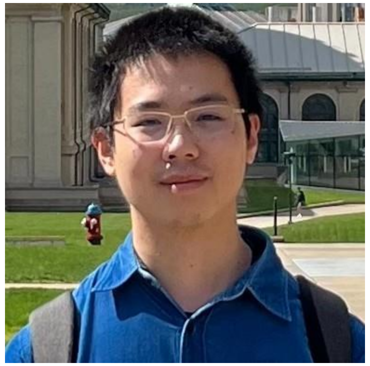
Analyzing Queues with Markovian Arrivals and Markovian Service

Isaac Grosof (Izzy) (UIUC -> Northwestern IEMS)

Designed with Mor Harchol-Balter (CMU)

Collaborators: Thank You!

Mor Harchol-B

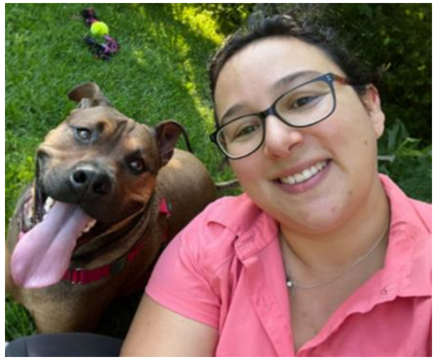


Yige Hong

Alan Scheller-Wolf

R. Srikant

Siva Theja Maguluri



Daniela Hurtado-Lange

Cameron Curtis

Seyed Irvani

Ziyuan Wang

Hayriye Ayhan

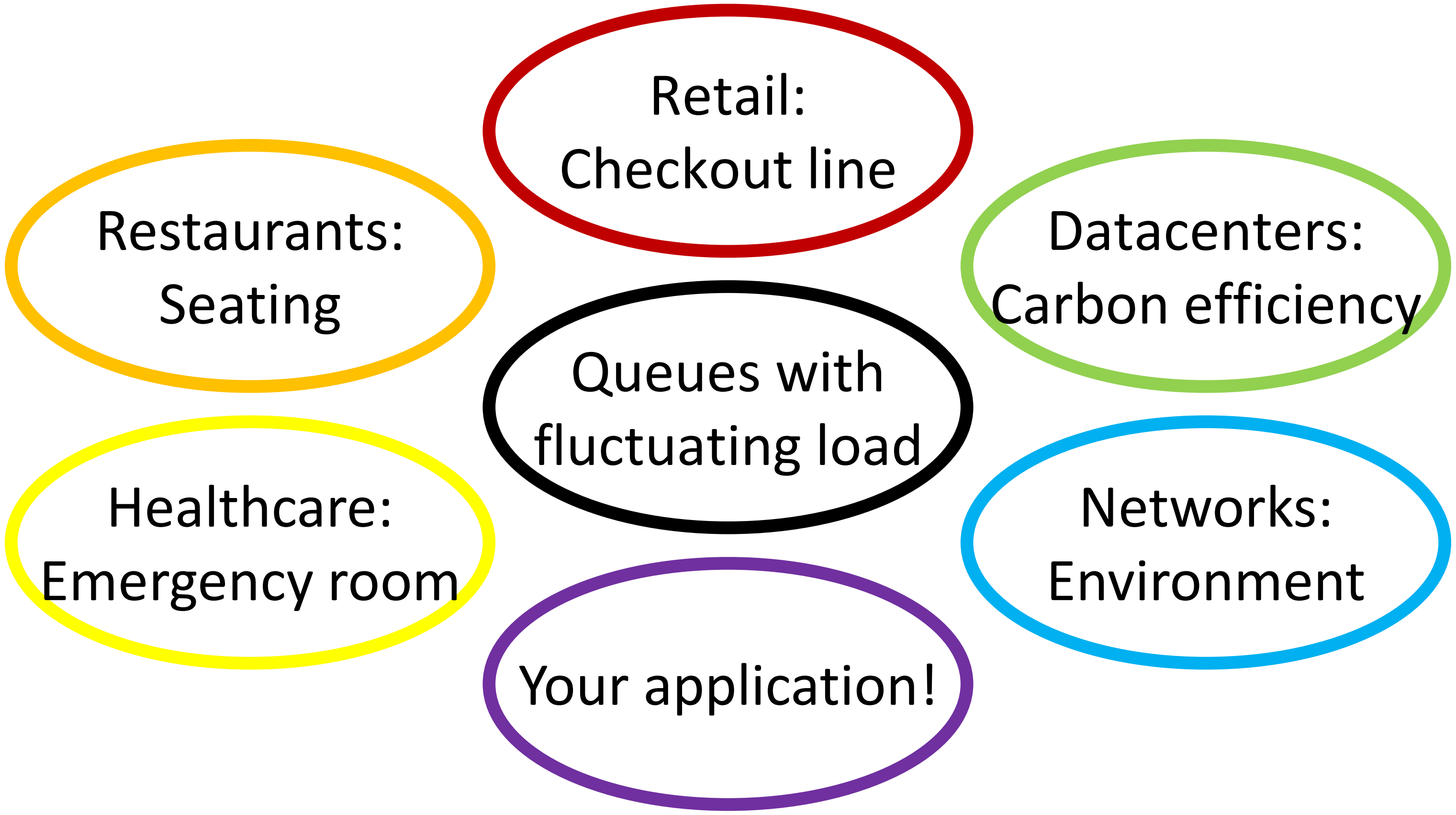
Ben Berg

Zhongrui Chen

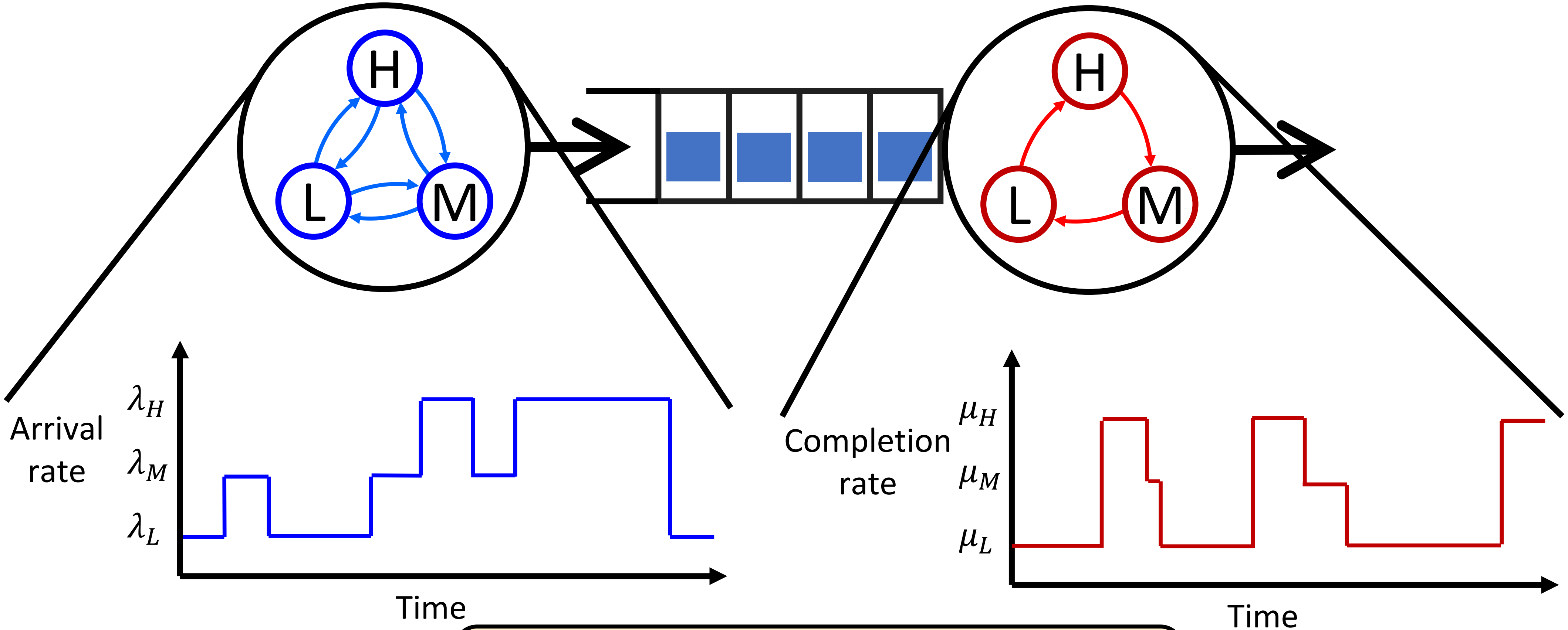
Fluctuating Load

Vast majority of queueing theory: i.i.d. arrivals, i.i.d. service, fixed load.

Reality: correlated arrivals, correlated service, fluctuating load.



Model: Markovian Arrivals & Markovian Service



Goal: Simple, explicit characterization of mean queue length, $E[Q]$

Outline

MAMS Model

Simple MAMS: 2-level Arrivals

Drift Method

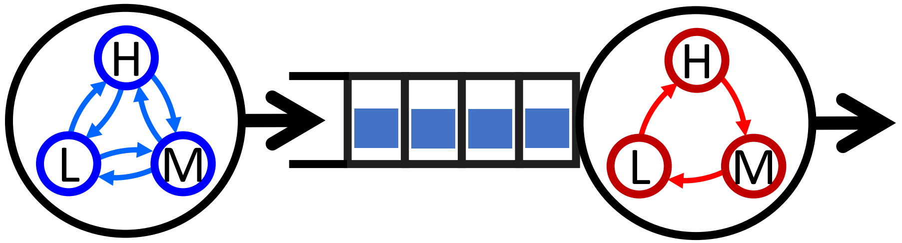
Relative Arrivals

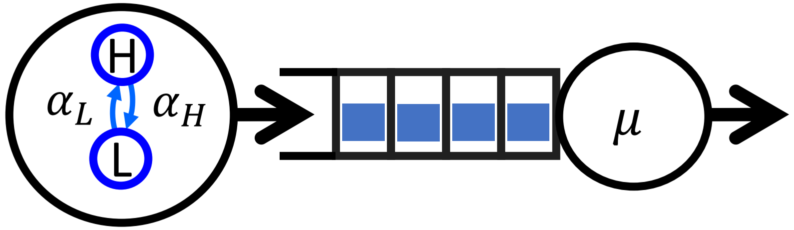
2-level results

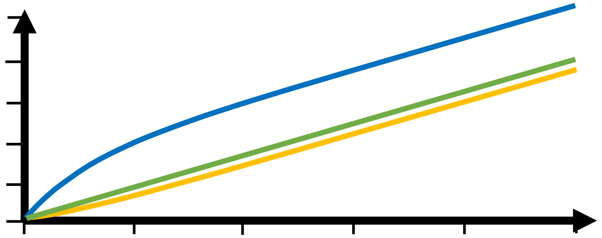
⟵ ⟵ ⟵ Break time ⟶ ⟶ ⟶

Generalizing to Full MAMS

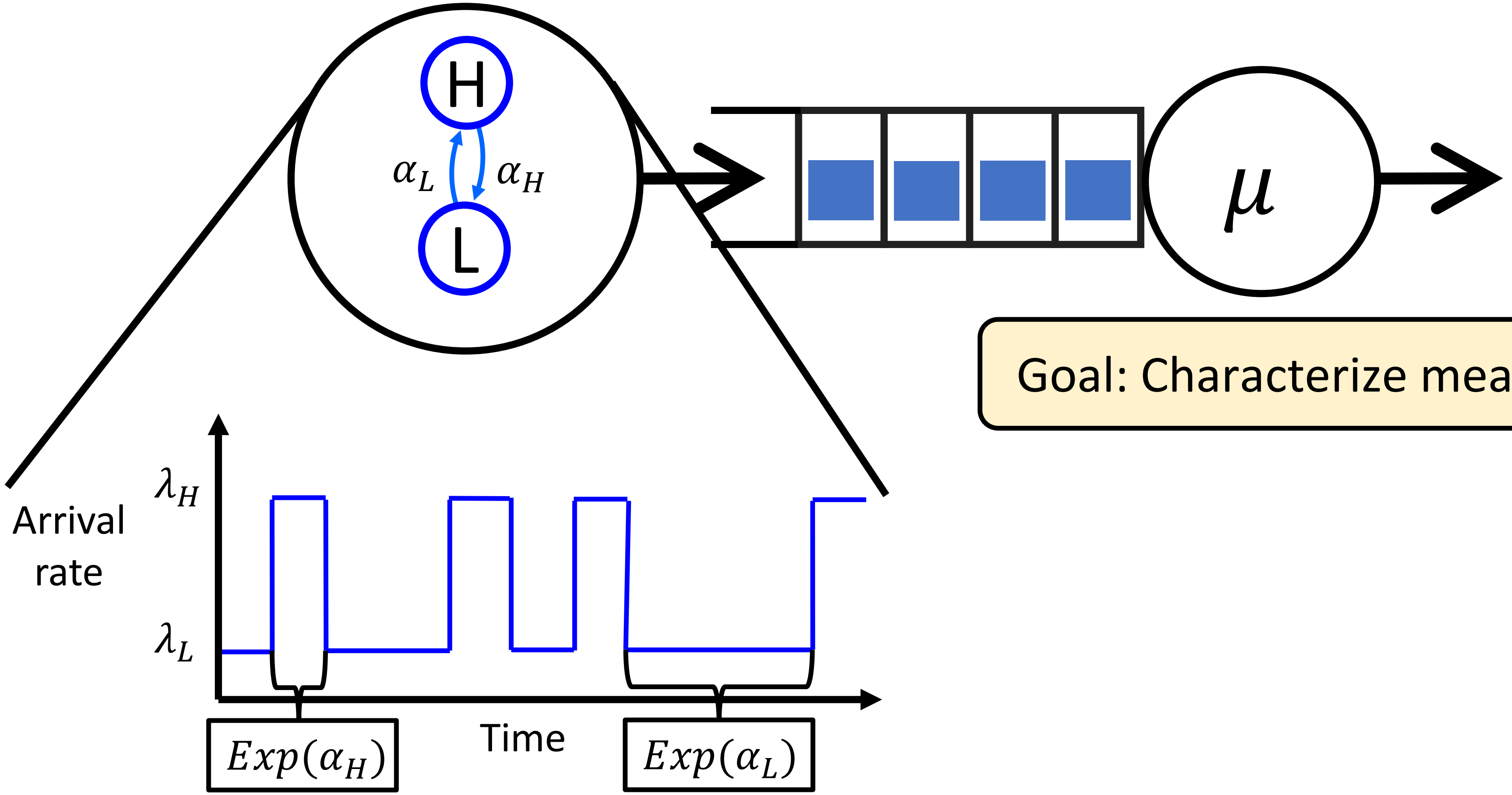
Applications!







