Scheduling for Multiclass Many-server Queues with Abandonment: the $c\mu/\theta$ Rule and its Generalizations

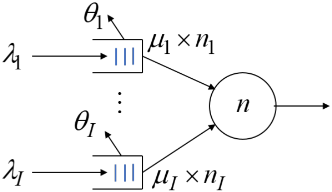
Nahum Shimkin

Technion – Israel Institute of Technology

YEQT 2018 - Young European Queueing Theorists XII
December 3-4, Toulouse, France

- The first part is joint work with Chanit Giat and Rami Atar (Technion).
- The second part is joint work with Zhenghua Long, Hailun Zhang and Jiheng Zhang (HKUST)

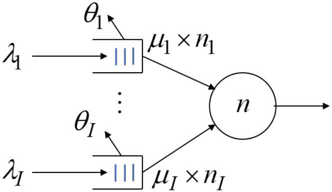
THE BASIC MODEL



Consider a queueing systems with:

- n identical servers
- Finite set $\mathcal{I} = \{1 \ldots I\}$ of customer classes
- Poisson arrivals, with rates λ_i, $i \in \mathcal{I}$
- Exponential service times, with means μ_i
- *Impatient customers:* exponential patience time, with mean θ_i

THE BASIC MODEL



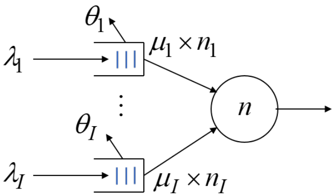
Consider a queueing systems with:

- n identical servers
- Finite set $\mathcal{I} = \{1 \dots I\}$ of customer classes
- Poisson arrivals, with rates λ_i, $i \in \mathcal{I}$
- Exponential service times, with means μ_i
- *Impatient customers:* exponential patience time, with mean θ_i

We focus here on the case of an overloaded system:

$$\sum_i \frac{\lambda_i}{\mu_i} > 1$$

COST PARAMETERS



- Waiting cost parameter c_i

Our cost function:

$$J(T) = \frac{1}{T}\mathbb{E}\int_0^T \sum_{i=1}^{I} c_i Q_i(t)dt$$

(for large T).

What about abandonment penalties?

What about abandonment penalties?

Consider the cost function

$$J(T) = \frac{1}{T}\mathbb{E}\int_0^T \sum_{i=1}^{I}(c_i Q_i(t) + \gamma_i dN_i^{aban}(t))dt$$

What about abandonment penalties?

Consider the cost function

$$J(T) = \frac{1}{T}\mathbb{E}\int_0^T \sum_{i=1}^{I}(c_i Q_i(t) + \gamma_i dN_i^{aban}(t))dt$$

Since patience is exponentially-distributed,

$$\mathbb{E}(dN_t^{aban}(t)) = \theta_i Q_i(t)\,,$$

and this cost reduces to the previous one with $c_i \leftarrow c_i + \gamma_i\theta_i$.

PRIORITY RULES

- For the single-server queue with no abandonment, the optimal scheduling policy is the celebrated $c\mu$ index rule [Cox & Smith 1951, etc.]
- For the same queue with convex delay costs, the generalized $c\mu$ rule is asymptotically optimal under the heavy-traffic diffusion regime [Van Mieghem 1995].

PRIORITY RULES

- For the single-server queue with no abandonment, the optimal scheduling policy is the celebrated $c\mu$ index rule [Cox & Smith 1951, etc.]
- For the same queue with convex delay costs, the generalized $c\mu$ rule is asymptotically optimal under the heavy-traffic diffusion regime [Van Mieghem 1995].
- We wish to find a simple scheduling policy, which is close to optimal under suitable conditions.

FLUID SCALING

- We consider the case of many servers, namely $n \to \infty$.
- Accordingly, we let $\lambda_i^n = n\lambda_i$.
  μ_i, θ_i and the cost parameters are not scaled.

FLUID SCALING

- We consider the case of many servers, namely $n \to \infty$.
- Accordingly, we let $\lambda_i^n = n\lambda_i$.
  μ_i, θ_i and the cost parameters are not scaled.

The scaled cost function:

$$J^n(T) = \frac{1}{nT}\mathbb{E}\int_0^T \sum_{i=1}^{I} c_i Q_i^n(t)dt \quad n, T \to \infty$$

FLUID SCALING

- We consider the case of many servers, namely $n \to \infty$.
- Accordingly, we let $\lambda_i^n = n\lambda_i$.
  μ_i, θ_i and the cost parameters are not scaled.

