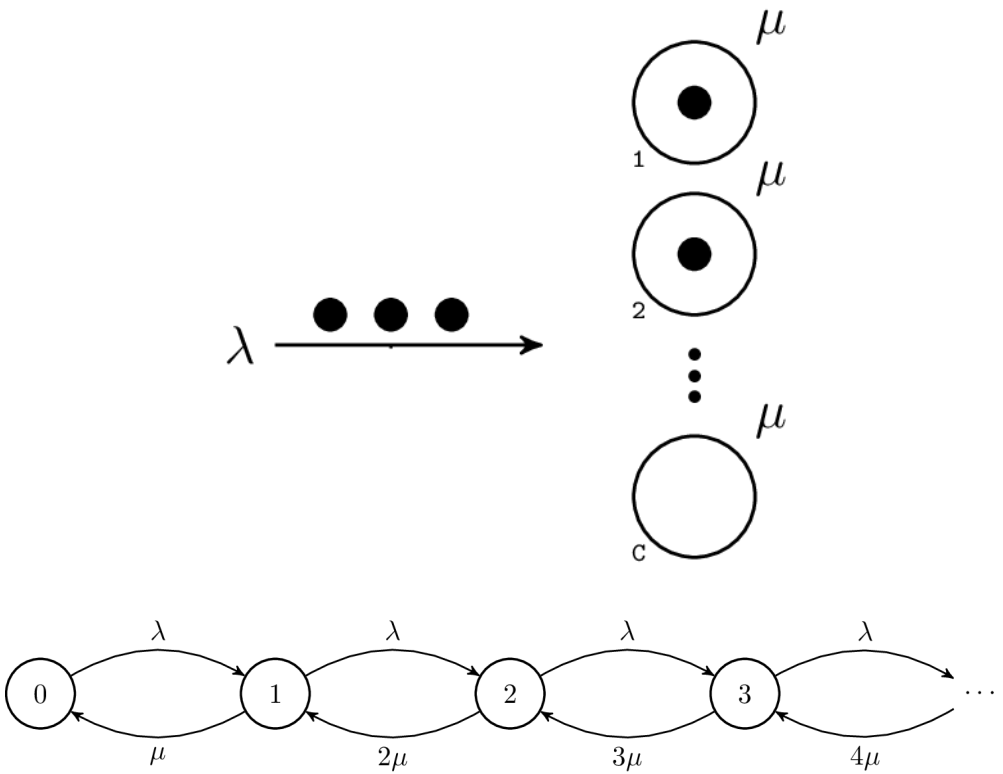


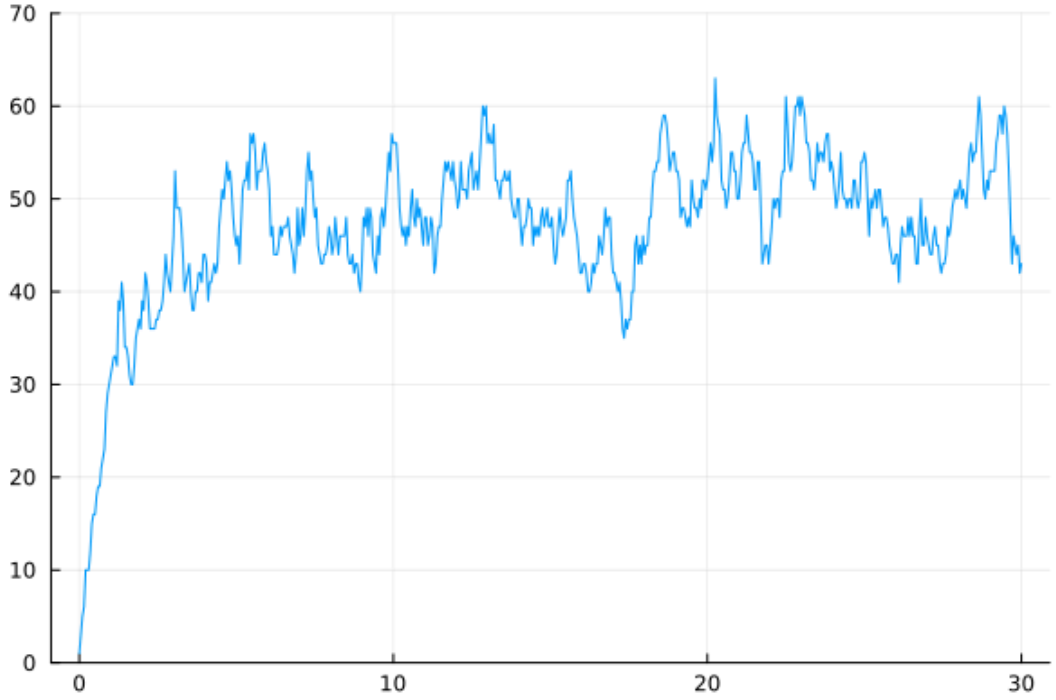
Queueing models

How normal people see queues:

Queueing models

How I see queues:

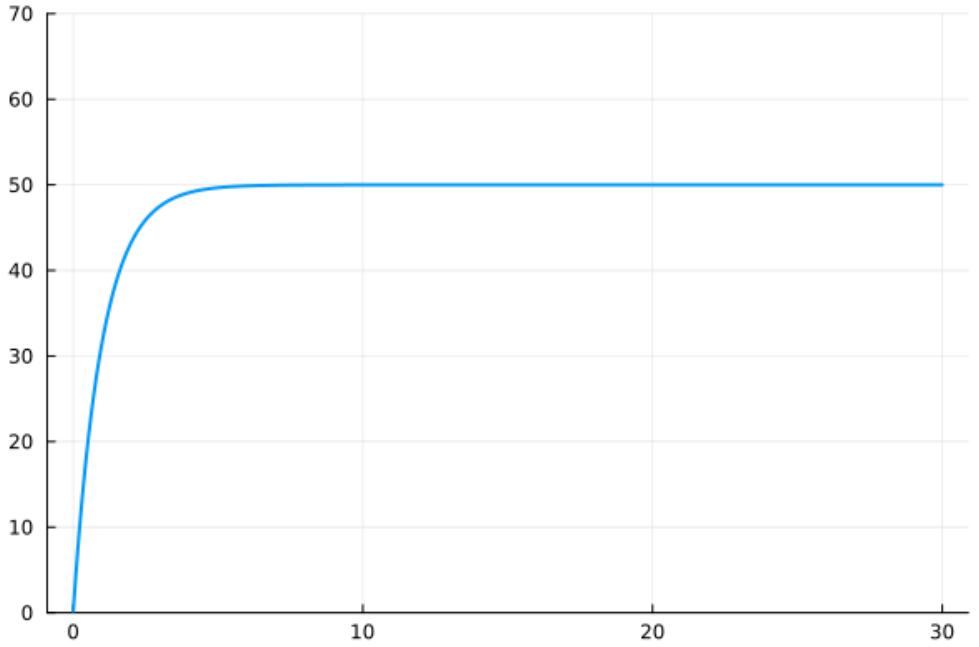




Queueing models

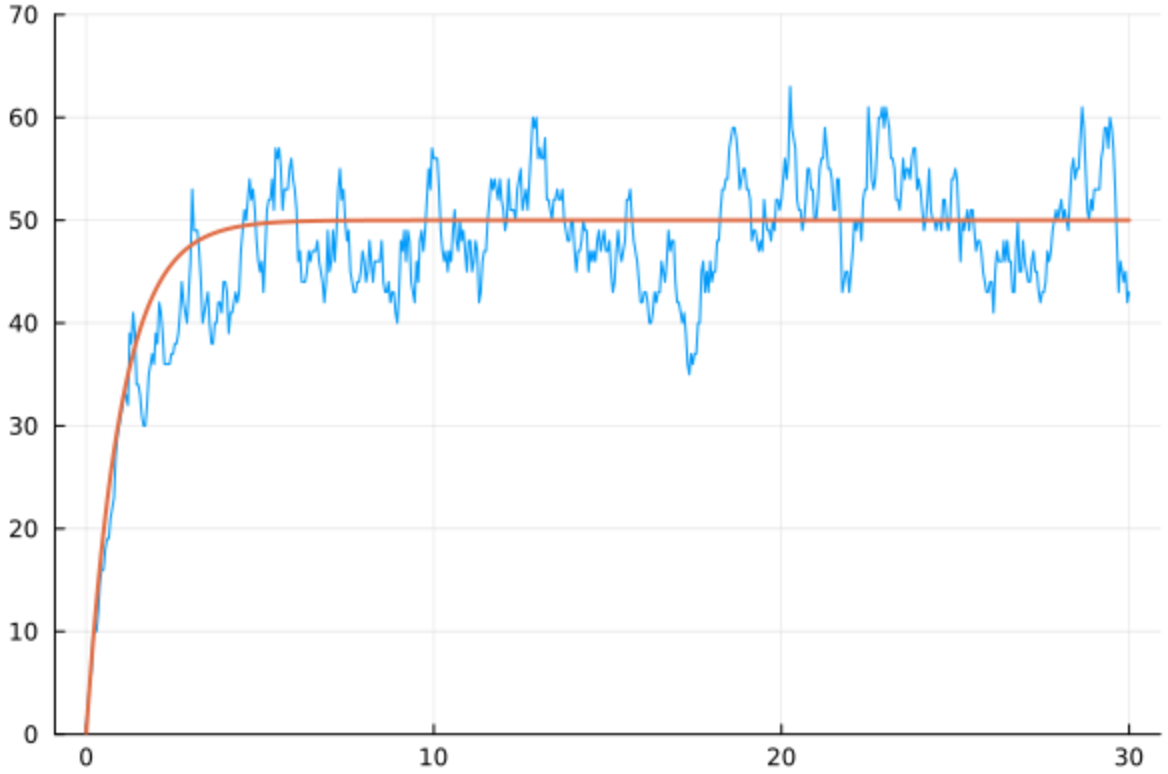
How **Fernando** sees queues:

$$\dot{x} = \lambda - \mu x$$



Queueing models

These are two viewpoints of the same thing!



- In fact there is a nice theory connecting both approaches [Kurtz et al.]

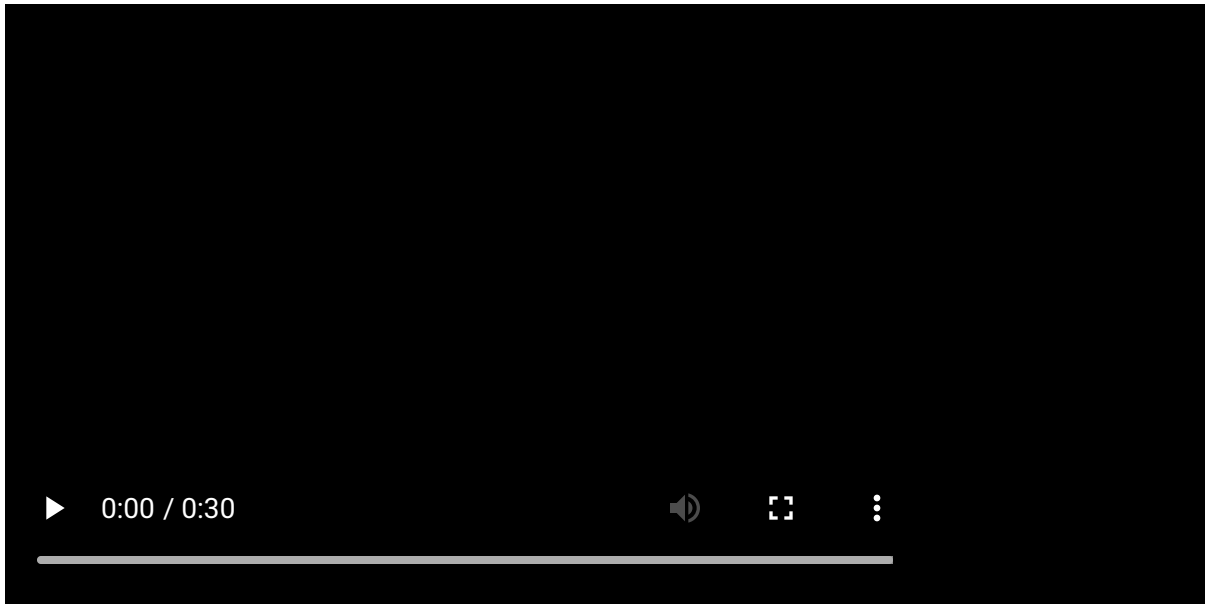
Back to Internet stability...

