Self-optimising state-dependent routing in parallel queues

Ilze Ziedins

Joint work with:
Heti Afimeimounga, Lisa Chen, Mark Holmes, Wiremu Solomon,
Niffe Hermansson, Elena Yudovina

The University of Auckland
Auckland, Monday, 8.30 a.m., predicted traffic
(downloaded 6 July 2013)
Auckland, Monday 10 June, 8.30 a.m., actual traffic
Which route/mode of transport to take?

- Individual choice (selfish routing) vs. social optimum
- User equilibrium vs. system optimum
- Probabilistic routing vs. state-dependent routing.
User equilibrium
Wardrop or user equilibrium

The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Wardrop, J.G. (1952)

Each user has an infinitesimal effect on the system.
Parallel queues

Network with collection $R$ of $N$ routes from $A$ to $B$.

Probabilistic routing – user optimal/equilibrium policies

$p_r$ = probability of taking route $r$, with $p_r \geq 0$, $\sum_r p_r = 1$.  
$p = (p_1, p_2, \ldots, p_N)$  
$W_r(p) =$ expected transit time via route $r \in R$.

At a user equilibrium, $p^{EQ}$, there exists $c$ such that

$$W_r(p^{EQ}) = c \quad \text{if } p_r^{EQ} > 0$$

$$\geq c \quad \text{if } p_r^{EQ} = 0.$$
State dependent routing – user optimal/equilibrium policies

A decision policy $\mathcal{D}$ is a partition of state space, $\mathcal{S}$, into sets $D_r, r \in R$ such that if system is in state $n \in D_r$ when a user arrives, then they take route $r$.

For a policy $D \in \mathcal{D}$ and $n \in \mathcal{S}$, $z_r^D(n)$ = expected time to reach the destination for a general user, if system is in state $n$ immediately prior to their arrival, and they choose to take route $r$.

A policy $D \in \mathcal{D}$ is a user optimal policy or user equilibrium if for each $n \in \mathcal{S}$

$$n \in D_r \implies z_r^D(n) \leq z_s^D(n) \text{ for all } s \neq r, s \in R.$$
Downs-Thomson network
Downs-Thomson network

$Q_1$: 1 server, $\mu_1$

$Q_2$: $\infty$ server, $\mu_2$

Two Poisson arrival streams – dedicated users to queue 2 at rate $\lambda_2$,
– general users at rate $\lambda$.

General users choose route – either probabilistic or state-dependent routing.

$Q_1$ single server queue (·/M/1), exponential service times, mean $1/\mu_1$.
$Q_2$ batch service $\infty$ server queue, service times with mean $1/\mu_2$.

Downs(62), Thomson(77), Calvert(97), Afimeimouna, Solomon, Z(05,10)
• Single server queue – private transportation (e.g. cars).
  – delay increases as load increases

• Batch service queue – public transportation (e.g. shuttle bus).
  – delay decreases as load increases
  – frequency of service increases as load increases

• This version of model first proposed by Calvert (1997) as queueing network version of transportation model that gives rise to the Downs Thomson paradox.

• Paradox is that delays for all users can increase when capacity of private transportation (roading) is increased. First observed by Downs (1962) and Thomson (1977).

• Afimeimounga, Solomon, Z (2005, 2010)
Downs-Thomson network –
probabilistic routing
$Q_1$: 1 server, $\mu_1$

$Q_2$: $\infty$ server, $\mu_2$

$Q_1$ single server queue ($\cdot/M/1$). Expected delay $W_1 = \frac{1}{\mu_1 - \lambda p}$

$Q_2$ batch service $\infty$ server queue. Expected delay $W_2 = \frac{1}{\mu_2} + \frac{N-1}{2(\lambda_2 + \lambda(1-p))}$

Both $W_1$ and $W_2$ are increasing in $p$. 
$\mu_1 = 0.8$

$\lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3$

$W_1, \ldots, W_2, \ldots$
\[ \mu_1 = 0.8 \]
\[ \lambda = 1, \lambda_2 = 0.1, \mu_2 = 1, N = 3 \]
\[ W_1, \quad W_2, \quad \]
\[ \begin{align*} 
\mu_1 &= 0.8 \\
\lambda &= 1, \ \lambda_2 = 0.1, \ \mu_2 = 1, \ N = 3
\end{align*} \]

\[ W_1, \cdots, \cdots, W_2, \cdots, \cdots, \cdots \]
$\mu_1 = 0.8$

$\lambda = 1, \lambda_2 = 0.1, \mu_2 = 1, N = 3$

$W_1, \cdots, W_2, \cdots$
\[ \mu_1 = 0.8 \]

\[ \lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3 \]