The scaled cost function:

$$J^n(T) = \frac{1}{nT}\mathbb{E}\int_0^T \sum_{i=1}^{I} c_i Q_i^n(t)dt \quad n, T \to \infty$$

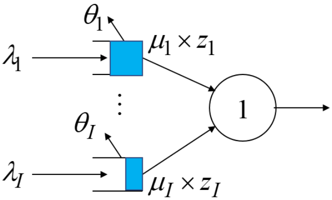
- Many-server fluid approximations of queueing systems with abandonments were studied, among others, by [Mandelbaym, Massey & Reiman 1998], [Whitt 2004] $(M/M/n+M)$.
  [Whitt 2006] suggested a heuristic model for the $G/GI/n+G$ queue.
- *Control* problems in the queueing regime were consdiered for example in [Bassamboo, Harrison & Zeevi 2007], who considered suboptimal routing and admission control policies that track the solution of the fluid model.

OUR PLAN

- Use a simplified fluid model to get some ideas for effective policies.
- Translate these policies to the original (stochastic) system.

THE FLUID MODEL

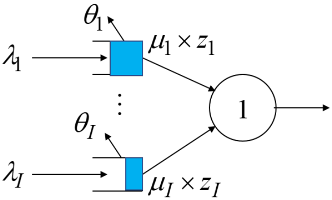
- Let us scale the arrival, departure and abandonment processes by $\frac{1}{n}$, assume that they are stationary, and focus on their rates. We arrive heuristically at the following static fluid model:



where $\sum_i z_i \leq 1$.

THE FLUID MODEL

- Let us scale the arrival, departure and abandonment processes by $\frac{1}{n}$, assume that they are stationary, and focus on their rates. We arrive heuristically at the following static fluid model:



  where $\sum_i z_i \leq 1$.

- Flow balance equations (with fixed queue lengths):

$$\lambda_i = z_i \mu_i + \theta_i q_i$$

  if $\lambda_i \geq z_i \mu_i$, otherwise $q_i = 0$.

THE FLUID LP PROBLEM

- Our optiimization problem:

$$\min_{\{z_i\}} \sum_i c_i q_i$$

s.t. $\lambda_i = \mu_i z_i + \theta_i q_i;\;\; z_i \geq 0,\; \sum_i z_i \leq 1;\;\; q_i \geq 0 \implies z_i \leq \frac{\lambda_i}{\mu_i}$

THE FLUID LP PROBLEM

- Our optiimization problem:

$$\min_{\{z_i\}} \sum_i c_i q_i$$

s.t. $\lambda_i = \mu_i z_i + \theta_i q_i; \;\; z_i \geq 0, \sum_i z_i \leq 1; \;\; q_i \geq 0 \implies z_i \leq \frac{\lambda_i}{\mu_i}$

- Substituting for q_i:

$$\sum_i c_i q_i = \sum_i c_i \frac{\lambda_i - \mu_i z_i}{\theta_i} = (\dots) - \sum_i z_i \frac{c_i \mu_i}{\theta_i}$$

- The solution now is obvious...

THE FLUID SOLUTION

- Renumber the classes in *decreasing* order of the index $\dfrac{c_i \mu_i}{\theta_i}$

THE FLUID SOLUTION

- Renumber the classes in *decreasing* order of the index $\dfrac{c_i \mu_i}{\theta_i}$
- Set
$$(z_1, \ldots, z_I) = (\tfrac{\lambda_1}{\mu_1}, \ldots, \tfrac{\lambda_{k-1}}{\mu_{k-1}}, z_k, 0, \ldots, 0)$$
where
$$k = \min\{j : \textstyle\sum_1^j \frac{\lambda_i}{\mu_i} > 1\},\ z_k = 1 - \textstyle\sum_1^{k-1} z_i$$

  This yields
$$(q_1, \ldots, q_I) = (0, \ldots, 0, q_k > 0, \tfrac{\lambda_{k+1}}{\theta_{k+1}}, \ldots \tfrac{\lambda_I}{\theta_I})$$

THE FULL INDEX

- Substituting $c_i \leftarrow c_i + \gamma_i \theta_i$ gives

$$\frac{c\mu}{\theta} \to \frac{(c+\gamma\theta)\mu}{\theta} = (\frac{c}{\theta} + \gamma)\mu$$

- Clearly, priority is given to customers with high waiting cost, long patience, high abandonment cost, and high service rate.