- We have a nice queueing model under *exponential* assumptions...
- ...but file sizes are not exponentially distributed.
- **Can we prove stability for general file sizes?**

The state space

Problem: you need a new state space which is *distributed* in nature…

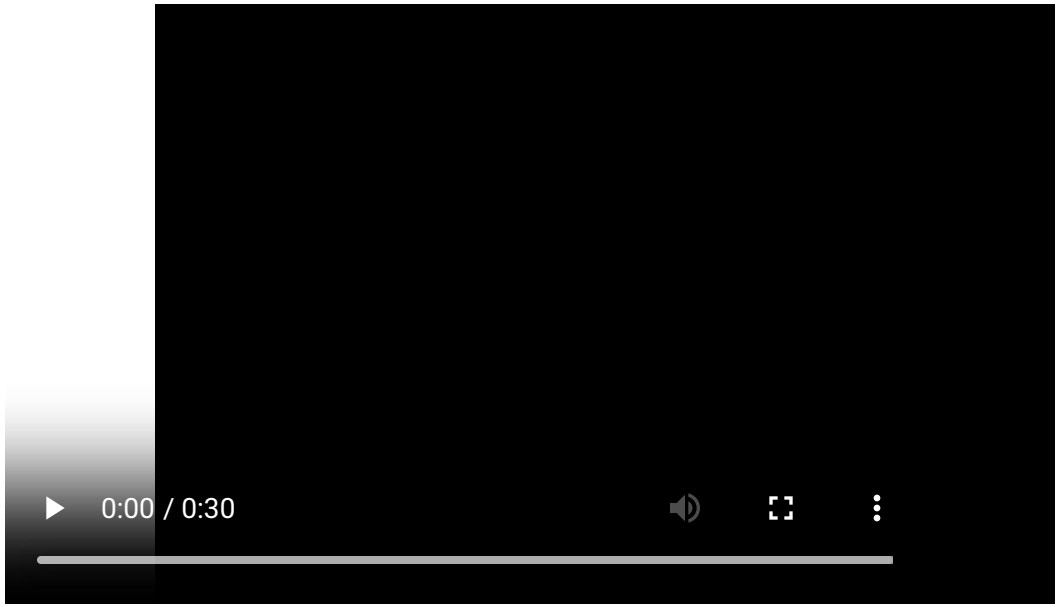
- Because you have to keep track of the *residual* services.



Fluid approximation

Idea: Replace the point masses by their empirical CCDF…

- …and then treat that like a fluid quantity…



Evolution equation

If $F_i(t, \sigma)$ is the CCDF of jobs present on route i, then its dynamics are:

$$\frac{\partial F_i}{\partial t} = \lambda_i G_i(\sigma) + \mathbf{x}^*(\mathbf{n(t)}) \frac{\partial F_i}{\partial \sigma} \quad i = 1, \ldots, N$$

- λ_i is the arrival rate on route i.
- $G_i(\sigma)$ is the CCDF of the job sizes on route i.
- $\mathbf{n}(t) = (n_i(t)) = (F_i(t, 0))$.
- $\mathbf{n} \mapsto \mathbf{x}^*$ is the congestion control mapping.

Lyapunov function

Let $V(F_1, \ldots, F_N)$ be given by:

$$V(F_1, \ldots, F_N) := \sum_{i=1}^{N} \frac{1}{\tilde{\rho}_i^{\alpha}} \int_0^{\infty} [F_i(t, \sigma)]^{\alpha+1} w_i(\sigma)\, d\sigma$$

- Here $\tilde{\rho}_i = \rho_i(1 + \delta)$.
- $w_i(\sigma)$ is a spatial weight function with $w_i(0) = 1$.

The Internet is stable!

> **Theorem** For the above fluid dynamics with congestion control feedback chosen from the $\alpha-$fair family, $U_i(x) := \frac{x^{1-\alpha}}{1-\alpha}$
> The function V is strictly decreasing along trajectories, provided that the natural stability condition $R\rho < c$ holds.

- So the internet is indeed stable!
- And moreover, the congestion control algorithms that we have been designing are **throughput optimal!** (i.e. they are able to stabilize the entire capacity region)

Network Stability under Alpha Fair Bandwidth Allocation with General File Size Distribution

Fernando Paganini*, Ao Tang†, Andrés Ferragut* and Lachlan L. H. Andrew‡
* Universidad ORT, Montevideo, Uruguay. † Cornell University, Ithaca, NY.
‡ Swinburne University of Technology, Australia.

