CYBER-PHYSICAL CLOUD COMPUTING LAB
UNIVERSITY OF CALIFORNIA, BERKELEY

Management of Spatial Queuing Systems: The Case of Gated Policies

Jiangchuan Huang and Raja Sengupta

Working Papers
CPCC-WP-2012-07-02

July 2012

Management of Spatial Queuing Systems: The Case of Gated Policies [*]

Jiangchuan Huang
Systems Engineering Group
Civil and Environmental Engineering
University of California at Berkeley
Berkeley, CA 94720
jiangchuan@berkeley.edu

Raja Sengupta
Systems Engineering Group
Civil and Environmental Engineering
University of California at Berkeley
Berkeley, CA 94720
sengupta@ce.berkeley.edu

ABSTRACT

We discuss the management of m-vehicle spatial queuing systems based on the mathematical results introduced in [10, 11] extending the m-vehicle Dynamic Traveling Repairman Problem (m-DTRP) introduced by Bertsimas et al. [2]. The m-DTRP is focused on the mean response time (E) for tasks. We consider the reduction of this measure a management objective but also include other considerations such as the variance of response time, its distribution, system predictability and transparency as management objectives. We present computational techniques to approximate the distribution of response time, compute the distributions for gated polling policies with task sequencing being done by several policies in the literature. The distributions of response time are then used to compute the $E-V-$ load profiles for the different policies. These profiles provide the basis for management of the spatial queuing system. We show first how to estimate the probability a given task will complete by a certain time, conditioned on the system state, so that a customer can know what to expect of the system at any given time. We think of this as predictability or transparency. We study how resource parameters such as service time, vehicle speed, and the number of vehicles affect the expected value and variance of response time. Multiple vehicles are handled by generating a Voronoi tessellation of the operating region. Finally, we model agencies by Cobb-Douglas utility functions and show how such a model enables one to choose the spatial queuing policy appropriate to an agency.

Keywords

$E-V$ analysis, queuing theory, DTRP, Economy of Scale, Administrative and resource control

1. INTRODUCTION

We extended in [10, 11] the results of the m-vehicle Dynamic Traveling Repairman Problem (m-DTRP) introduced by Bertsimas et al. [2] by augmenting the performance measure of this problem from the expectation of the response time, T, defined as the difference between task completion time and arrival time, to the expectation and variance ($E-V$) of T and the distribution of T when possible. We called this extension the $E-V$ analysis of the m-Vehicle Spatial Queuing Problem (m-SQP) and showed that it is possible.

Scheduling policies are necessary when we need to allocate tasks to resources. When there are more tasks than resources (servers or vehicles) in a certain horizon, a queue builds up. Usually tasks can be substantially variable in their sizes, arrival times, and locations due to incompleteness of processing information or the inherent process variability [5], the task flow features in time [15], and physical location requirements in space [18]. These uncertainties make the response time highly variable under a scheduling policy. Sarin et al. [20, 21] stated the impact of uncertainty in scheduling and the need for efficient modeling of stochastic scheduling problems, as well as the need to devise effective scheduling strategies to counter the impact of uncertainty (or variability). They focused on expected value and the variance of a performance measure of evaluation, e.g. T. The scheduling environments they considered include scheduling on sigle (or parallel) machine(s) and permutation flow shops / job shops with unlimited intermediate storage. Srirangacharyulu et al. [22] considered the problem of minimizing the completion time variance of n jobs in single and multi-machine systems with deterministic processing times. De et al. [5] defined stochastic optimality and stochasitic efficiency in terms of the entire probability distribution of the performance measure of interest, and examined the identification of expectation-variance efficient job sequences in the flow-time problem introduced in [1].