THE FULL INDEX

- Substituting $c_i \leftarrow c_i + \gamma_i \theta_i$ gives

$$\frac{c\mu}{\theta} \to \frac{(c+\gamma\theta)\mu}{\theta} = (\frac{c}{\theta} + \gamma)\mu$$

- Clearly, priority is given to customers with high waiting cost, long patience, high abandonment cost, and high service rate.

- We note that in [Ayesta, Jacko & Novak 2017], the same index (with some additional cost terms) is derived using the Whittle index for restless bandits.

Back to the Stochastic System:

The fluid solution suggests at least two different implementations in the stochastic system:

Back to the Stochastic System:

The fluid solution suggests at least two different implementations in the stochastic system:

- Fixed server assignment: Apply a fraction $\approx z_i^*$ of the servers to queue i.
  Advantages: $\surd$ Easy to implement $\surd$ General applicability

Back to the Stochastic System:

The fluid solution suggests at least two different implementations in the stochastic system:

- Fixed server assignment: Apply a fraction $\approx z_i^*$ of the servers to queue i.
  Advantages: $\surd$ Easy to implement $\surd$ General applicability
- Fix priority rule: Assign servers to waiting customers with the highest $\frac{c\mu}{\theta}$ index. (Preemptive or nonpreemptive.)

Back to the Stochastic System:

The fluid solution suggests at least two different implementations in the stochastic system:

- Fixed server assignment: Apply a fraction $\approx z_i^*$ of the servers to queue i.
  Advantages: $\surd$ Easy to implement $\surd$ General applicability
- Fix priority rule: Assign servers to waiting customers with the highest $\frac{c\mu}{\theta}$ index. (Preemptive or nonpreemptive.)
  Advantages:
  $\surd$ No server idleness
  $\surd$ Policy does not depend on (λ_i)

ASYMPTOTIC OPTIMALITY

- Denote by v^* the optimal value of the fluid LP problem.
- Recall that $J^{n,T}(\pi) = \frac{1}{nT}\mathbb{E}^\pi \int_0^T \sum_{i=1}^I c_i Q_i^n(t)dt$
- Let π^0 denote the $c\mu/\theta$ index policy (preemptive or non-preemptive).

ASYMPTOTIC OPTIMALITY

- Denote by v^* the optimal value of the fluid LP problem.
- Recall that $J^{n,T}(\pi) = \frac{1}{nT}\mathbb{E}^{\pi} \int_0^T \sum_{i=1}^I c_i Q_i^n(t)dt$
- Let π^0 denote the $c\mu/\theta$ index policy (preemptive or non-preemptive).
- We show first that

$$\liminf_{T\to\infty} \liminf_{n\to\infty} J^{n,T}(\pi_n) \geq v^*$$

  for any sequence $\{\pi_n\}$ of policies (not necessary stationary).
- We further show that for π^0 the limits exist and equal v^*.

[Atar, Giat & Sh. 2010]

ASYMPTOTIC OPTIMALITY

- Denote by v^* the optimal value of the fluid LP problem.
- Recall that $J^{n,T}(\pi) = \frac{1}{nT}\mathbb{E}^\pi \int_0^T \sum_{i=1}^I c_i Q_i^n(t) dt$
- Let π^0 denote the $c\mu/\theta$ index policy (preemptive or non-preemptive).
- We show first that

$$\liminf_{T\to\infty} \liminf_{n\to\infty} J^{n,T}(\pi_n) \geq v^*$$

  for any sequence $\{\pi_n\}$ of policies (not necessary stationary).
- We further show that for π^0 the limits exist and equal v^*.

  [Atar, Giat & Sh. 2010]

- We repeat the above for the ergodic cost: Limits taken in the opposite order.

  [Atar, Giat & Sh. 2011]

The Controlled Process and General Policies

- The processes involved (for given n - omitting the n superscript)
  - A_i, D_i, R_i: cumulative number of arrivals / service completions / reneging on $[0, t]$.
  - X_i, Q_i, Z_i: Number of jobs in the system / queue (unserved) / service at t; $X_i = Q_i + Z_i$.