Simple MAMS: 2-level arrivals



Goal: Characterize mean queue length, $E[Q]$

Prior work on Markov-modulated arrivals

Computational Methods

Generating functions
[Yechiali & Naor ‘71],
[Gupta et al. ‘06]

Matrix analytic methods
[Neuts ‘78], [Ramaswami ‘80],
[Latouche & Ramaswami ‘99], …

Symbolic Results

m-step drift
Heavy-traffic, semi-closed form
[Mou & Maguluri ’20]

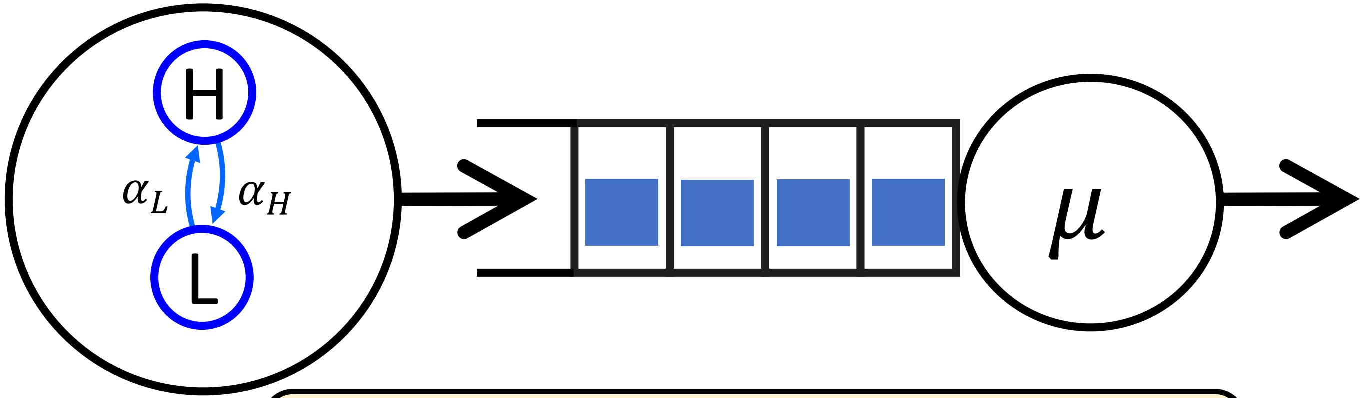
Structural, monotonicity,
convexity results [Gupta et al. ‘06],
[Vesilo, Harchol-Balter, Scheller-Wolf’21]

Simple formula for $E[Q]$

No prior results

Today:
Relative arrivals + Drift [GHH’24]
Results in half of a tutorial!

Simple MAMS: 2-level arrivals



Goal: Simple, explicit characterization of mean queue length, $E[Q]$

Q1: $P(H)$? $P(L)$?

Q2: Arrival rate λ?

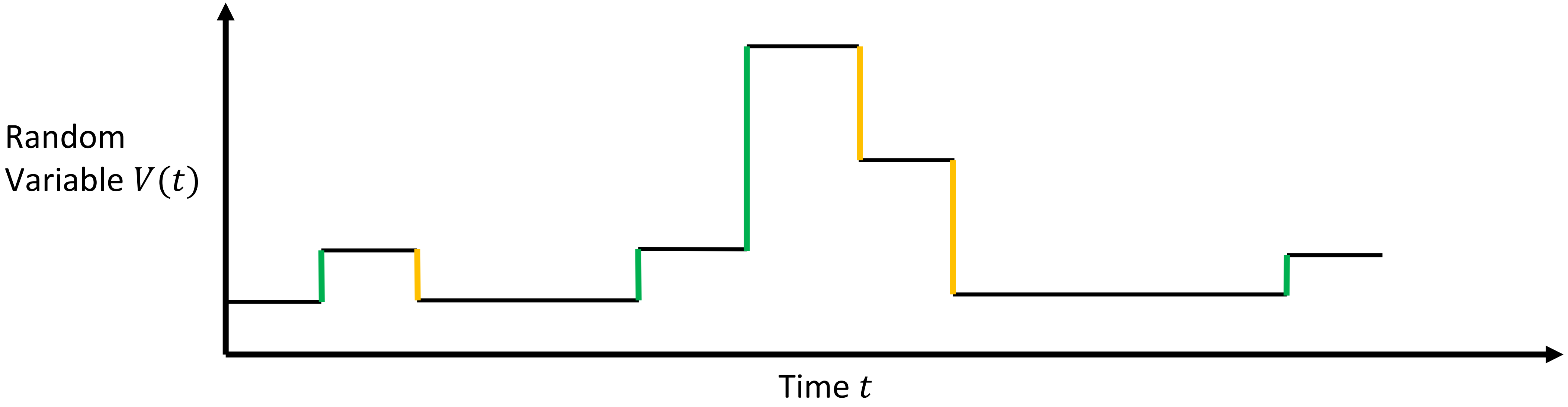
A1: $P(H) = \frac{\alpha_L}{\alpha_L+\alpha_H}, P(L) = \frac{\alpha_H}{\alpha_L+\alpha_H}$

A2: $\frac{\lambda_H\alpha_L+\lambda_L\alpha_H}{\alpha_L+\alpha_H}$

Drift Method: Background

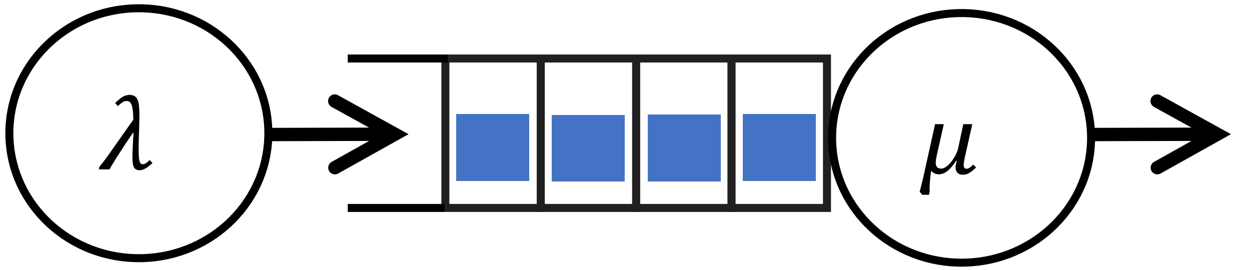
Choose any random variable V.

Idea: Increases match decreases

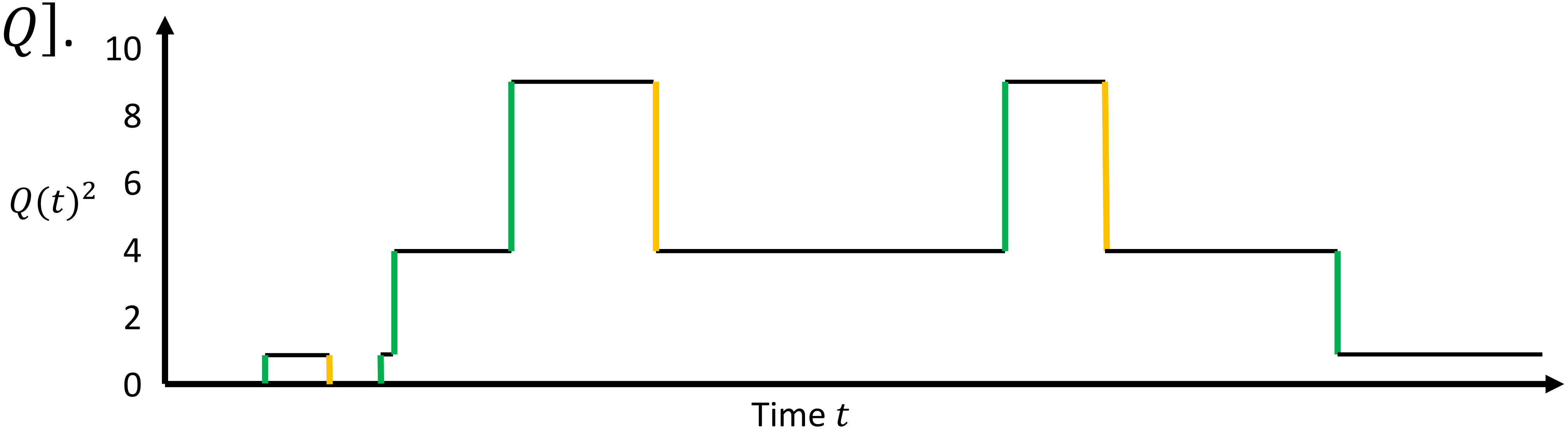


Thm: In stationarity, expected increase matches expected decrease

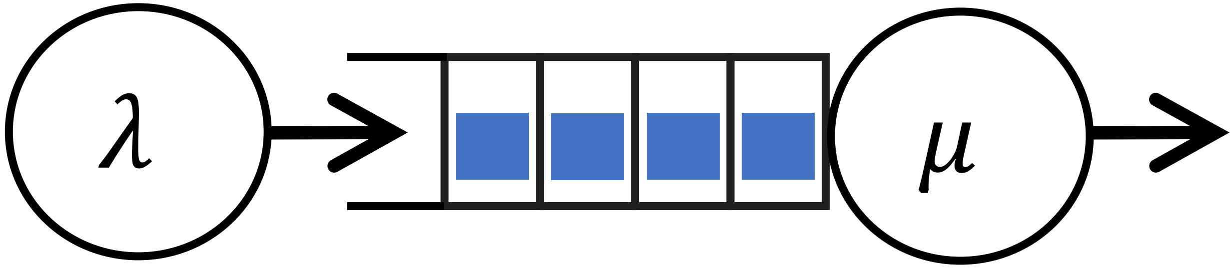
Drift Method: M/M/1

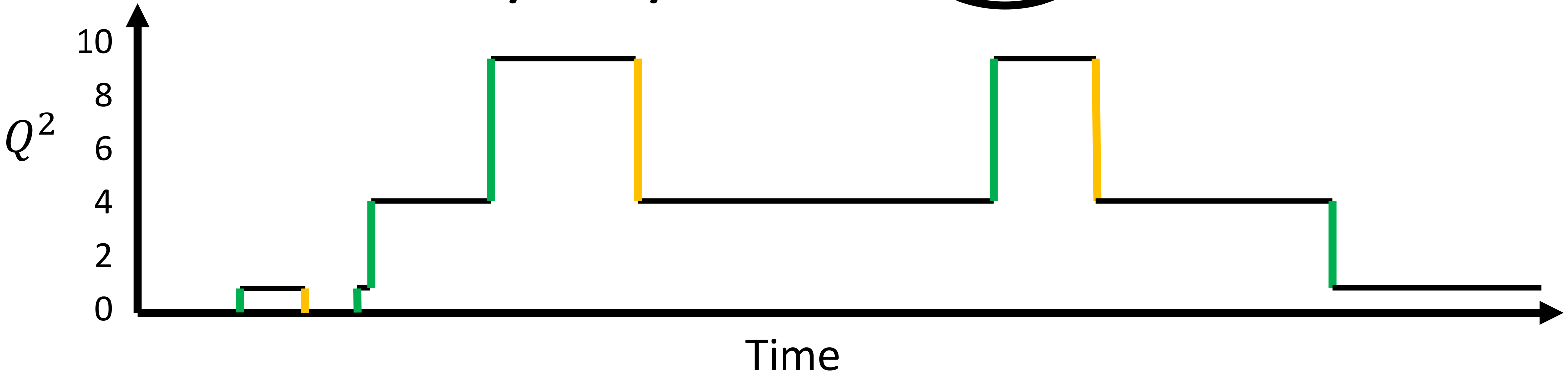


Example: Random variable Q^2. Let's use drift method to calculate $E[Q]$.



Drift Method: M/M/1





Suppose $Q(t) = q$. What rates of change of Q^2? What amounts of change?

Increases: Arrivals

Q3: Rate of arrivals? Change in Q^2?

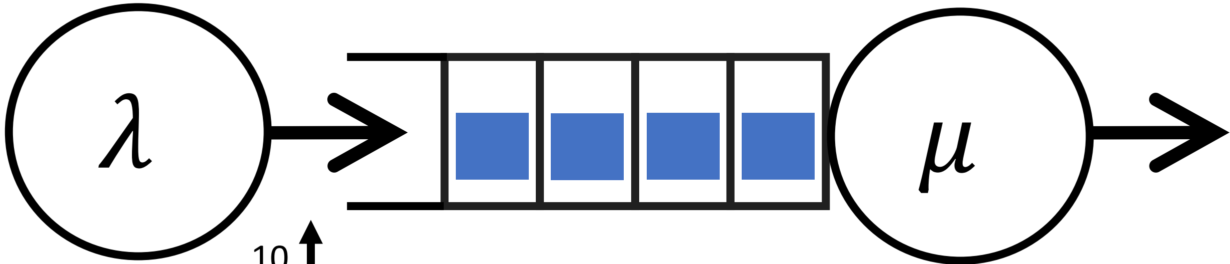
A3: Rate λ, change $2q + 1$

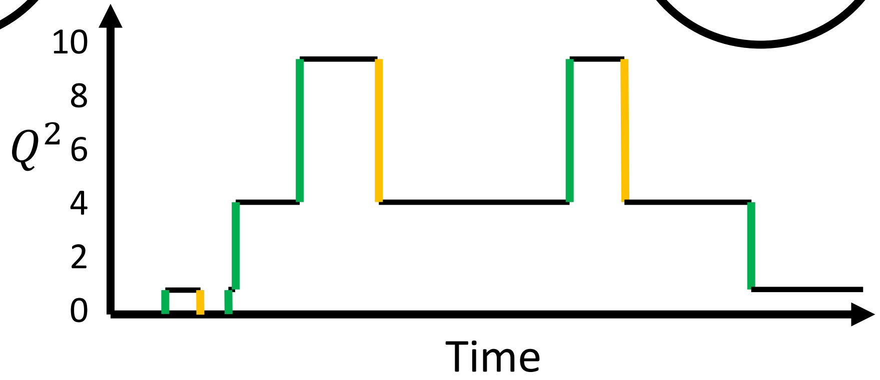
Decreases: Completions

Q4: Rate of completions? Change in Q^2?

A4: Rate μ, change $-2q + 1$, 0 except if $q = 0$

Drift Method: M/M/1





A3: Arrivals: Rate λ, change $2q + 1$

A4: Completions: Rate μ, change $-2q + 1$, except if $q = 0$

Q5: Expected increase, in stationarity?
Expected decrease, in stationarity?