***Abstract*—Rate allocation among a fixed set of end-to-end connections in the Internet is carried out by congestion control, which has a well established model: it optimizes a concave network utility, a particular case of which is the alpha-fair bandwidth allocation. This paper studies the slower dynamics of connections themselves, that arrive randomly in the network and are served at the allocated rate. It has been shown that under the condition that the mean offered load at each link is less than its capacity, the resulting queueing system is stochastically stable, for the case of exponentially distributed file-sizes. The conjecture that the result holds for general file-size distributions has remained open, and is very relevant since heavy-tailed distributions are often the best models of Internet file sizes.**

In this paper, building on existing fluid models of the system, we use a partial differential equation to characterize the dynamics. The equation keeps track of residual file size and therefore is suitable for general file size distributions. For alpha fair bandwidth allocation, with any positive alpha parameter, a Lyapunov function is constructed with negative drift when the offered load is less than capacity. With this tool we answer the conjecture affirmatively in the fluid sense: we prove asymptotic convergence to zero of the fluid model for general file-size distributions of finite mean, and finite-time convergence for those of finite $p > 1$ moment. In the stochastic sense, we build on recent work that relates fluid and stochastic stability subject to a certain light-tailed restriction. We further provide the supplementary fluid stability argument to establish the conjecture for this class that includes phase-type distributions. Results are supplemented by illustrative network simulations at the packet level.

rates depend on bandwidth allocation, assumed to occur at a faster time-scale. This leads to a basic *stability* question, first posed by De Veciana et al. [7]: under which connection level demands (job arrival rate and mean workload) is the resulting queueing process stable? The answers given in [7] apply to Poisson arrivals and exponentially distributed job sizes, and max-min fair or proportionally fair bandwidth allocation. In this case the numbers of connections per route form a Markov chain, which is shown to be stable (i.e., ergodic) under the natural stability condition: namely, that the *mean load in each link of the network is strictly less than the link capacity.* In a subsequent paper by Bonald and Massoulié [3], these results were generalized to the α-fair case; other utility functions are considered in [26]. Further extensions include relaxing the time-scale separation [14] and relaxing the model of fixed-capacity links [15].

We note that the natural condition is not sufficient for all allocation policies, such as when the network seeks to maximize instantaneous throughput ($\alpha = 0$, see [3]) or under certain forms of prioritization. For a demonstrative example we refer to Section VI. Verloop et al. [24] show that the form of scheduling that is optimal for a single link (shortest remaining processing time, SRPT) can cause networks to be unstable even if the maximum individual link utilization approaches 0. The intuition is that a flow on a multi-link path

The beginning of a friendship...

The beginning of a friendship...

The beginning of a friendship...

The beginning of a friendship...

The beginning of a friendship...

And a fruitful one...

And still going...

And still going...

Dynamics and Optimization in Spatially Distributed Electrical Vehicle Charging

Fernando Paganini, *Fellow, IEEE*

***Abstract*—We consider a spatially distributed demand for electrical vehicle recharging, that must be covered by a fixed set of charging stations. Arriving EVs receive feedback on transport times to each station, and waiting times at congested ones, based on which they make a selfish selection. This selection determines total arrival rates in station queues, which are represented by a fluid state; departure rates are modeled under the assumption that clients have a given sojourn time in the system. The resulting differential equation system is analyzed with tools of optimization. We characterize the equilibrium as the solution to a specific convex program, which has connections to optimal transport problems, and also with road traffic theory. In particular a price of anarchy appears with respect to a social planner's allocation. From a dynamical perspective, global convergence to equilibrium is established, with tools of Lagrange duality and Lyapunov theory. An extension of the model that makes customer demand elastic to observed delays is also presented, and analyzed with extensions of the optimization machinery. Simulations to illustrate the global behavior are presented, which also help validate the model beyond the fluid approximation.**

***Index Terms*—Electrical vehicle charging, optimization, transportation networks, distributed algorithms/control.**

This solution, while not centrally planned, can nevertheless be characterized in terms of a suitable optimization problem, which has been helpful to understand the *Price of Anarchy*, i.e. the gap between this equilibrium and the social welfare optimum [18], and to propose means (e.g. tolls) to mitigate it. While much of this classical analysis concerns equilibrium, *dynamic* studies of road traffic networks are also extensive, see e.g. [5] and references therein.

In this paper we consider a new application area, the operation of an Electrical Vehicle (EV) charging infrastructure. In particular, we are interested in public facilities situated in parking lots, where EV chargers are made available for temporary use [13]. This development has motivated an active area of research, within which we distinguish different problems: (i) the operation of a *single* facility of this kind, in particular the scheduling of charging opportunities taking into account EV deadlines and installation limitations [10], [25]; (ii) integration of EV charging to the smart grid [14], [23]; (iii) facility location problems, i.e. where to deploy EV charging [9], [15].

And still going...

Dynamics and Optimization in Spatially Distributed Electrical Vehicle Charging

Fernando Paganini, *Fellow, IEEE,* and Andres Ferragut

***Abstract*—We consider a spatially distributed demand for electrical vehicle recharging, that must be covered by a fixed set of charging stations. Arriving EVs receive feedback on transport times to each station, and waiting times at congested ones, based on which they make a selfish selection. This selection determines total arrival rates in station queues, which are represented by a fluid state; departure rates are modeled under the assumption that clients have a given sojourn time in the system. The resulting differential equation system is analyzed with tools of optimization. We characterize the equilibrium as the solution to a specific convex program, which has connections to optimal transport problems, and also with road traffic theory. In particular a price of anarchy appears with respect to a social planner's allocation. From a dynamical perspective, global convergence to equilibrium is established, with tools of Lagrange duality and Lyapunov theory. An extension of the model that makes customer demand elastic to observed delays is also presented, and analyzed with extensions of the optimization machinery. Simulations to illustrate the global behavior are presented, which also help validate the model beyond the fluid approximation.**