The Controlled Process and General Policies

- The processes involved (for given n - omitting the n superscript)
  - A_i, D_i, R_i: cumulative number of arrivals / service completions / reneging on $[0, t]$.
  - X_i, Q_i, Z_i: Number of jobs in the system / queue (unserved) / service at t; $X_i = Q_i + Z_i$.
- Stochastic Primitives: $(\tilde{A}_i, \tilde{D}_i, \tilde{R}_i) \sim$ Independent Poisson processes, with rates $n\lambda_i$, μ_i, θ_i; and IC's $X_i(0)$.
- Define

$$D_i(t) = \tilde{D}_i(\int_0^t Z_i(s)ds), \quad R_i(t) = \tilde{R}_i(\int_0^t Q_i(s)ds)$$

The Controlled Process and General Policies

- The processes involved (for given n - omitting the n superscript)
  - A_i, D_i, R_i: cumulative number of arrivals / service completions / reneging on $[0, t]$.
  - X_i, Q_i, Z_i: Number of jobs in the system / queue (unserved) / service at t; $X_i = Q_i + Z_i$.
- Stochastic Primitives: $(\tilde{A}_i, \tilde{D}_i, \tilde{R}_i) \sim$ Independent Poisson processes, with rates $n\lambda_i$, μ_i, θ_i; and IC's $X_i(0)$.
- Define

$$D_i(t) = \tilde{D}_i(\int_0^t Z_i(s)ds), \quad R_i(t) = \tilde{R}_i(\int_0^t Q_i(s)ds)$$

item Additional relations:

$$X_i(t) = X_i(0) + (A_i - D_i - R_i)(t)$$

$$Q_i \geq 0, \quad 0 \leq Z_i \leq n$$

Policies

- A policy is now defined implicitly as any tuple
  $$\pi^n = (D_i^n, R_i^n, X_i^n, Q_i^n, Z_i^n)$$
  that satisfies the above-mentioned relations.
- The implied policies include history-dependent, non-stationry policies – and in fact also non-causal policies.

Policies

- A policy is now defined implicitly as any tuple
$$\pi^n = (D_i^n, R_i^n, X_i^n, Q_i^n, Z_i^n)$$
that satisfies the above-mentioned relations.
- The implied policies include history-dependent, non-stationry policies – and in fact also non-causal policies.

Fluid scaling:

- The time-dependent fluid model is obtained as the limit in $n \to \infty$ of the scaled processes $(\frac{1}{n}D_i^n, \frac{1}{n}R_i^N ....)$ (whenever the limits exist).
- When these processes converge to a constant, we obtain the static model discussed above.

Non-exponential Patience Distributions

General Patience Distributions

- Fluid models for (many-server) queues with abandonment and generally-distributed patience become more complicated, as they require measure-valued processes todescribe the (remaining) patience of customers in the queue.
- The fluid limit of a multiclass queueing system with $G/GI/n + GI$ queues under fixed priority policies was analyzed in Atar, Kaspi & Sh. (2014), extending the approach of Kaspi & Ramanan (2011), Kang & Ramanan (2012) to the multiclass case.

General Patience Distributions

- Fluid models for (many-server) queues with abandonment and generally-distributed patience become more complicated, as they require measure-valued processes todescribe the (remaining) patience of customers in the queue.
- The fluid limit of a multiclass queueing system with $G/GI/n + GI$ queues under fixed priority policies was analyzed in Atar, Kaspi & Sh. (2014), extending the approach of Kaspi & Ramanan (2011), Kang & Ramanan (2012) to the multiclass case.
- We outline here some initial results in the fluid model that pertains to the simpler $G/M/n + GI$ case, along with *nonlinear* holding costs.

Elements of the Fluid Model

- F_i is the patience distribution of class i, with hazard-rate function h_i.
- $X_i(t) = Q_i(t) + B_i(t)$ is the number of class-i customers in the system (# in queue + # in service).

Elements of the Fluid Model

- F_i is the patience distribution of class i, with hazard-rate function h_i.
- $X_i(t) = Q_i(t) + B_i(t)$ is the number of class-i customers in the system (# in queue + # in service).
- Cost function:

$$J_T(\pi) = \frac{1}{T} \sum_{i=1}^{I} \left[\int_0^T C_i\left(Q_i(s)\right) ds + \gamma_i R_i(T) \right].$$

where $C_i(q)$ is a no-decreasing holding cost function, and $R_i(T)$ is the number of abandonmens by time T.