A5: Increase: $\lambda(2E[Q] + 1)$,
Decrease: $\mu(-2E[Q] + P(Q > 0))$
$= -2\mu E[Q] + \lambda$

Q6: Sum to 0, solve for $E[Q]$.

A6: $\lambda(2E[Q] + 1) - 2\mu E[Q] + \lambda = 0$
$2(\lambda - \mu)E[Q] + 2\lambda = 0$
$E[Q] = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1-\rho}$

Drift Method: Formalize

Formalize random variable with test function!

Function f mapping system states to real values. $f(q) = q^2$.

Formalize "increases and decreases": Instantaneous generator!

$$G \circ f(q) = \lim_{t \to 0} \frac{1}{t} E\left[f\big(Q(t)\big) - f(q) \middle| Q(0) = q\right]$$

For countable-state CTMC: Just rate of change times amount of change!

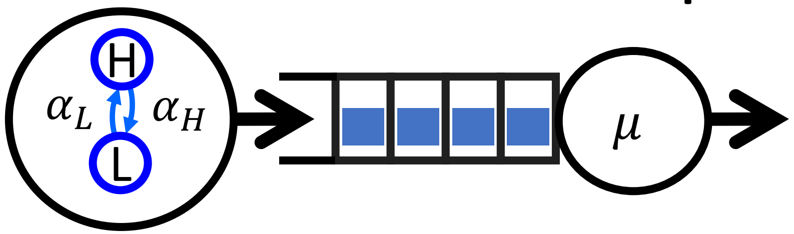
Formalize drift theorem:

Thm: For any f such that $E[f(Q)] < \infty$, the stationary drift is zero:

$$E[G \circ f(Q)] = 0.$$

Drift Method: 2-level, first attempt

Same random variable: Q^2



Four ingredients to drift: Rate of arrivals, change due to arrivals, rate of completions, change due to completions.

Q7: What's different between the M/M/1 and the 2-level?

A7: Rate of arrivals is now state dependent!

Let $Y(t) = y$ denote the arrival state. Y in stationarity.

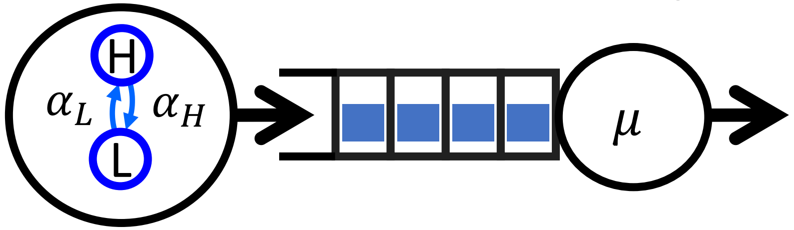
State $Q(t) = q, Y(t) = H$: Drift due to arrivals is $\lambda_H(2q + 1)$

State $Q(t) = q, Y(t) = L$: Drift due to arrivals is $\lambda_L(2q + 1)$

Q8: Expected drift due to arrivals, in stationarity?

A8: $2E[\lambda_Y Q] + \lambda$

Drift Method: 2-level, first attempt



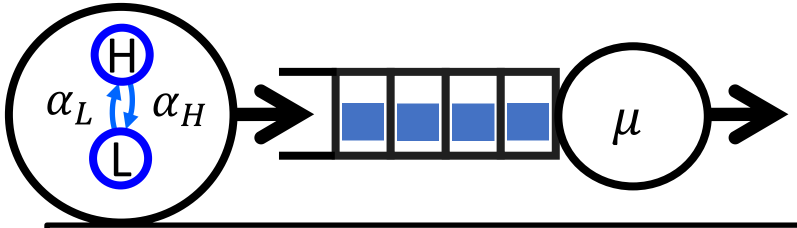
Apply key theorem ($E[G \circ Q^2] = 0$):

$$2E[\lambda_Y Q] + \lambda - 2\mu E[Q] + \lambda = 0$$
$$E[\lambda_Y Q] - \mu E[Q] + \lambda = 0$$

Conclusion: Q^2 doesn't work for the 2-level system.

Key idea of drift method: Find the right random variable/test function for your system.

We need to smooth out the arrival rates, get an $E[Q]$ drift term.

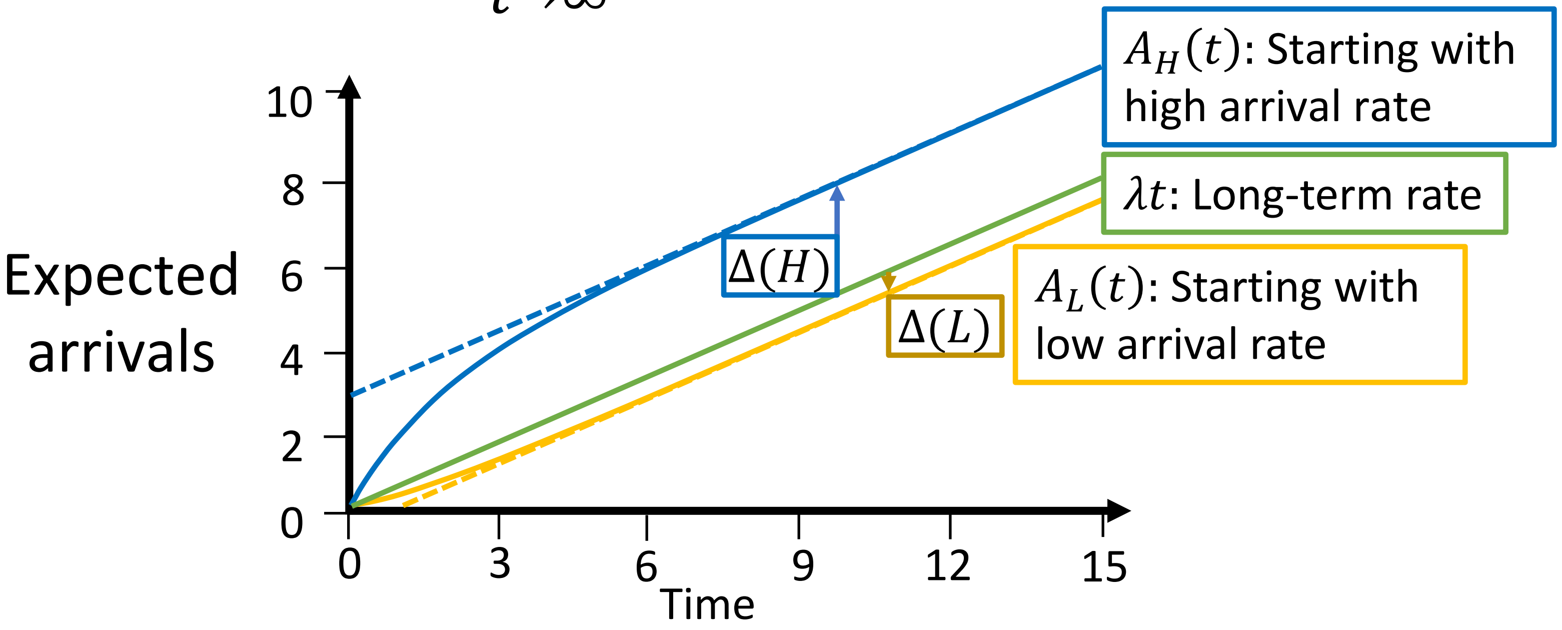


New idea: Relative arrivals

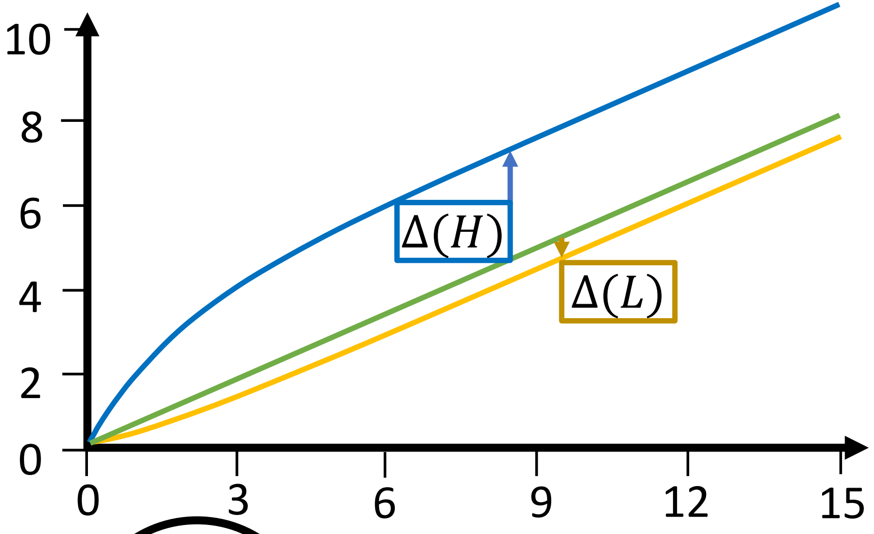
Setting:
$\lambda_H = 2, \lambda_L = 0.2, \alpha_H = 0.5,$
$\alpha_L = 0.1, \mu = 1, \lambda = 0.5$

Define $A_y(t)$ to be the number of arrivals up to time t, with the system initialized in arrival state y.

Relative arrivals: $\Delta(y) = \lim_{t\to\infty}\left(E\left[A_y(t)\right] - \lambda t\right)$



Relative Arrivals: Calculate

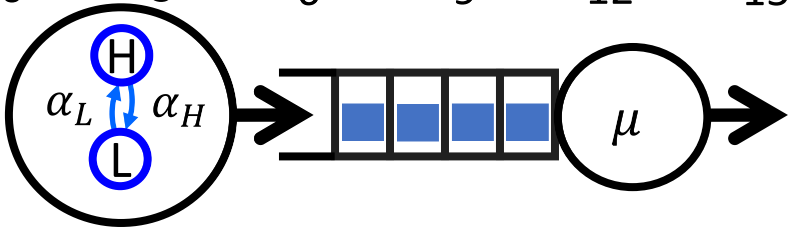


Equivalently, $\Delta(y)$ is the relative value of a Markov Reward Process with reward λ_y.

Can calculate $\Delta(y)$ using Poisson Equation:

$$\Delta(H) = \frac{\lambda_H - \lambda}{\alpha_H} + \Delta(L)$$

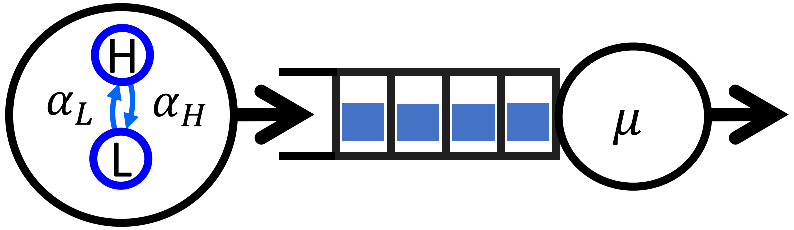
Another key fact: $E[\Delta(Y)] = 0$.



$$\Delta(y) = \lim_{t\to\infty}\left(E\left[A_y(t)\right] - \lambda t\right)$$

$$\Delta(L) = \frac{\lambda - \lambda_H}{\alpha_L + \alpha_H}\frac{\alpha_L}{\alpha_H}, \Delta(H) = \frac{\lambda - \lambda_L}{\alpha_L + \alpha_H}\frac{\alpha_H}{\alpha_L}$$

Drift of Relative Arrivals



What is the drift of $\Delta(y)$? Recall: $\Delta(H) = \frac{\lambda_H - \lambda}{\alpha_H} + \Delta(L)$

Q9: What makes y change?

Q9: Changing between H and L

Q10: At what rate does it change?

Q10: α_y

Q11: By how much does $\Delta(y)$ change?

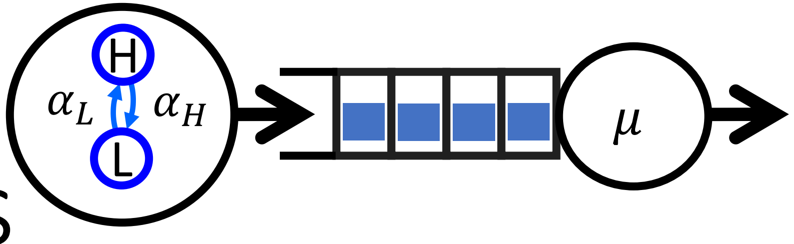
Q11: $\frac{\lambda - \lambda_y}{\alpha_y}$

Q12: What is the drift, $G \circ \Delta(y)$?

Q12: $\lambda - \lambda_y$

Alternate definition of $\Delta(y)$: "the test function with drift $G \circ \Delta(y) = \lambda - \lambda_y$"

Drift Method + Relative Arrivals



Queue length: $G \circ q = \lambda_y - \mu + \mu 1\{q = 0\}$

Relative arrivals: $G \circ \Delta(y) = \lambda - \lambda_y$

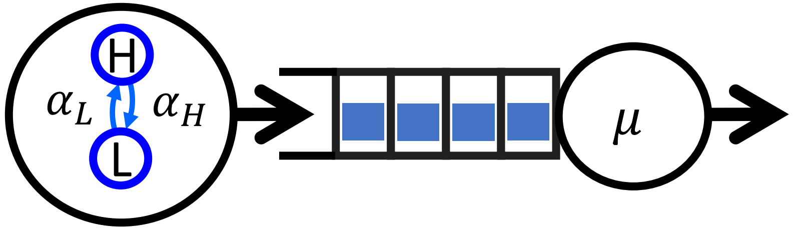
Constant drift: $G \circ \big(q + \Delta(y)\big) = \lambda - \mu + \mu 1\{q = 0\}$

If we can get constant drift, we can get linear drift, we can get $E[Q]$.