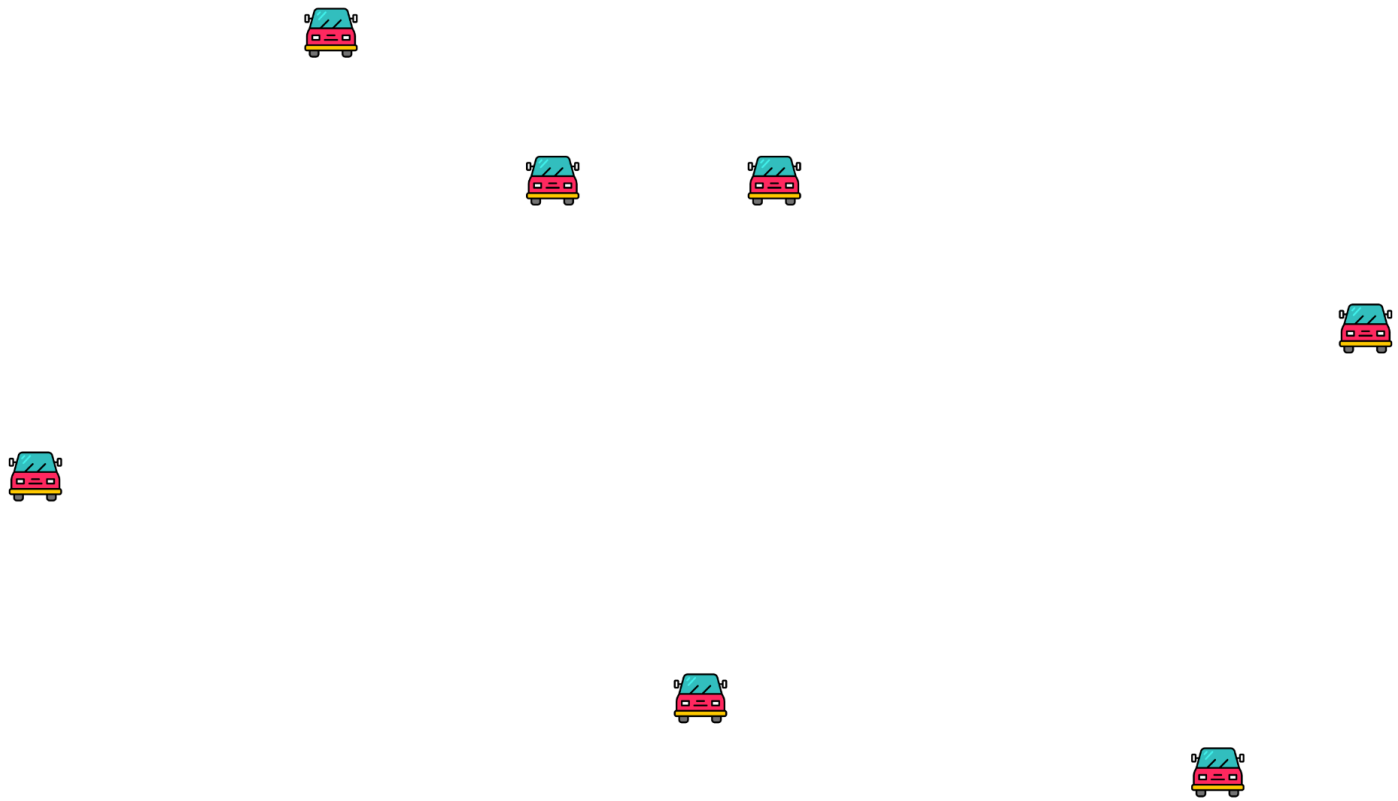
***Index Terms*—Electrical vehicle charging, optimization, transportation networks, distributed algorithms/control.**

This solution, while not centrally planned, can nevertheless be characterized in terms of a suitable optimization problem, which has been helpful to understand the *Price of Anarchy*, i.e. the gap between this equilibrium and the social welfare optimum [18], and to propose means (e.g. tolls) to mitigate it. While much of this classical analysis concerns equilibrium, *dynamic* studies of road traffic networks are also extensive, see e.g. [5] and references therein.

In this paper we consider a new application area, the operation of an Electrical Vehicle (EV) charging infrastructure. In particular, we are interested in public facilities situated in parking lots, where EV chargers are made available for temporary use [13]. This development has motivated an active area of research, within which we distinguish different problems: (i) the operation of a *single* facility of this kind, in particular the scheduling of charging opportunities taking into account EV deadlines and installation limitations [10], [25]; (ii) integration of EV charging to the smart grid [14], [23]; (iii) facility location problems, i.e. where to deploy EV charging [9], [15].