Steady State Fluid Model

- For a given non-idling scheduling policy π, suppose $Q_i(t) \to q_i$, and $B_i(t) \to b_i$ (actually one implies the other).
- Then $0 \leq b_i \leq \lambda_i/\mu_i$, $\sum_{i=1}^{I} b_i = n$, and

$$q_i = \lambda_i \int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds$$

Steady State Fluid Model

- For a given non-idling scheduling policy π, suppose $Q_i(t) \to q_i$, and $B_i(t) \to b_i$ (actually one implies the other).
- Then $0 \le b_i \le \lambda_i/\mu_i$, $\sum_{i=1}^{I} b_i = n$, and

$$q_i = \lambda_i \int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds$$

- Therefore,

$$\lim_{T\to\infty} J_T(\pi) = \sum_{i=1}^{I} J_i(b_i)$$

where

$$J_i(b_i) = C_i\Big(\lambda_i \int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds\Big) + \gamma_i(\lambda_i - b_i\mu_i).$$

Fluid Optimization Problem

In terms of the steady state of the fluid model, we obtain the optimization problem:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{I} J_i(b_i) \\
\text{subject to} \quad & \sum_{i=1}^{I} b_i \leq n, \\
& 0 \leq b_i \leq \frac{\lambda_i}{\mu_i},\ i = 1, \ldots, I.
\end{aligned}
\tag{1}
$$

The decision variables b_i's can be intuitively understood as the amount of service resources that are assigned to class i customers in the long run.

Fluid Optimization Problem: The concave Case

- Suppose that the holding cost functions C_i is *concave*, and the patience hazard rate functions h_i are *nondecreasing*. Then the optimization problem is concave.

Fluid Optimization Problem: The concave Case

- Suppose that the holding cost functions C_i is *concave*, and the patience hazard rate functions h_i are *nondecreasing*. Then the optimization problem is concave.
- In that case the optimal solution is at the extreme point of the feasible region, which implies a fixed priority rule. In particular, there exists a fixed priority rule π^* such that each $B_i(t)$ converges to the optimal solution b_i^*.
- Hence, the average cost $J_T(\pi^*)$ converges to the optimal steady state solution as $T \to \infty$.

The convex Case

- Suppose that the holding cost functions C_i are *convex*, and the patience hazard rate functions h_i are *nonincreasing*. Then the optimization problem is convex.
- Assuming further strict convexity and an interior solution, the KKT optimality conditions for this problem imply

$$
\begin{aligned}
P_i(b_i) &:= \frac{C_i'\big(\lambda_i \int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds\big)\mu_i}{h_i(F_i^{-1}(1-b_i\mu_i/\lambda_i))} + \gamma_i\mu_i + \alpha_i\mu_i - \beta_i\mu_i \\
&= constant
\end{aligned}
$$

along with $\sum_i b_i = n$.

The convex Case - Generalized $c\mu/\theta$ rule.

- This motivates us to consider the following *dynamic* priority rule:

  At time t, assign priority in decreasing order of $P(B_i(t))$

The convex Case - Generalized $c\mu/\theta$ rule.

- This motivates us to consider the following *dynamic* priority rule:

  At time t, assign priority in decreasing order of $P(B_i(t))$

- Under this policy, each $B_i(t)$ converges to the optimal solution b_i^*.
- Hence, the average cost $J_T(\pi^*)$ converges to the optimal steady state solution as $T \to \infty$.

A General Priority Rule: The Target-Setting Policy

- Let (b_i^*) be an optimal solution of the steady-state optimization problem.
- Consider the time-varying priority rule

  At time t, assign priority in decreasing order of $P_i(t)$

  where

$$P_i(t) = b_i^* - B_i(t)$$

A General Priority Rule: The Target-Setting Policy

- Let (b_i^*) be an optimal solution of the steady-state optimization problem.
- Consider the time-varying priority rule

  At time t, assign priority in decreasing order of $P_i(t)$

  where
  $$P_i(t) = b_i^* - B_i(t)$$

- Then similar convergence properties hold, namely $B_i(t) \to b*$, and $J_T(\pi) \to J^*$.