Magic test function/random variable: $q^2 + 2q\Delta(y)$

$$G \circ \big(q^2 + 2q\Delta(y)\big) = G \circ q^2 + 2q\big(G \circ \Delta(y)\big) + 2\Delta(y)(G \circ q)$$

$$= \underbrace{2q(\lambda - \mu)}_{\text{Linear!}} + \underbrace{2\Delta(y)\big(\lambda_y - \mu + \mu 1\{q = 0\}\big) + \lambda_y + \mu - \mu 1\{q = 0\}}_{\text{Bounded!}}$$



2-level result

$$G \circ \left(q^2 + 2q\Delta(y)\right) = \underbrace{\boxed{2q(\lambda - \mu)}}_{\text{Linear!}} \underbrace{\boxed{\begin{array}{l} + 2\Delta(y)\left(\lambda_y - \mu + \mu 1\{q = 0\}\right) \\ + \lambda_y + \mu - \mu 1\{q = 0\} \end{array}}}_{\text{Bounded!}}$$

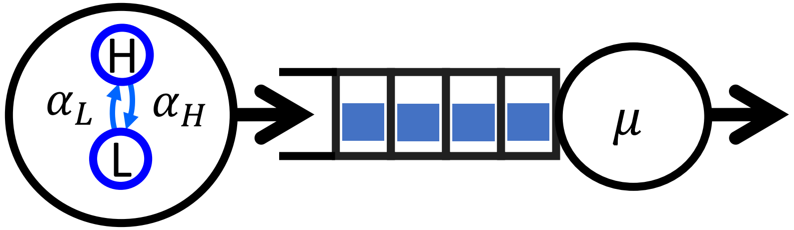
Fundamental drift theorem: $E\left[G \circ \left(Q^2 + 2Q\Delta(Y)\right)\right] = 0.$

$$0 = \boxed{2E[Q](\lambda - \mu)} + \boxed{2\mathrm{E}[\Delta(Y)\lambda_Y] + 2\mu E[\Delta(y)1\{Q = 0\}] + 2\lambda}$$

Thm:

$$\boxed{E[Q^{2-level}] = \frac{\rho}{1-\rho}\left(E[\Delta(Y)\lambda_Y]/\lambda + 1\right) + E[\Delta(Y) | Q = 0]}$$

Explicit result



Thm:

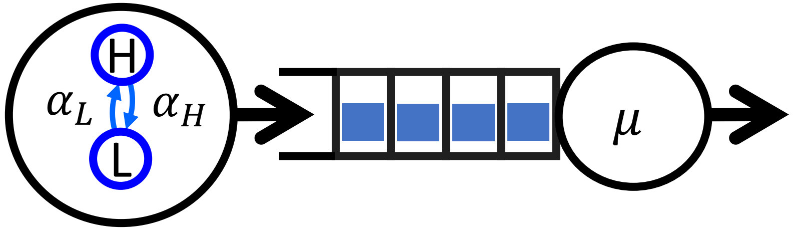
$$\boxed{E[Q^{2-level}] = \frac{\rho}{1-\rho}(E[\Delta(Y)\lambda_Y]/\lambda + 1) + E[\Delta(Y)|Q=0]}$$

$$E[\Delta(Y)\lambda_Y] = \frac{(\lambda_H - \lambda)(\lambda - \lambda_L)}{\alpha_H + \alpha_L}$$

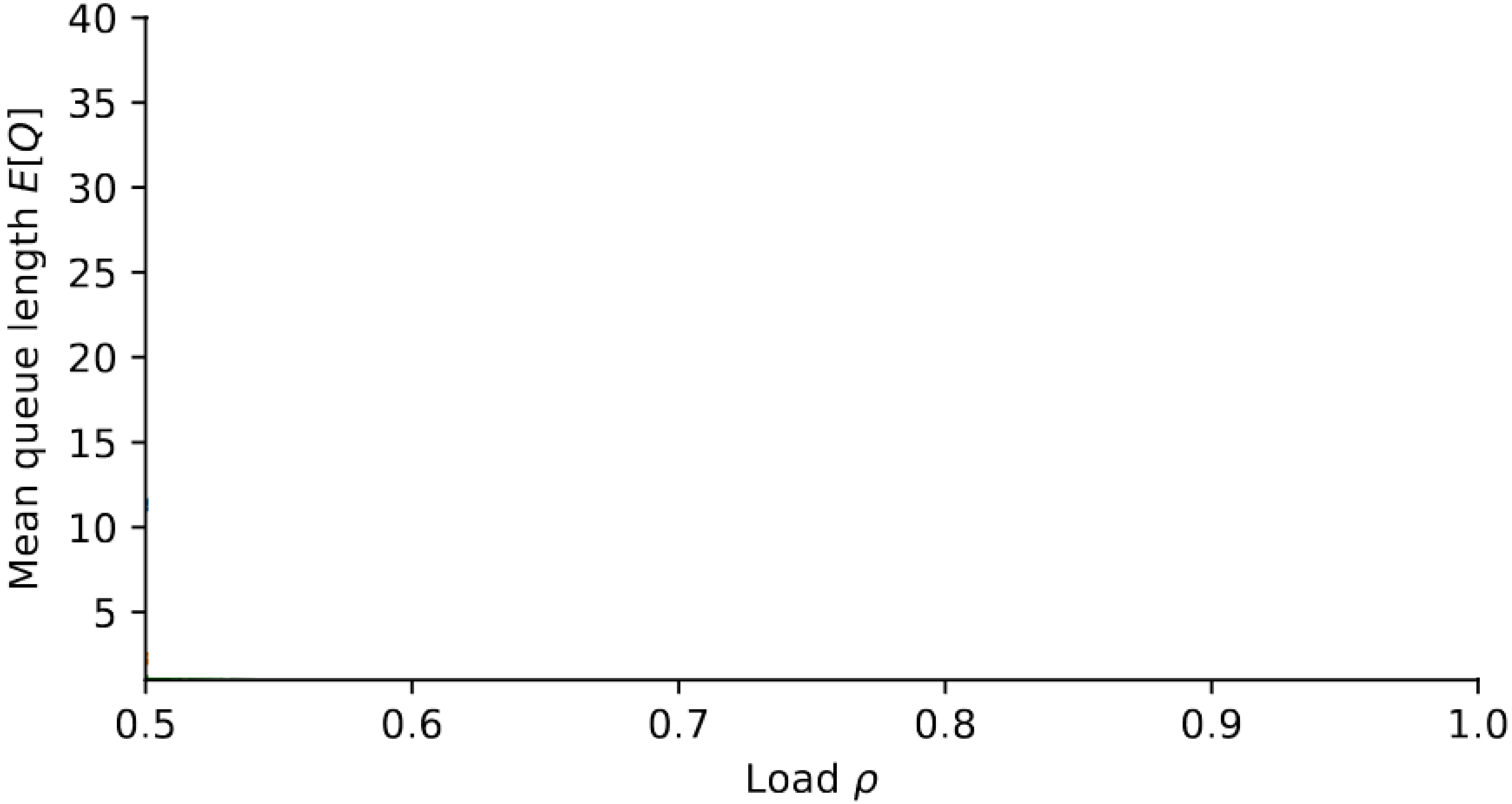
$$E[\Delta(Y)|Q=0] = \frac{\lambda_H - \lambda}{\alpha_H}\left(P(Y=H|Q=0) - \frac{\alpha_L}{\alpha_L + \alpha_H}\right)$$

Tight bounds, even just from $0 \leq P(Y=H|Q=0) \leq 1$!

"Analysis of Markovian Arrivals and Service with Applications to Intermittent Overload". Grosof, Hong, Harchol-Balter.



Compare to simulation



Setting: $\lambda_H = 4\rho, \lambda_L = 0.4\rho, \alpha_H = 5\alpha_L, \mu = 1$

Key Ideas

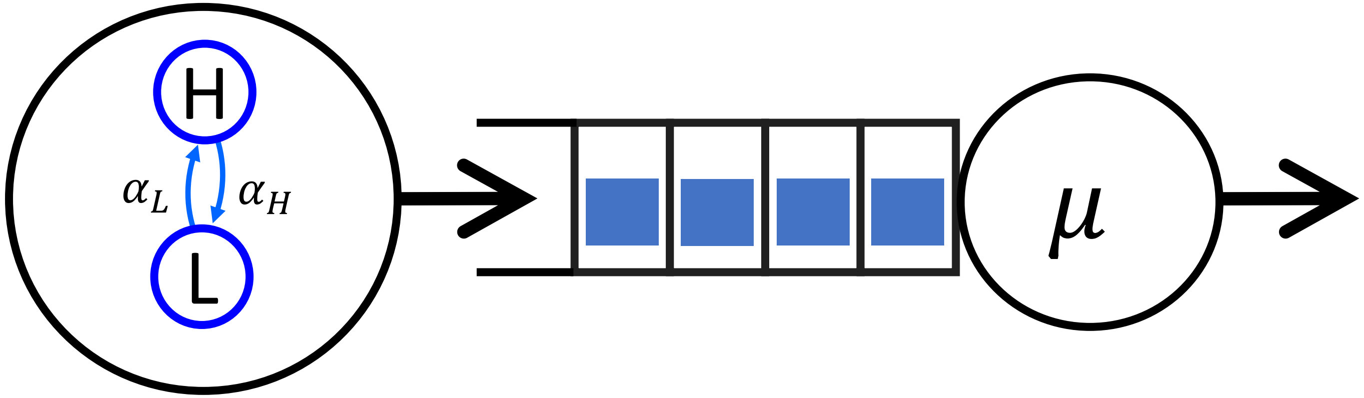
Design a random variable/test function to have just the right drift for what we want.

Separate the q part of the drift from the y part of the drift.

$$G \circ \left(q^2 + 2q\Delta(y)\right) = \underbrace{\boxed{2q(\lambda - \mu)}}_{\text{Linear!}} \underbrace{\boxed{\begin{aligned} &+ 2\Delta(y)\left(\lambda_y - \mu + \mu 1\{q = 0\}\right) \\ &+\lambda_y + \mu - \mu 1\{q = 0\} \end{aligned}}}_{\text{Bounded!}}$$

Very widely-applicable idea!

Break Time!



Thm:

$$E[Q^{2-level}] = \frac{\rho}{1-\rho}\left(E[\Delta(Y)\lambda_Y]/\lambda + 1\right) + E[\Delta(Y)|Q = 0]$$

Outline

Simple MAMS: 2-level Arrivals

Drift Method + Relative Arrivals

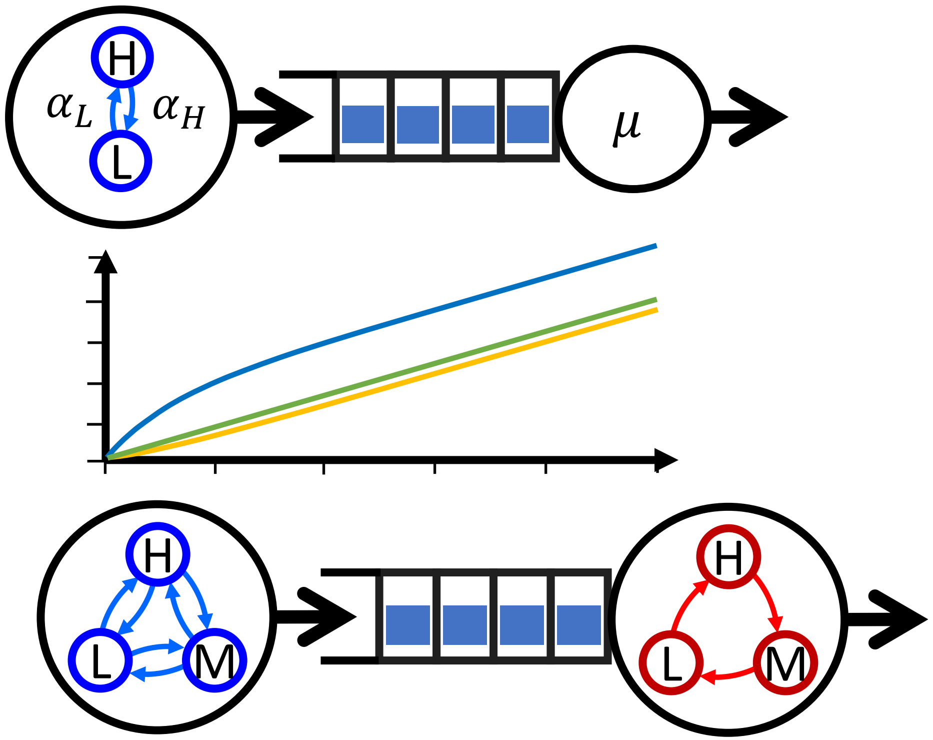
⟵ ⟵ ⟵ Break time ⟶ ⟶ ⟶

Generalizing to Full MAMS

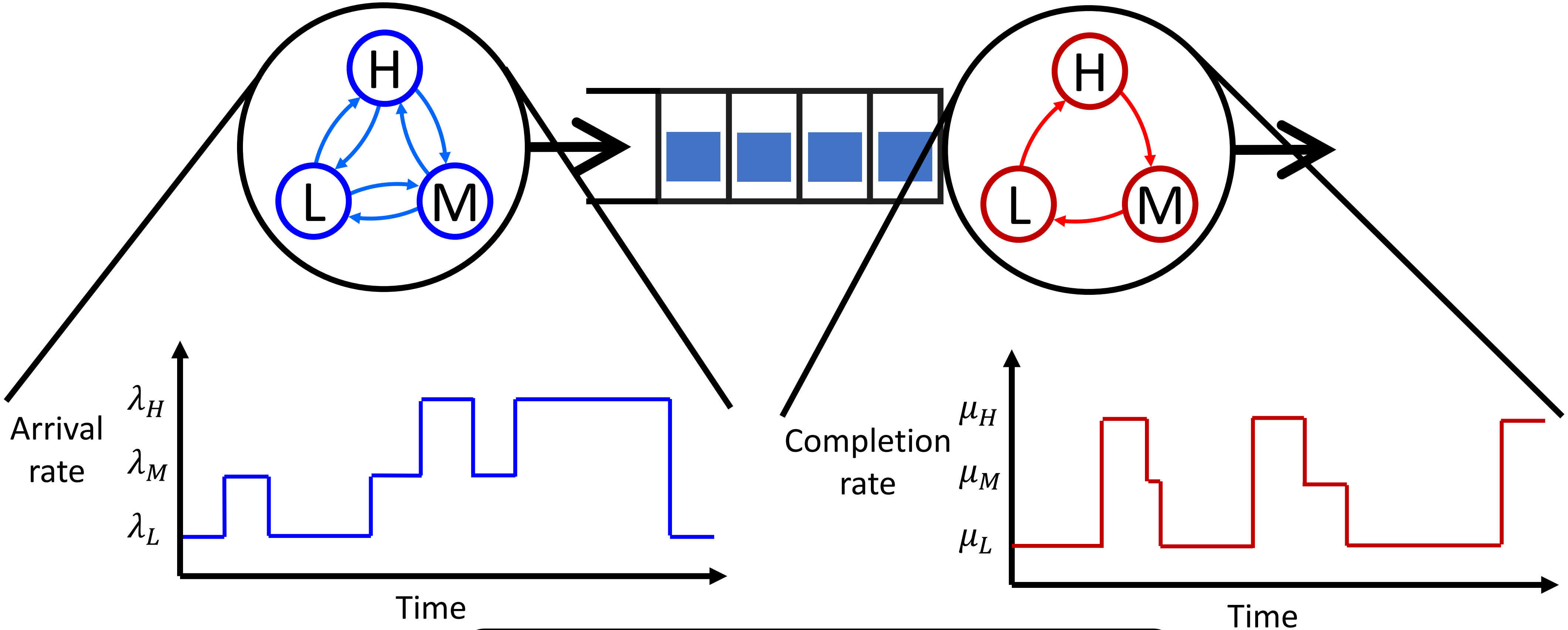
Applications: Fluctuating Load

Multiserver Jobs

Networks with Abandonment (e.g. Quantum switching network)