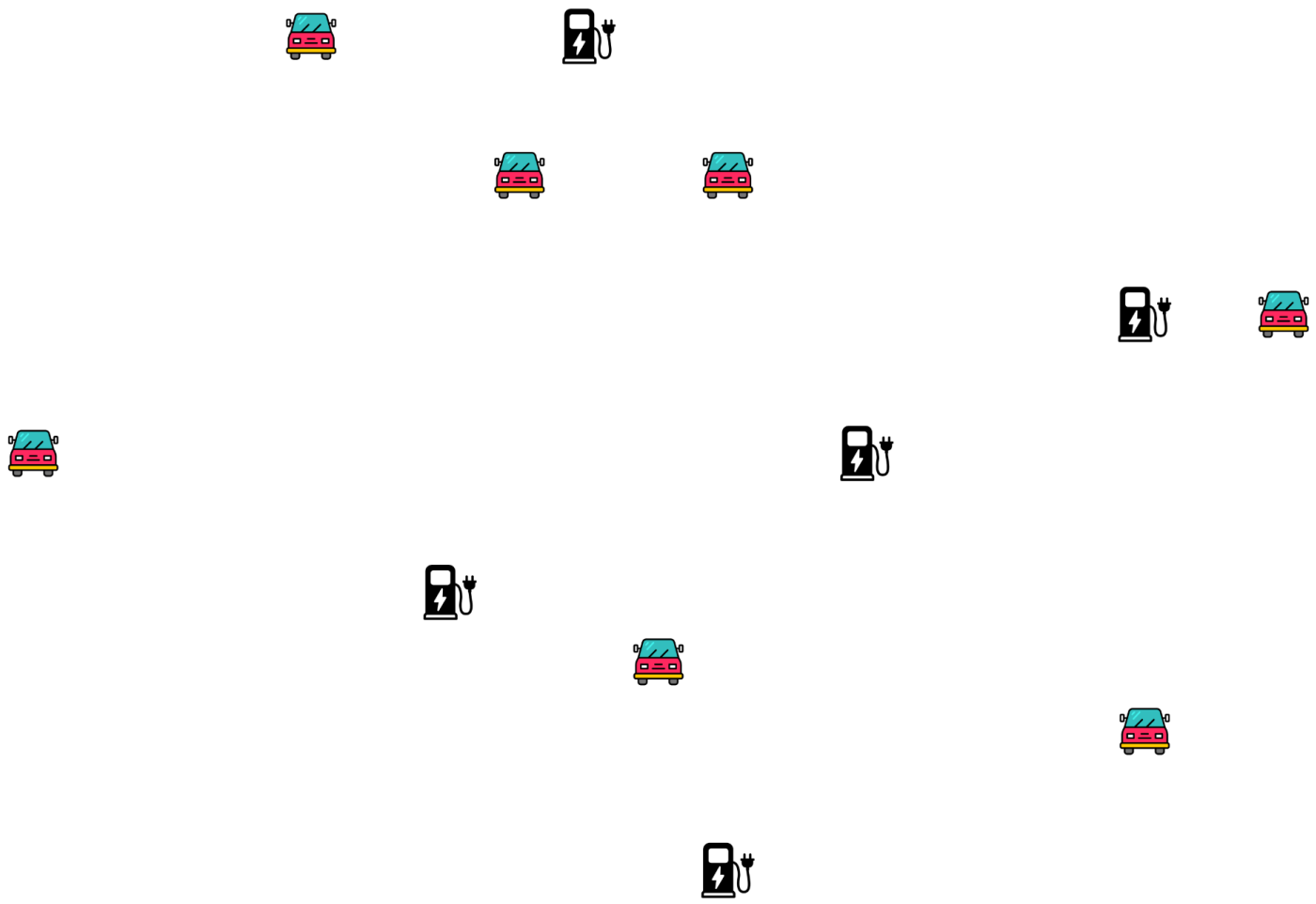
Spatial load balancing

Spatial load balancing



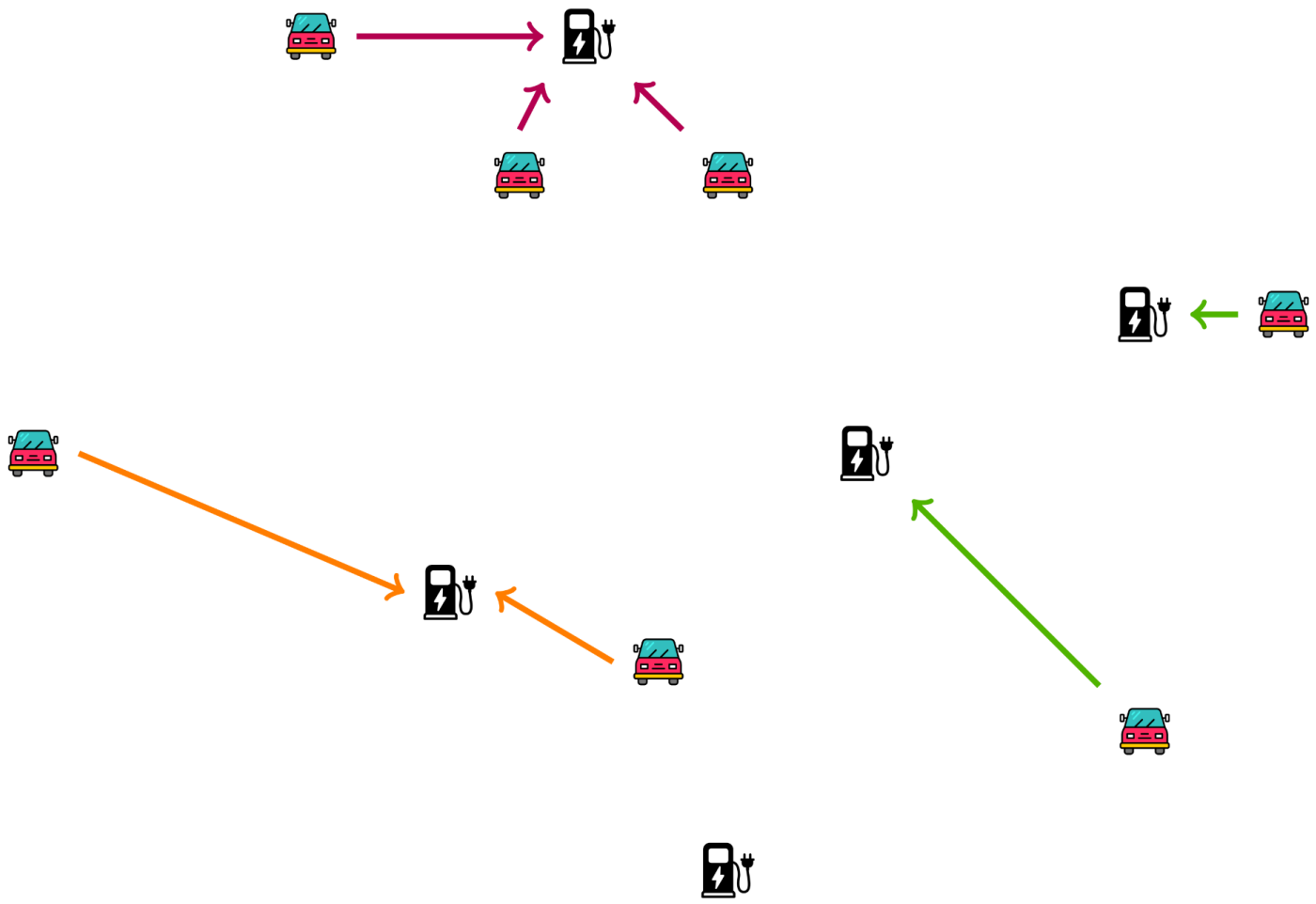
Spatially distributed requests

Spatial load balancing



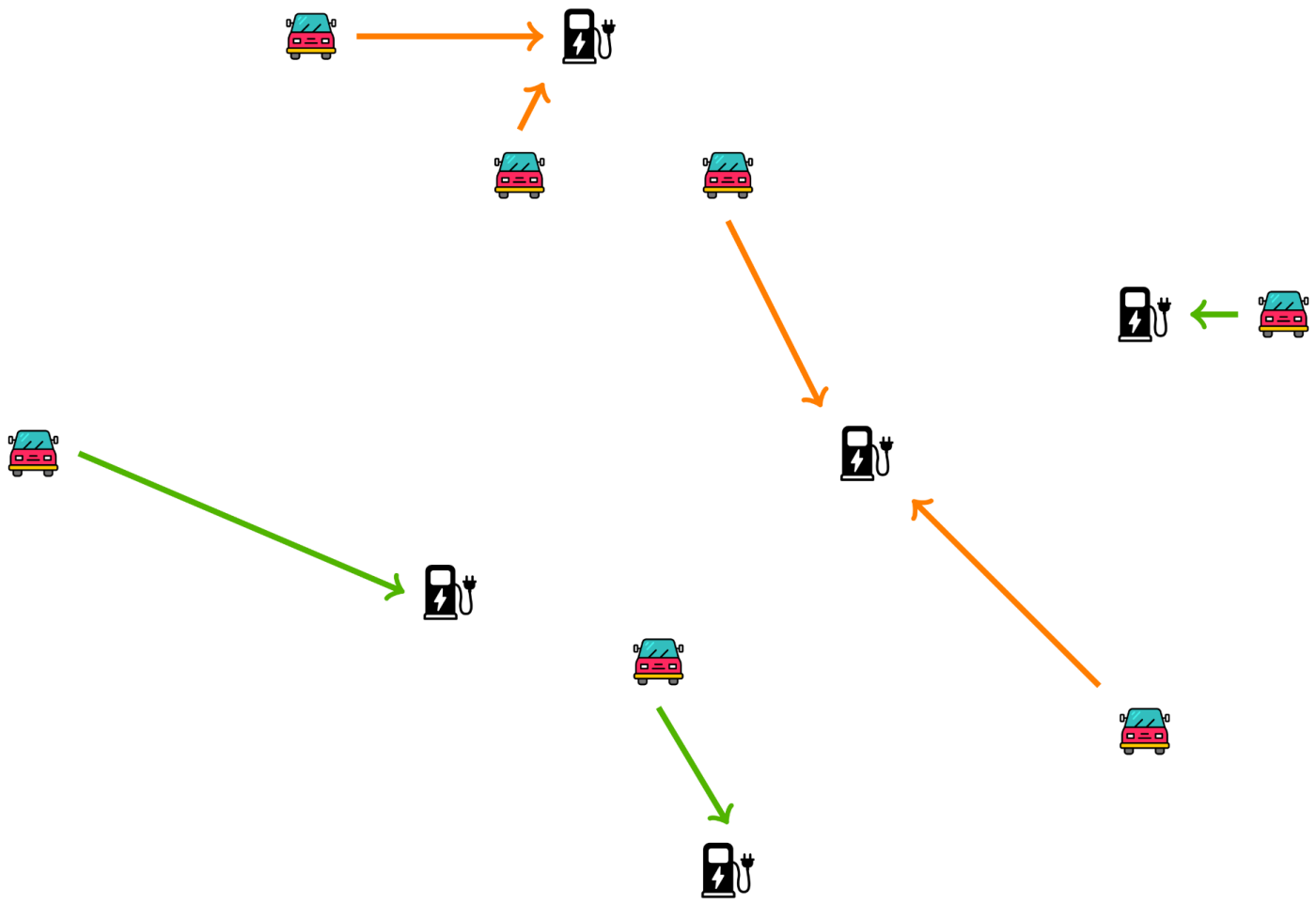
Spatially distributed servers

Spatial load balancing



Minimum distance assignment, does it make sense?

Spatial load balancing



Maybe it’s better to balance more evenly...

Model

- Assume a *finite* (but arbitrarily large) number of arrival locations $i = 1, \ldots, n$.
- Each location gets requests at rate r_i.
- Stations are at locations $j = 1, \ldots, m$, and have capacity c_j.
- You have a *travel cost* matrix:

  $$K = (\kappa_{ij})$$

  The **time** it takes demands to get from location i to location j

Optimization formulation

$$\min_{x_{ij} \geqslant 0} \sum_{i,j} \kappa_{ij} x_{ij},$$

subject to:

$$\sum_j x_{ij} = r_i \quad i = 1, \ldots, n,$$

$$\sum_i x_{ij} \leqslant c_j \quad j = 1, \ldots, m.$$

- $\delta_{ij} := x_{ij}/r_i$ is the *fraction* of traffic sent from location i to station j.
- So it's a variant of the Monge-Kantorovich mass transport problem, where demands are fixed and you have a destination capacity constraint.