\[ W_1, \quad - - - - - - - - , \quad W_2, \quad \underline{\rule{5cm}{0.1pt}} \]
\[ \mu_1 = 0.8 \]
\[ \lambda = 1, \lambda_2 = 0.1, \mu_2 = 1, N = 3 \]
\[ W_1, \quad \quad \quad \quad \quad \quad \quad \quad W_2, \]
\[ \mu_1 = 0.8, 0.95 \]
\[ \lambda_1 = 1, \lambda_2 = 1, \mu_2 = 1, N = 3 \]
\[ W_1, \quad W_2, \]
\[ \mu_1 = 0.8, 0.95 \]
\[ \lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3 \]
\[ W_1, - - - - - - - - , W_2, \]
\[ \mu_1 = 0.8, 0.95, 1.05 \]
\[ \lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3 \]
\[ W_1, \quad \quad \quad \quad \quad W_2, \quad \quad \quad \quad \quad \]
\[
\mu_1 = 0.8, 0.95, 1.05 \\
\lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3 \\
W_1, \quad W_2,
\]
\[
\lambda = 1, \lambda_2 = .1, \mu_2 = 1, N = 3
\]
\[
W = p^{EQ}W_1 + (1 - p^{EQ})W_2
\]
\[ \lambda = 1, \lambda_2 = 0.1, \mu_2 = 1, N = 3 \]

\[ W = p^{EQ}W_1 + (1 - p^{EQ})W_2 \]

\[ W = \min_p pW_1 + (1 - p)W_2 \]
Consequences of individual choice

- Network performance may be poorer than expected
- Adding capacity may lead to worse performance
Downs-Thomson network –
state dependent routing
State dependent policies

\[ X_1(t) = \text{number of customers in queue 1} \]
\[ \text{(including customer in service)} \]

\[ X_2(t) = \text{number of customers waiting for service in queue 2} \]
\[ \text{(not including those in service)} \]

State space \( S = \mathbb{Z}_+ \times \{0, 1, 2, \ldots, N - 1\} \).

Process \( X_D \) operating under decision policy \( D \) has transition rates:

\[
\begin{align*}
\mathbf{n} \rightarrow \begin{cases} 
\mathbf{n} - e_1 & \text{at rate } \mu_1 I_{\{n_1 > 0\}} \\
\mathbf{n} + e_1 & \text{at rate } \lambda I_{\{n \in D_1\}} \\
(n_1, (n_2 + 1) \mod N) & \text{at rate } \lambda_2 + \lambda I_{\{n \in D_2\}}
\end{cases}
\end{align*}
\]

where \( I_A = 1 \) if \( A \) occurs, and \( I_A = 0 \) otherwise.

A policy \( D \in \mathcal{D} \) is a user optimal policy or user equilibrium if

\[ \mathbf{n} \in D_1 \iff z_1^D(n) < z_2^D(n) \quad \text{for all } \mathbf{n} \in S. \]
Points in $D_1$ – ●. Points in $D_2$ – ○.

Unique user optimal policy for

$N = 10, \lambda = 1.5, \lambda_2 = 0.5, \mu_1 = 2, \mu_2 = 1$.

A policy $D \in \mathcal{D}$ is monotone if $D$ satisfies

(A) $n \in D_2 \Rightarrow n + e_1 \in D_2$ for all $n \in S$ and

(B) $n \in D_2 \Rightarrow n + e_2 \in D_2$ for all $n \in S$
Properties

- A user optimal policy exists and is unique (no randomization needed).
- The user optimal policy is monotone.
- The user optimal policy is monotone in the parameters $\lambda$, $\lambda_2$, $\mu_1$, $\mu_2$ in the following sense. Let $X^{(1)}$ and $X^{(2)}$ be two processes, with common batch size $N$ and user optimal policies $D^*(1)$, $D^*(2)$ respectively. If $\lambda^{(1)} \geq \lambda^{(2)}$, $\mu_1^{(1)} \leq \mu_1^{(2)}$, $\lambda_2^{(1)} \geq \lambda_2^{(2)}$ and $\mu_2^{(1)} \geq \mu_2^{(2)}$, then $D^*_1(1) \subset D^*_1(2)$.
- Proof uses a coupling argument.
- As part of the proof show monotonicity of $z_2^D(n)$ in $\lambda$, $\lambda_2$, $\mu_1$, $\mu_2$; and in the decision policy.
Expected transit times under user optimal policy for state-dependent routing (———–), and probabilistic routing (−−−−−−−).

λ = 1, λ₂ = 0.1, μ₂ = 1, N = 3 for 0 ≤ μ₁ ≤ 3.
Expected transit times under user optimal policy for state-dependent routing (———–), and probabilistic routing (−−−−−−−)

\(\lambda = 1, \lambda_2 = 0.1, \mu_2 = 1, N = 3\) for \(0 \leq \mu_1 \leq 3\).
Variations
Two batch-service queues

Expected transit times under user optimal policy for state-dependent routing (-----), and probabilistic routing (-----)

\[ \lambda = 4, \lambda_1 = 3, \lambda_2 = 1, \mu_2 = 2, N_1 = N_2 = 5 \] for \( 0 \leq \mu_1 \leq 6 \).

Chen, Holmes, Z(2011)
Other variations

Processor-sharing queues

- Iterative procedure may converge to periodic orbit
- User equilibrium doesn’t always possess monotonicity properties
- Randomization needed

Braess’s paradox

- State dependent routing mitigates worst effects here as well
  Cohen, Kelly (1990), Calvert, Solomon, Z (1997)
Some final comments

• Do user equilibria exist more generally under state dependent routing, and if yes, when are they unique?

• How to overcome poor performance at user equilibria?

• Does more information lead to shorter delays in general? Effects of partial information

• Add monetary and other costs to the problem, as well as delays

• Convergence issues – effect of delayed information.

• Differing information and/or policies for different customer classes
  Argument for investment in public transport, using public transport