General Markovian Arrivals & Markovian Service

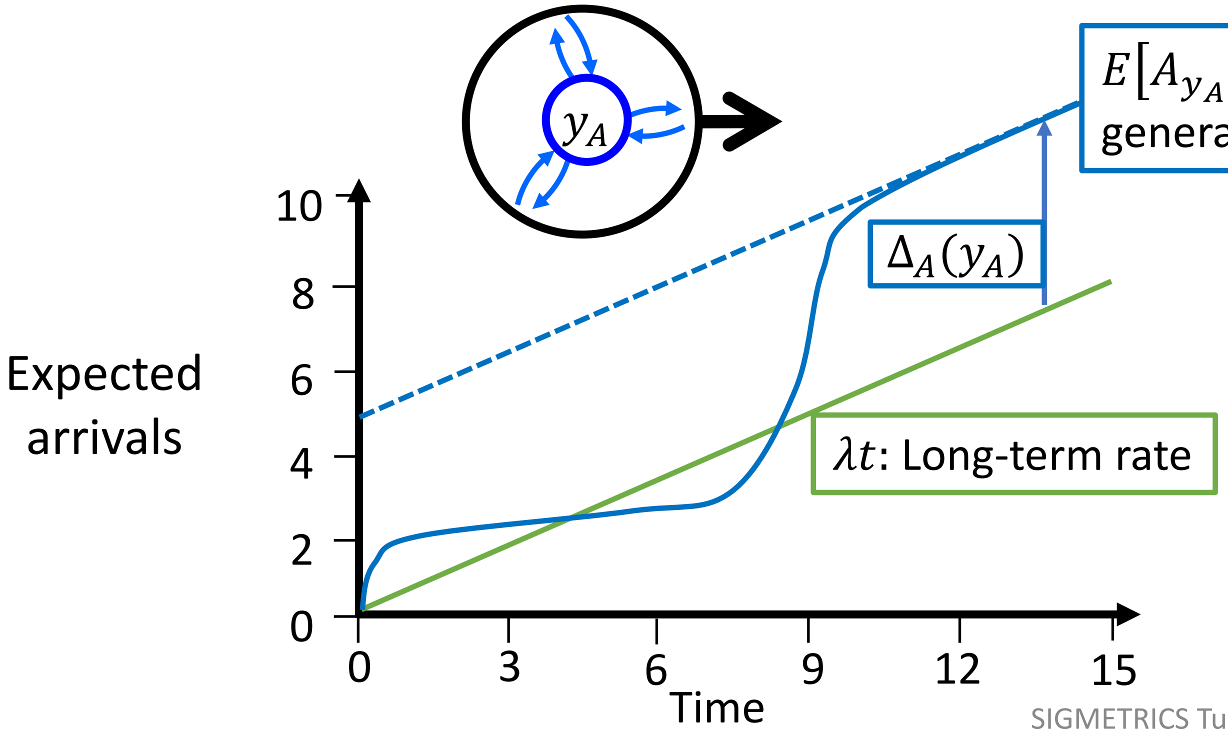


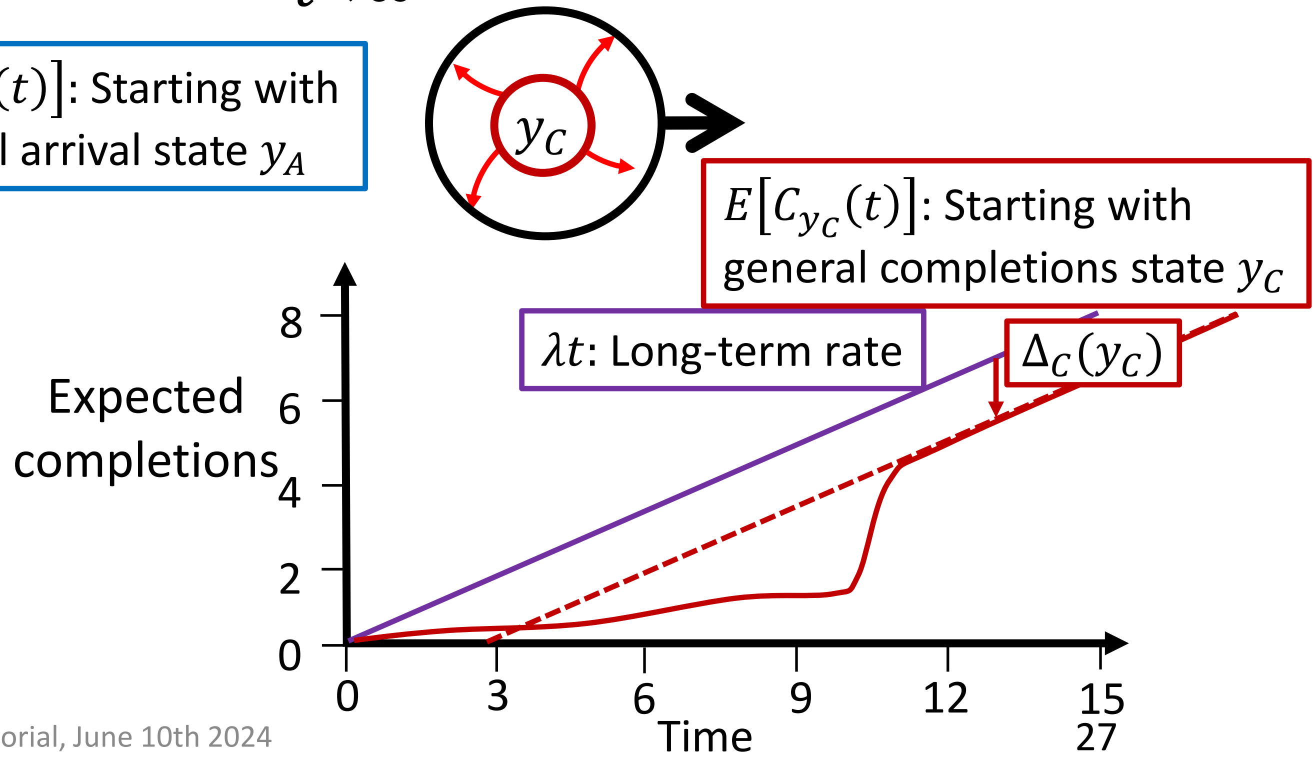
Goal: Simple, explicit characterization of mean queue length, $E[Q]$

Generalizing: Relative Arrivals & Completions

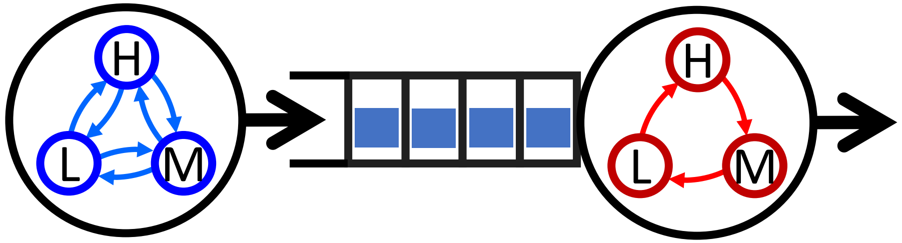
Relative Arrivals: Same definition: $\Delta_A(y_A) = \lim_{t\to\infty}\left(E\left[A_{y_A}(t)\right] - \lambda t\right)$

Relative Completions: Same idea: $\Delta_C(y_C) = \lim_{t\to\infty}\left(E\left[C_{y_C}(t)\right] - \lambda t\right)$





Drift for MAMS



Queue length: $G \circ q = \lambda_{y_A} - \mu_{y_C} + \mu_{y_C} 1\{q = 0\}$

Relative arrivals: $G \circ \Delta_A(y_A) = \lambda - \lambda_{y_A}$

Same drift as before!

Relative completions: $G \circ \Delta_C(y_C) = \mu - \mu_{y_C}$

Constant drift: $G \circ \big(q + \Delta_A(y_A) - \Delta_C(y_C)\big) = \lambda - \mu + \mu_{y_C} 1\{q = 0\}$

If we can get constant drift, we can get linear drift, we can get $E[Q]$.

Magic test function/random variable: $q^2 + 2q\Delta_A(y_A) - 2q\Delta_C(y_C)$

$$G \circ \big(q^2 + 2q\Delta_A(y_A) - 2q\Delta_C(y_C)\big) = \underbrace{\boxed{2q(\lambda - \mu)}}_{\text{Linear!}} + \underbrace{\boxed{f(y_A, y_C, 1\{q = 0\})}}_{\text{Bounded!}}$$

General MAMS Result

Thm:

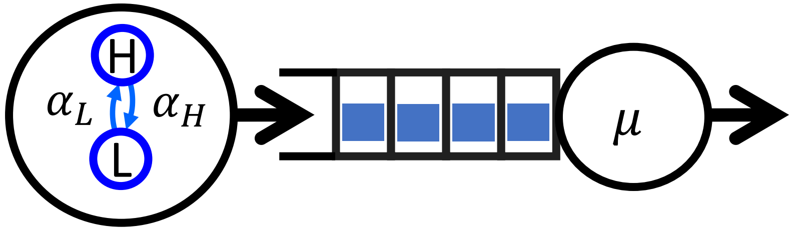
$$\boxed{E[Q^{MAMS}] = \frac{E\left[\Delta_{\mathrm{A}}(Y_A)\lambda_{Y_A}\right]/\mu + E\left[\Delta_C(Y_C)\mu_{Y_C}\right]/\mu + \rho}{1-\rho} + E_U[\Delta_A(Y_A) - \Delta_C(Y_C)]}$$

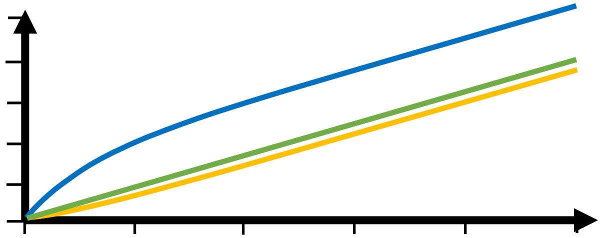
$E_U[\cdot]$: Expectation over moments of unused service.

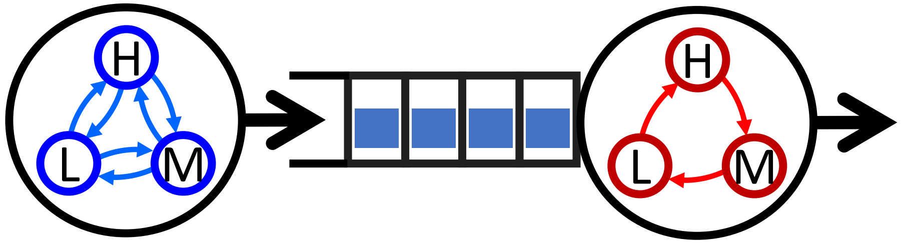
Tight bounds, even just from

$$E_U[\Delta_A(Y_A) - \Delta_C(Y_C)] \in \left[\Delta_A^{\min} - \Delta_C^{\max}, \Delta_A^{\max} - \Delta_C^{\min}\right]$$

Outline



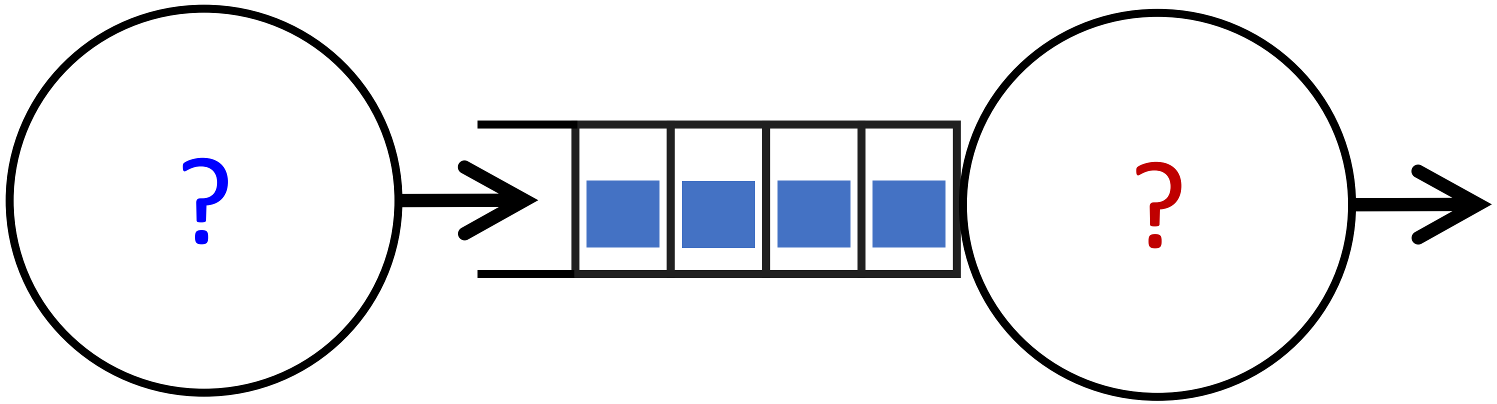




- Simple MAMS: 2-level Arrivals
- Drift Method + Relative Arrivals
- ⟵ ⟵ ⟵ Break time ⟶ ⟶ ⟶
- Generalizing to Full MAMS
- Applications: Fluctuating Load
- Multiserver Jobs
- Networks with Abandonment (e.g. Quantum switching network)