Accounting for users' choices

You can solve this as a "central planner", but users may decentralize their own choices.

- Assume q_j represents the **queue** size at station j.
- Define $\mu_j(q_j)$ as the **queueing delay** at station j.
- Then *greedy* users will choose, **in feedback**, the station that minimizes:

$$j^* = \arg\min_i \{\kappa_{ij} + \mu_j(q_j), j = 1, \dots, m\}$$

- This complicates things due to the switching nature of the dynamics.
  - So simplify this by using a *soft-min* function.

Fluid model of the queueing dynamics

Using the *bucket analogy* again, we get these dynamics for the entire system:

$$\begin{cases} \dot{q}_j = \sum_{i=1}^{M} x_{ij} - \frac{1}{T} q_j & j = 1, \dots, m \\ \mu_j(q_j) = T \left[1 - \frac{c_j}{q_j} \right]^+ & j = 1, \dots, m \\ x_{ij} = r_i \delta_{ij} & i = 1, \dots, n, \ j = 1, \dots, m \end{cases}$$

- Again r_i and c_j are the network parameters.
- T represents the *average sojourn time* the customers are willing to stay.
- δ_{ij} is the soft-min function:

$$\delta_{ij}(K, \mu) = \frac{e^{-\frac{1}{\varepsilon}(\kappa_{ij} + \mu_j)}}{\sum_k e^{-\frac{1}{\varepsilon}(\kappa_{ik} + \mu_j)}}$$

Optimization formulation

$$\min_{x_{ij} \geqslant 0} \sum_{i,j} \kappa_{ij} x_{ij} + \sum_{j} \beta_j(q_j) + \varepsilon \sum_{i} \mathcal{H}(\delta^i)$$

subject to:

$$\sum_{j} x_{ij} = r_i, \ \forall i \quad \sum_{i} x_{ij} = \frac{q_j}{T} \ \forall j$$

- Here β_j is a *penalty function* accounting for queueing delay:

$$\beta_j(q_j) = \frac{1}{T} \int_0^{q_j} \mu_j(\sigma) d\sigma.$$

- $\mathcal{H}$ is the entropy function of the choices, that acts as a regularization term yielding the soft-max.

Interpretation

- The above Problem is a *primal relaxation* of the earlier Monge-Kantorovich optimal mass transport…

- …that serves as a suitable model for the system when users have greedy choices in terms of **time**.

Main result

> **Theorem:** The greedy dynamics are globally asymptotically stable, and its equilibrium point is the solution of the relaxed approximation of the Monge-Kantorovich relaxed problem.

Main result

> **Theorem:** The greedy dynamics are globally asymptotically stable, and its equilibrium point is the solution of the relaxed approximation of the Monge-Kantorovich relaxed problem.

> **Proof idea:** Use Lagrangian decomposition, and take the dual function of the optimization problem as your Lyapunov function.

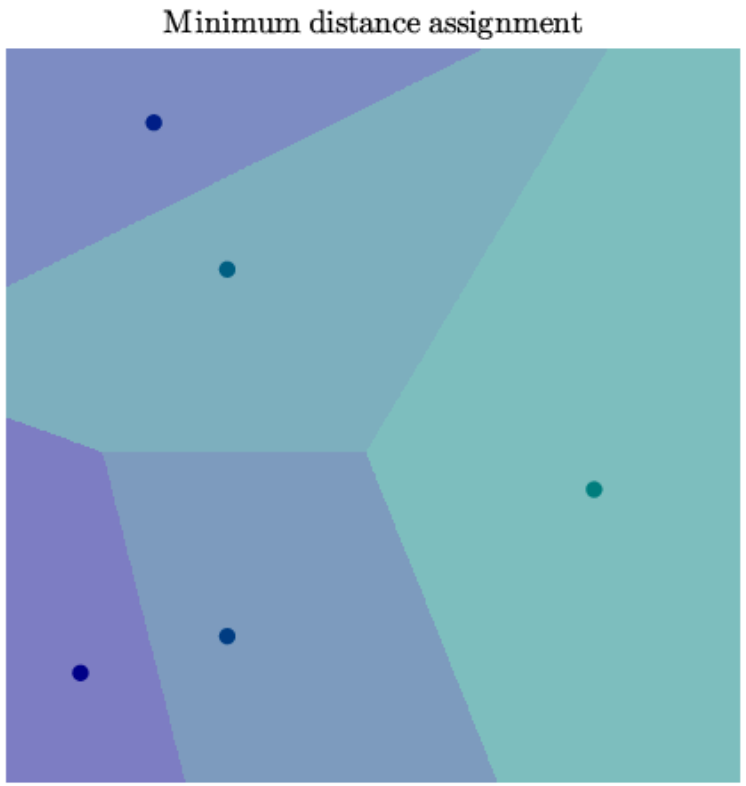
Main result

> **Theorem:** The greedy dynamics are globally asymptotically stable, and its equilibrium point is the solution of the relaxed approximation of the Monge-Kantorovich relaxed problem.

> **Proof idea:** Use Lagrangian decomposition, and take the dual function of the optimization problem as your Lyapunov function.

Of course, devil is in the details, but Fernando has a pact with the devil…

Simulations



Simulations

Simulations

Final remarks

> “If we knew what it was we were doing, it would not had been called research, would it?”
> — Albert Einstein

Final remarks

> “If we knew what it was we were doing, it would not had been called research, would it?”
> — Albert Einstein

> “At any moment there is only a fine layer between the 'trivial' and the impossible. Mathematical discoveries are made in this layer.”
> — A.N. Kolmogorov