Applications: Fluctuating Load

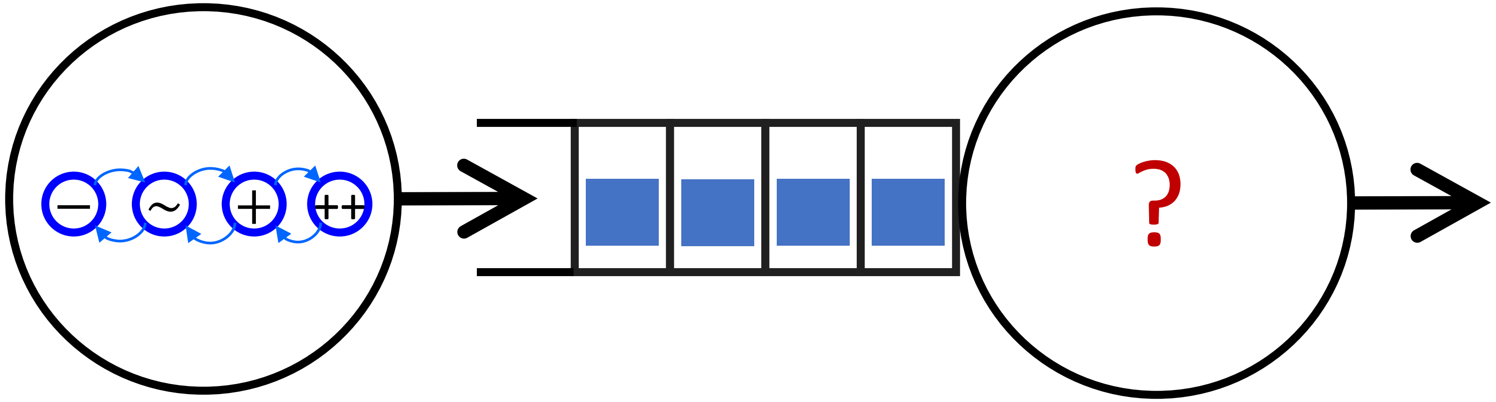
Example: Datacenter



Arrivals: Normal load (∼), off hours (−), peak load (+), rare event (+ +)

Applications: Fluctuating Load

Example: Datacenter

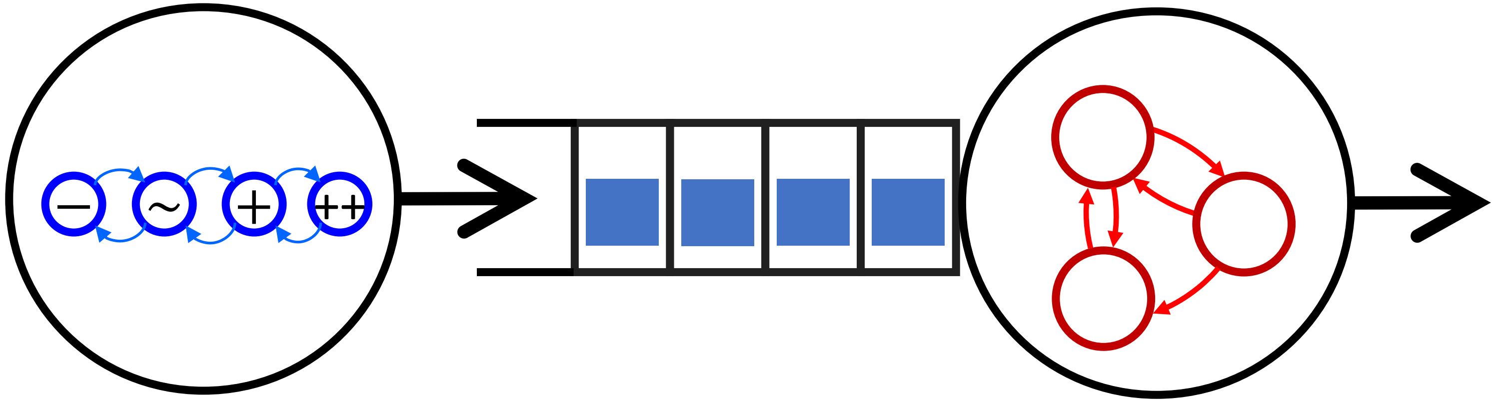


Arrivals: Normal load (~), off hours (–), peak load (+), rare event (+ +)

Service: Full operation (😊), maintenance (🛠), outage (✕)

Applications: Fluctuating Load

Example: Datacenter

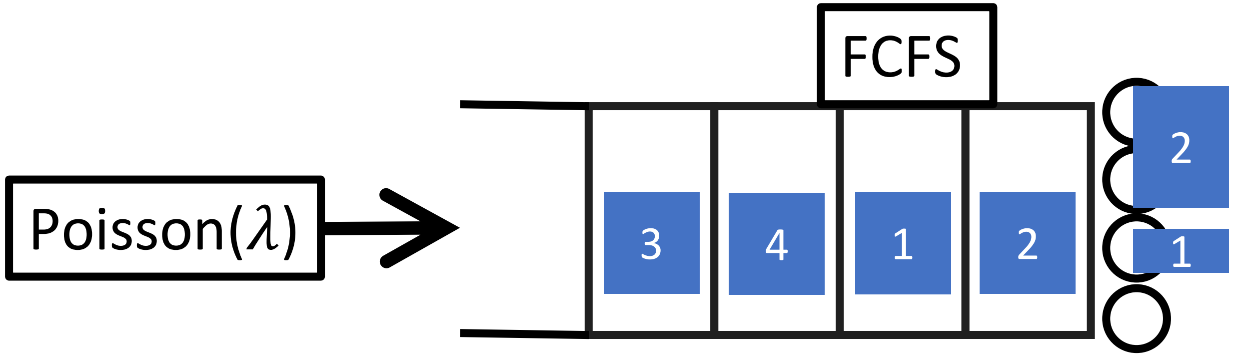


Arrivals: Normal load (∼), off hours (−), peak load (+), rare event (+ +)

Service: Full operation (😊), maintenance (🛠), outage (✕)

MAMS Model: Performance characterization from relative arrivals and relative completions.

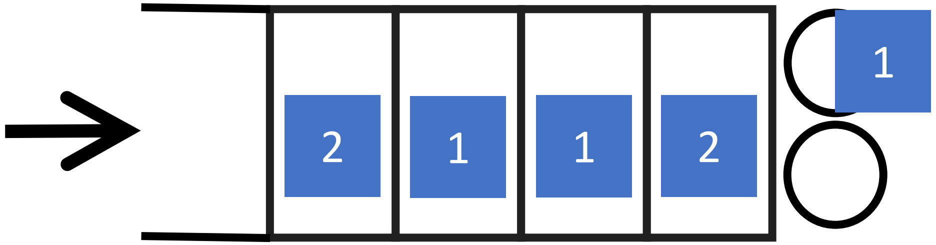
Application: Multiserver-job Model



Job: (duration, server need)

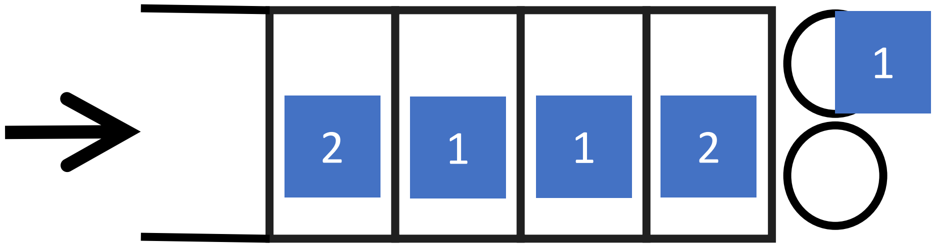
Sampled i.i.d. from joint distribution, phase-type job durations

Simple example: 2 servers.
Distribution: $(Exp(\mu_1), 1)$ & $(Exp(\mu_2), 2)$



Goal: Simple, explicit characterization of mean queue length, $E[Q]$

Applications: Multiserver Jobs



Jobs: $(Exp(\mu_1), 1), (Exp(\mu_2), 2)$

Service rate fluctuations!

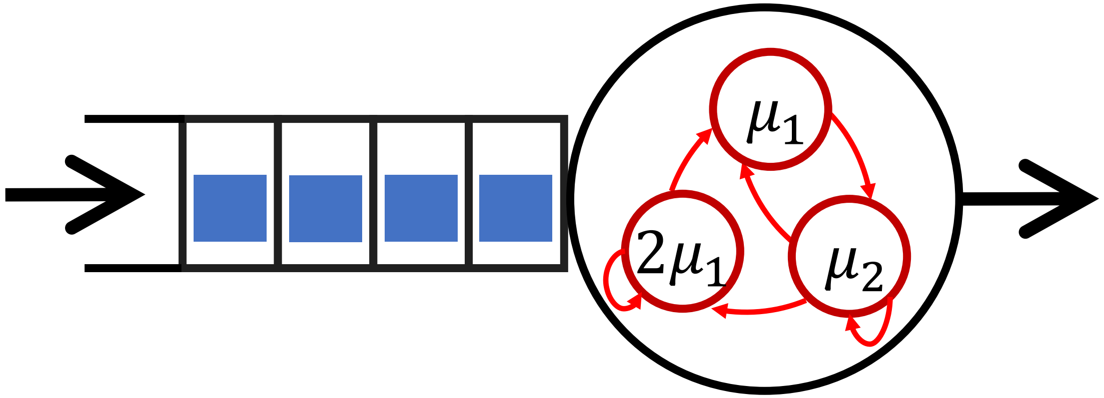
Internal fluctuation, not external.

Use MAMS anyways!

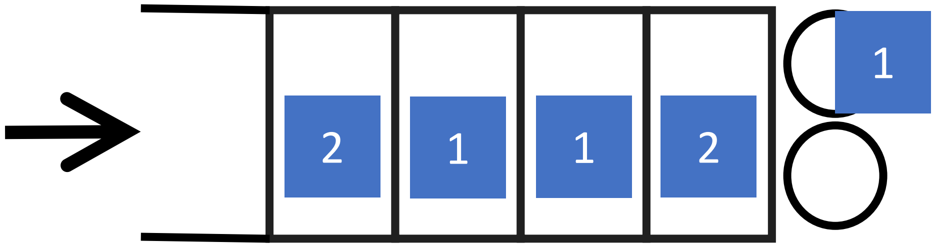
Q13: What service rates are possible?

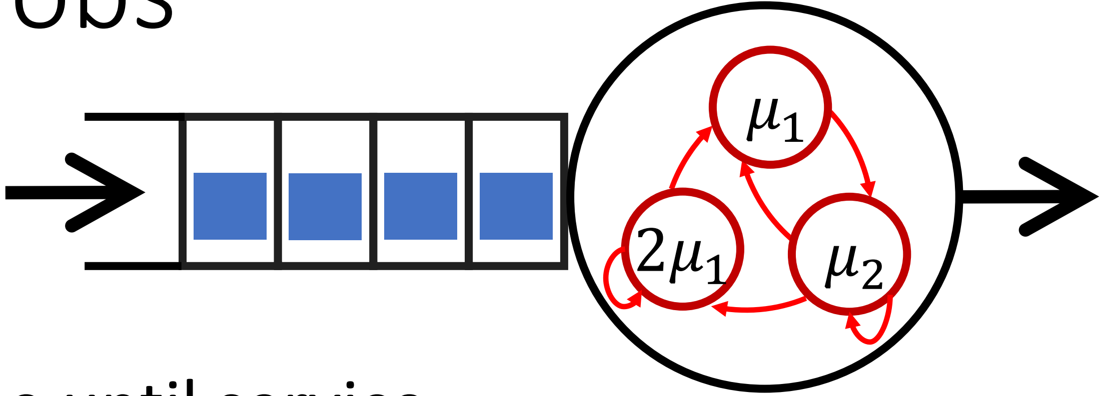
A13: $\mu_1, 2\mu_1, \mu_2$

Compare with MAMS:



Applications: Multiserver Jobs





Almost identical – delay sampling job type until service.

However: Different if no jobs in system. Or one job.

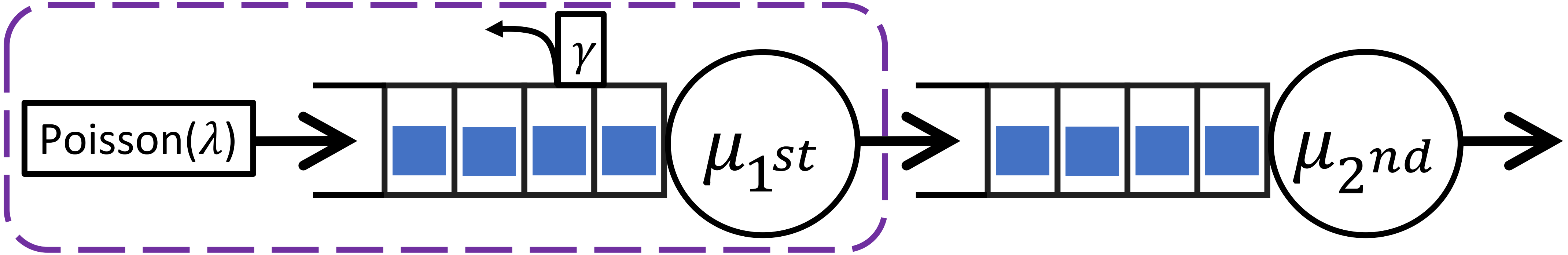
Same if 2+ jobs in system.

“Near-MAMS”

$\boxed{\text{Thm (RESET): } E[Q^{Near-MAMS}] = E[Q^{MAMS}] + O_\rho(1)}$

The RESET and MARC Techniques, with Application to Multiserver-Job Analysis. [GHHS ’23]

Application: Tandem queue with abandonment

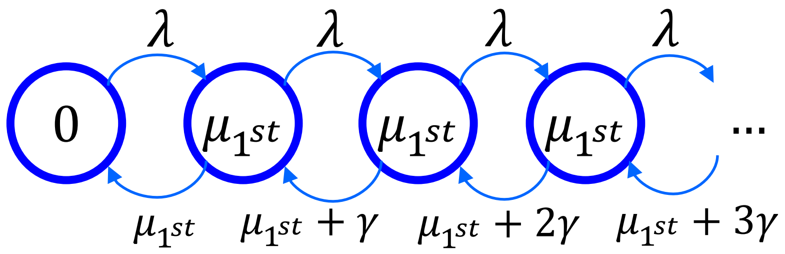


Coffeeshop: First, customers queue to order and pay. May abandon. Second, customers queue to pick up their drinks. No abandonment.

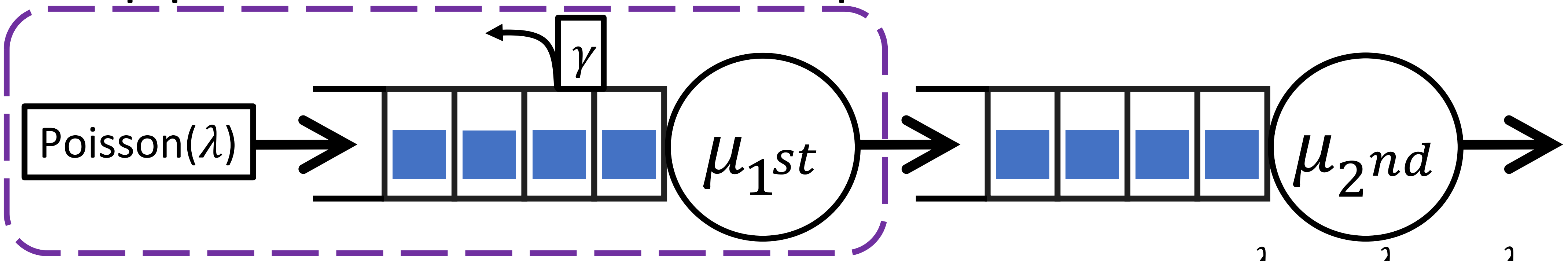
Idea: First queue as Markovian Arrival process!

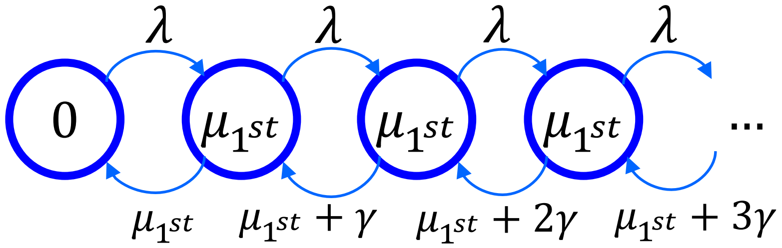
Q14: Draw Markov Chain of arrival rates to second queue

A14:



Application: Tandem queue with abandonment





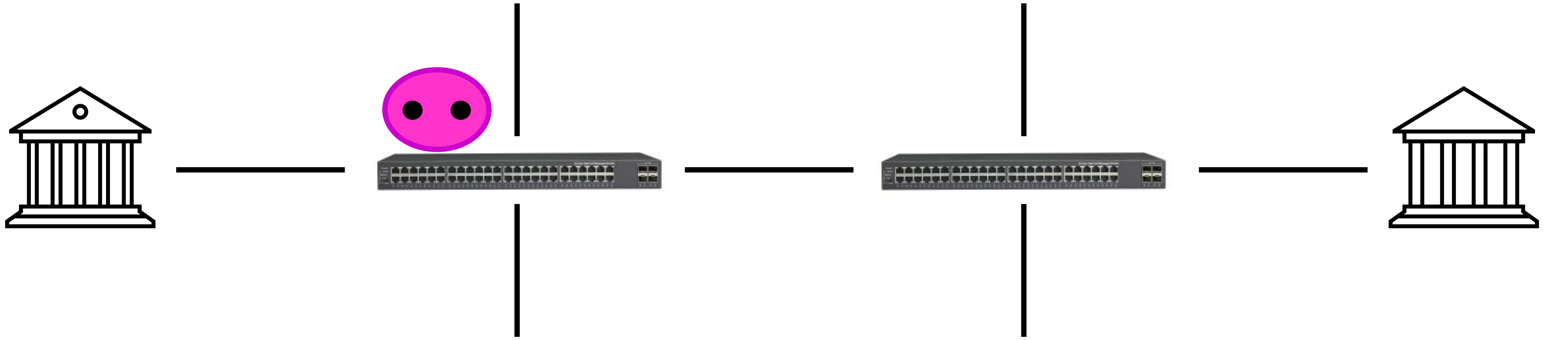
Infinite-MAMS – Uniform positive recurrence

Still define relative arrivals, $\Delta(y)$

Same mean queue length result:

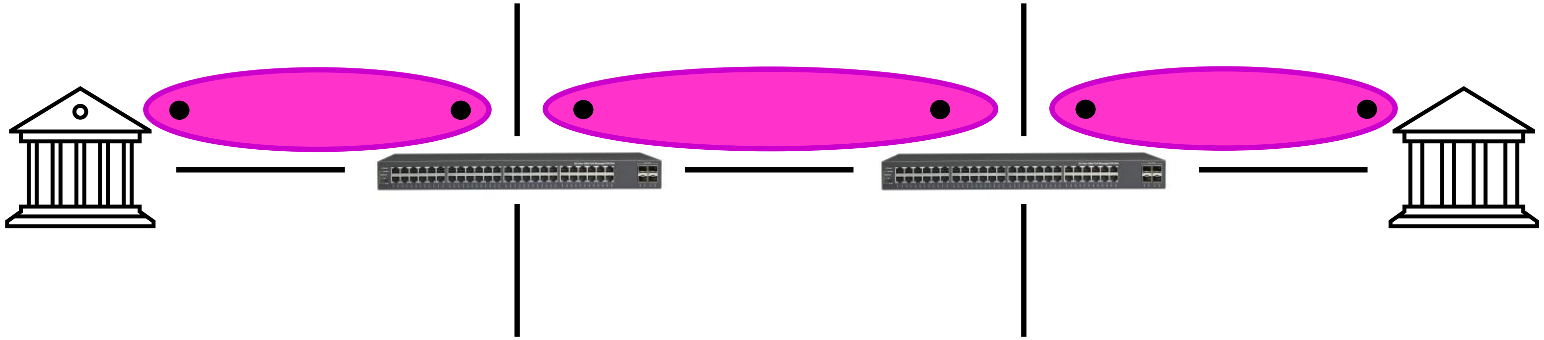
$$\boxed{E[Q^{2-level}] = \frac{\rho}{1-\rho}(E[\Delta(Y)\lambda_Y]/\lambda + 1) + \overbrace{E[\Delta(Y)|Q=0]}^{\text{More work to bound}}}$$

Application: Quantum switching network



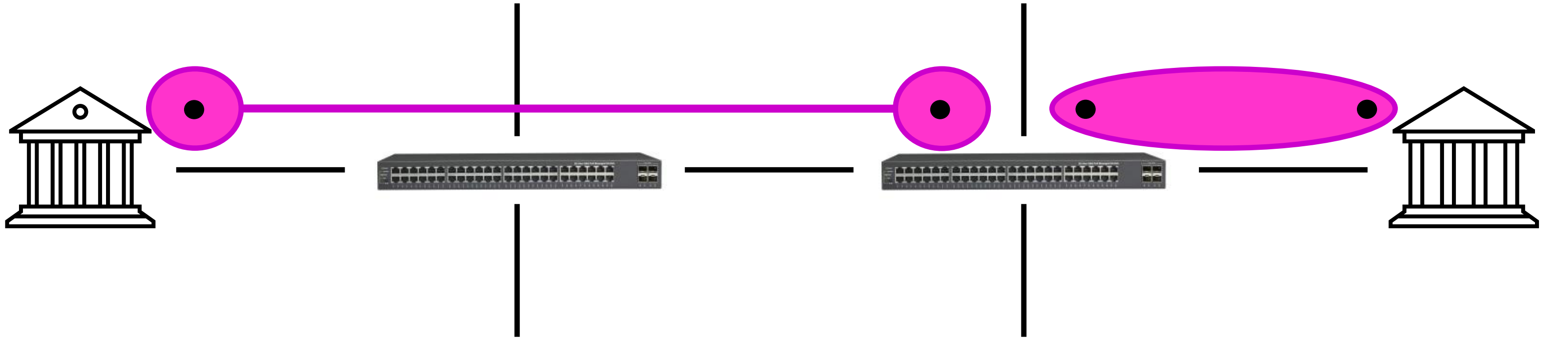
1. Switch generates entangled qubit-pairs

Application: Quantum switching network



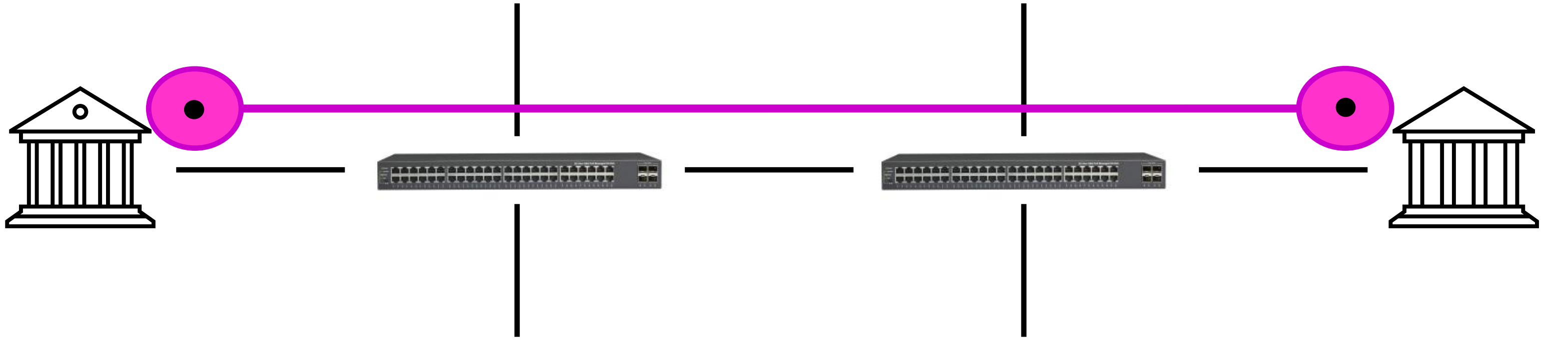
1. Switch generates entangled qubit-pairs
2. Switch transmits half of entangled pair

Application: Quantum switching network



1. Switch generates entangled qubit-pairs
2. Switch transmits half of entangled pair
3. Switch combines two entangled pairs to make longer-distance pair

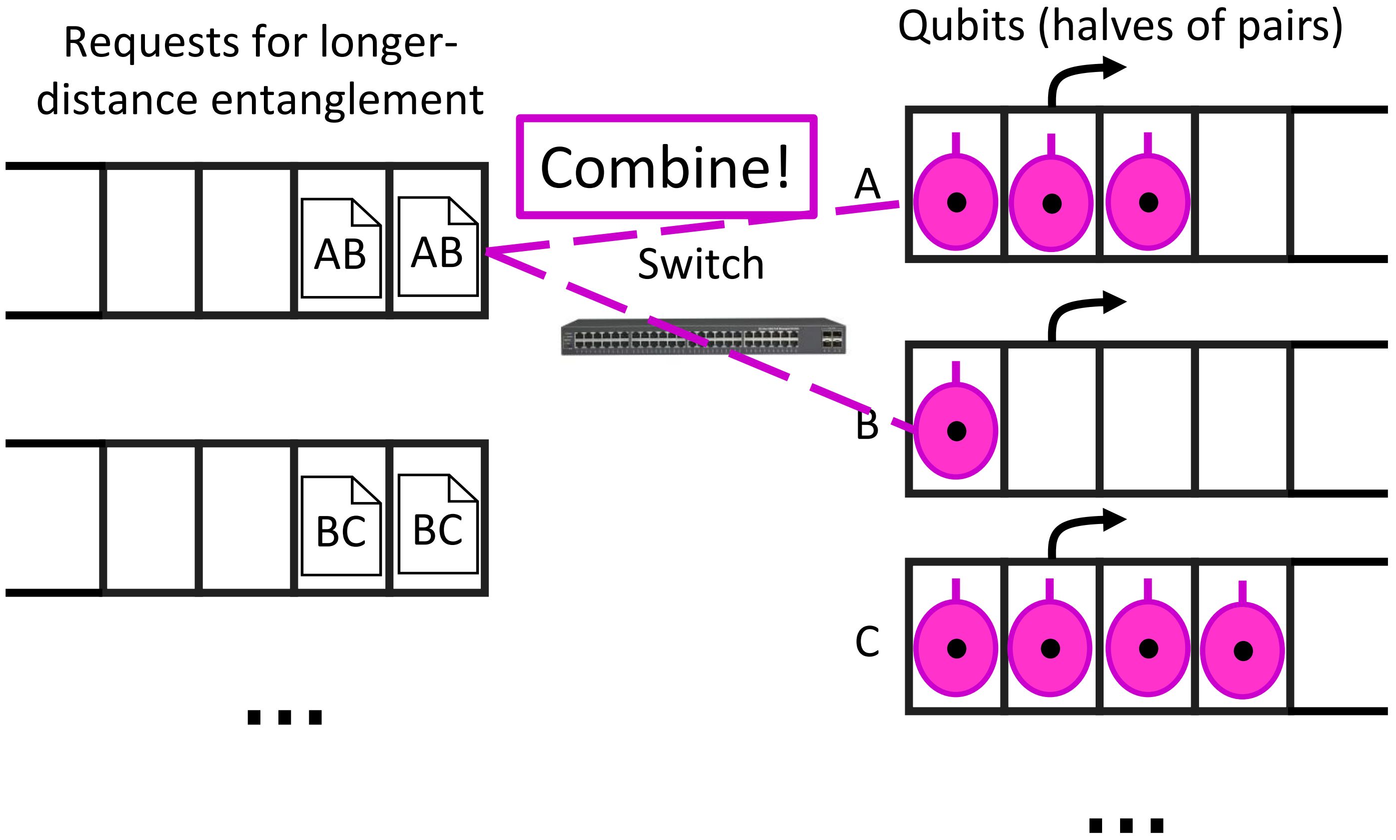
Application: Quantum switching network



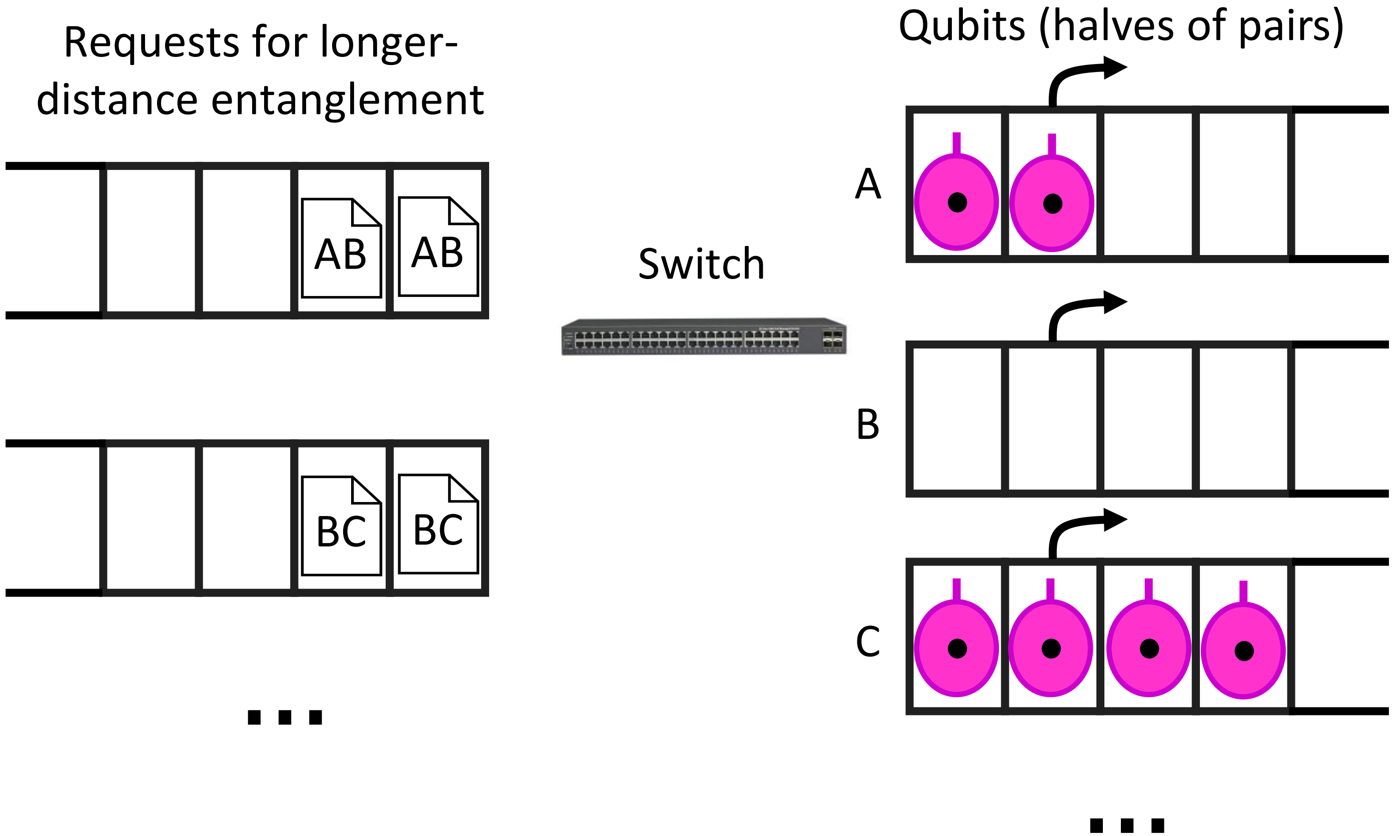
1. Switch generates entangled qubit-pairs
2. Switch transmits half of entangled pair
3. Switch combines two entangled pairs to make longer-distance pair

“Matching Queues with Abandonments in Quantum Switches: Stability and Throughput Analysis” [ZJM]

Quantum switching: Switch perspective

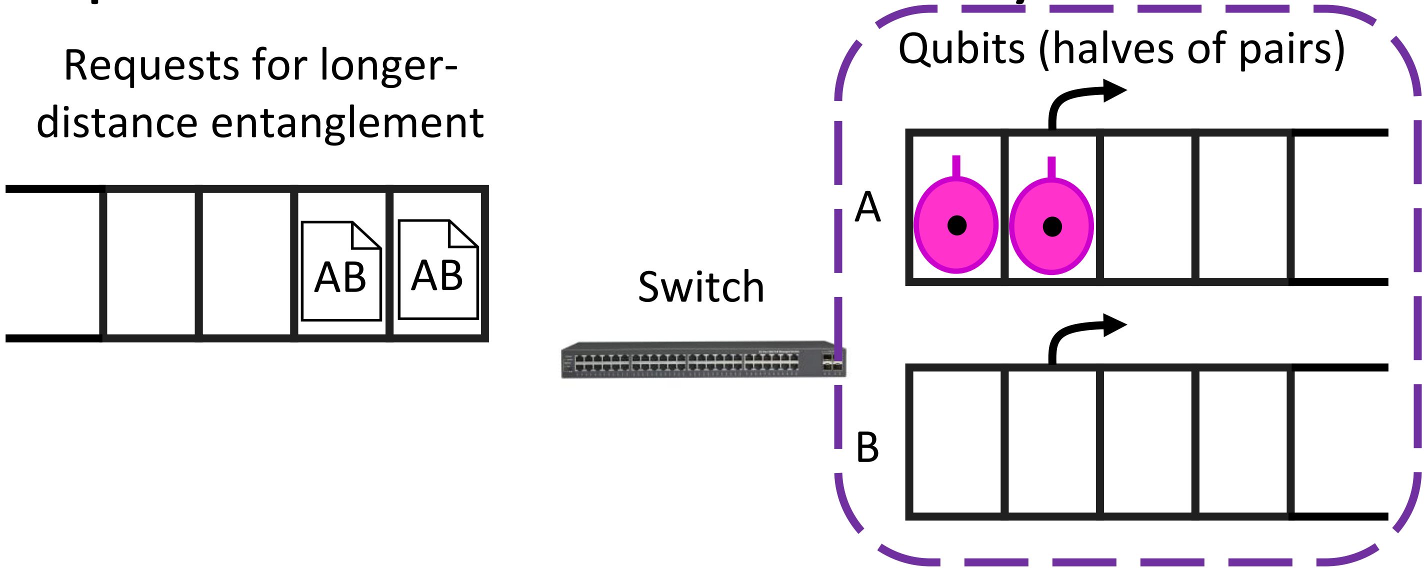


Quantum switching: Switch perspective



Goal: Simple, explicit characterization of mean queue length, $E[Q]$

Simplest quantum switch: Y-system



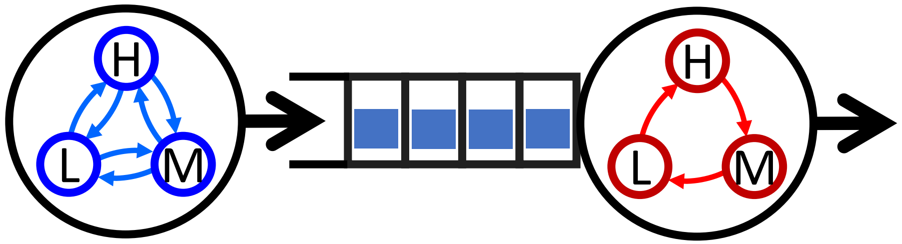
Idea: think of qubits as service-rate modulation process.

Uniform positive recurrence – Infinite-MAMS!

Alternative behavior when no requests – Near-MAMS!

Plan: Characterize $E[Q]$ using near-MAMS + infinite-MAMS

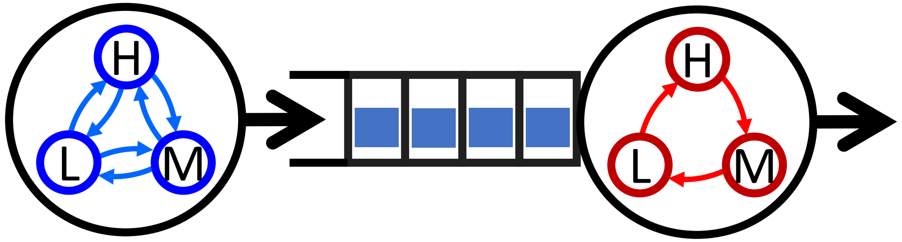
Further Directions



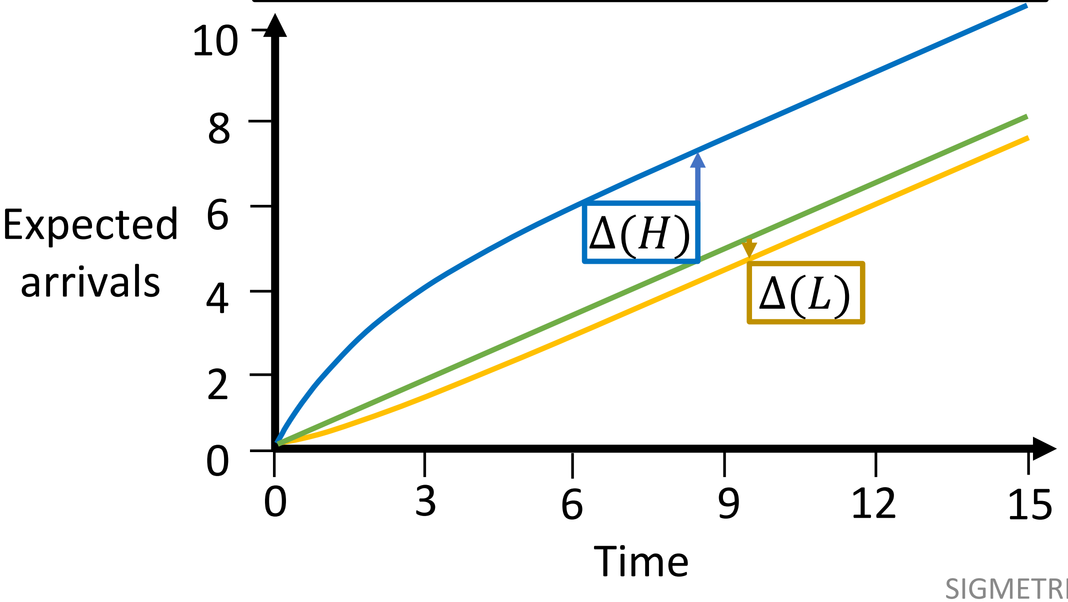
- Tail performance: $E[e^{-sQ}]$? $E\left[e^{-s(1-\rho)Q}\right]$?
- MAMS-work: Jobs have sizes, modulate work completion rate
- MAMS & drift concepts for scheduling
- Two MAMA papers on Friday:
  - 2:15pm: "Bounds on M/G/k Scheduling Under Moderate Load" [G., Wang]
  - 2:45pm: "Simple Policies for Multiresource Job Scheduling" [Chen, G., Berg]
- Your application/model/setting!

Conclusion

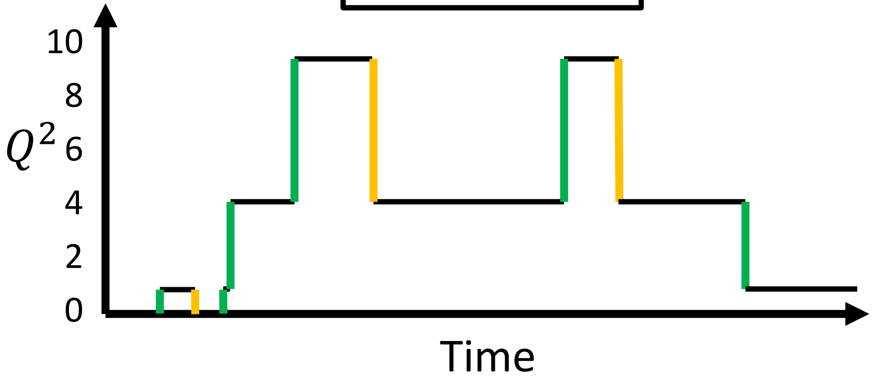
Markovian Arrivals, Markovian Service



Relative Arrivals, Relative Completions



Drift Method



Simple, explicit, tight characterization of mean queue length:

$$E[Q^{2-level}] = \frac{\rho}{1-\rho}(E[\Delta(Y)\lambda_Y]/\lambda + 1) + E[\Delta(Y)|Q = 0]$$

Applications: Fluctuating load, Multiserver jobs, Tandem queues with abandonment, Quantum switching!

Bonus: MAMS Plot!

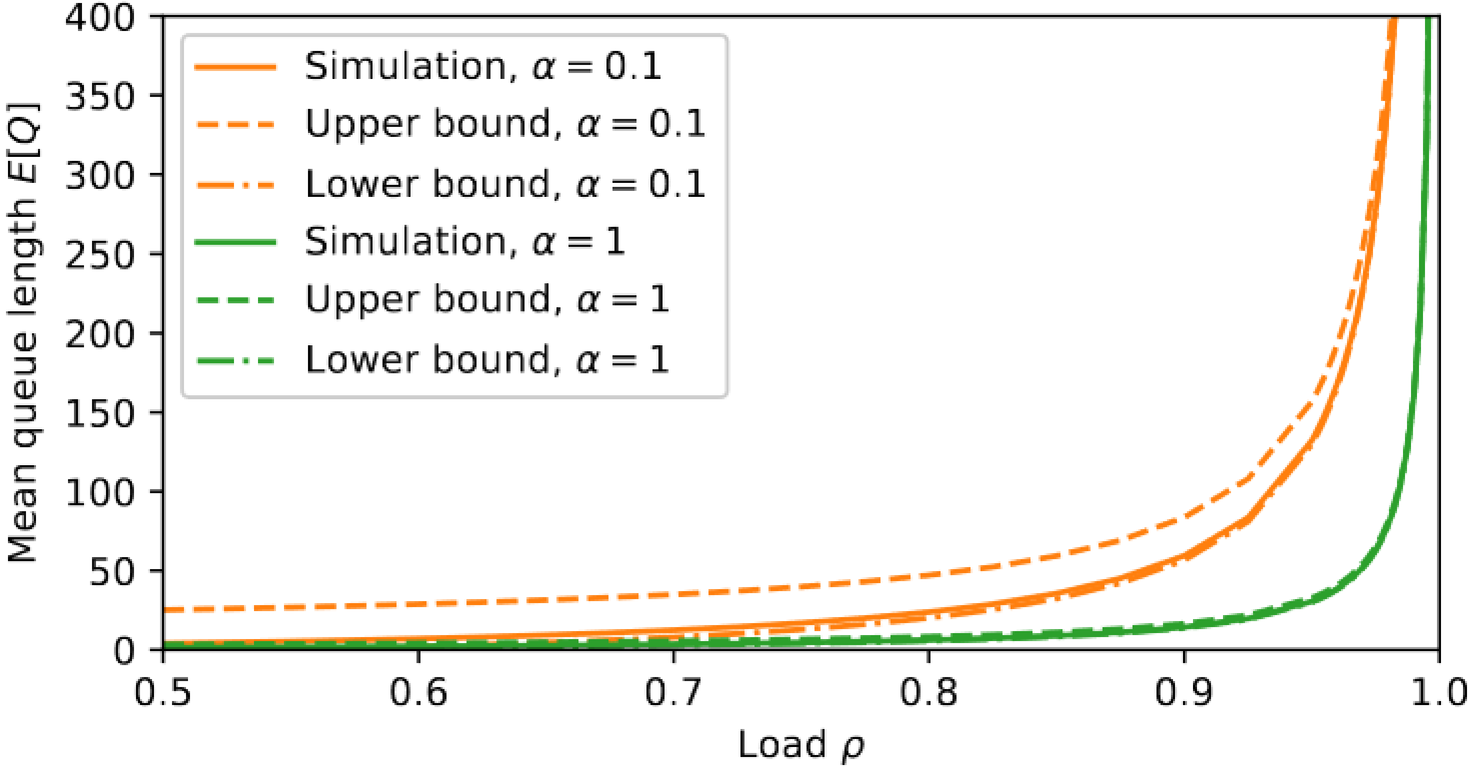


Figure 6: Setting: MAMS queue with three arrivals levels: $[0.3\rho, 2\rho, 2.2\rho]$, and three completions levels: [0.5, 1.0, 3.0]. The system remains in each arrival state and each service state for time $Exp(\alpha)$, then moves cyclically to the next rate in the list, wrapping around. Bounds given in Corollary 5.1. Simulated 10^9 arrivals.