We seek the equivalent of this literature for the m-vehicle spatial queuing problem and focus on the management of spatial queuing systems based on the expected value of task response time, its variance, system predictability and transparency. All of these are management objectives. We present computational techniques to approximate the distribution of response time, compute the distributions for gated policies with task sequencing being done by several policies in the

[*]This work was partly supported by NSF-CNS-1136141. The authors would like to thank Prof. Carlos F. Daganzo in Transportation, Civil and Environmental Engineering at UC Berkeley.

literature. The distributions of response time are then used to compute the $E - V -$ load profiles for the different policies. These profiles provide the basis for management of the spatial queuing system. We show first how to estimate the probability a given task will complete by a certain time, conditioned on the system state, so that a customer can know what to expect of the system at any given time. We think of this as predictability or transparency. We study how resource parameters such as service time, vehicle speed, and the number of vehicles affect the expected value and variance of response time. Multiple vehicles are handled by generating a Voronoi tessellation of the operating region. Finally, we model agencies by Cobb-Douglas utility functions and show how such a model enables one to choose the spatial queuing policy appropriate to an agency. This approach is borrowed from financial economics.

In financial economics, the application of expectation-variance analysis is quite common [9], [14]. Markowitz [13] and Elton et al. [7] used variance or standard deviation of return as a measure of risk. Accordingly, we explore $E[T] - Var[T]$ analysis for the m-SQP considering that the mean response time, $E[T]$, and the risk, measured by $Var[T]$, are in many respects related with the utility of each individual agency. We show in Section 4.3 that two policies at the same load level can be incomparable in the sense that one has high reward and high risk while the other has low reward but also low risk. Different agencies will have different preferences over values of $E[T]$ and $Var[T]$. A risk tolerant agency, e.g. a best-effort package delivery company, might find a policy with low E and relatively high V preferable while a risk averse agency, e.g. a UAV system in military use, might prefer to reduce V even if it means increasing E. We adopt the concept of *Utility Indifference Curves* (*UIC's*) [17], which are the contours of the utility function, to model the behavior of an agency. In general, each individual agency i) will prefer lower to higher $Var[T]$ for a given level of $E[T]$, and ii) will only accept a greater $Var[T]$ if compensated in the form of less $E[T]$. Thus we assume an agency is modeled by a set of UIC's, with each UIC showing different bundles of $E[T] - Var[T]$ or $1/E[T] - 1/Var[T]$ between which the agency is indifferent. One can equivalently refer to each point on the indifference curve as rendering the same level of utility (satisfaction) for the agency.

Consumer preference theory assumes preferences are *complete*, *reflexive*, *continuous*, *strongly monotonic*, and that UIC's exhibit *diminishing marginal rates of substitution* [3]. One utility function satisfying this is the well-known Cobb-Douglas utility function $U(y_1, y_2) = y_1^{\epsilon} y_2^{(1-\epsilon)}$ [6]. We assume $y_1 = \frac{1}{\sigma[T]}$ and $y_2 = \frac{1}{E[T]}$. $\epsilon \in (0,1)$ and $1-\epsilon$ are the *output elasticities* of $\frac{1}{\sigma[T]}$ and $\frac{1}{E[T]}$, respectively. ϵ models the risk aversion of an agency in our setting. The greater the value of ϵ, the more risk tolerant the agency. An agency satisfying consumer preference theory will always chose the $E - V$ that renders the highest utility, i.e., the $E - V$ contained in the UIC furthest away from the origin. Figure 1 shows the UIC's for $\epsilon = 0.5$. We see $UIC_3 \succ UIC_2 \succ UIC_1$, where $UIC_2 \succ UIC_1$ means UIC_2 is chosen when both UIC_1 and UIC_2 are available. We show how to use the $E - V$ of response time of different policies to help an agency select a policy appropriate to their own utility.

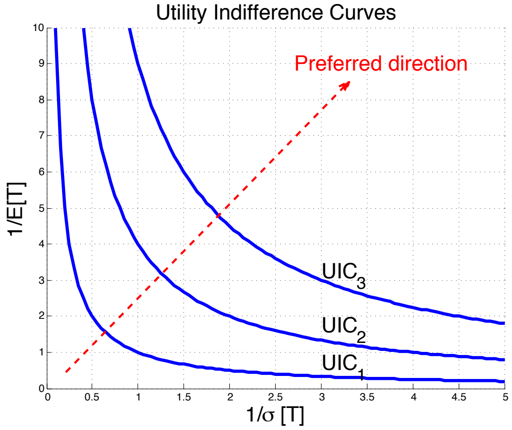

Figure 1: Utility Indifference Curves.

1.1 Problem Statement

For the convenience of the reader, we restate the definition of the m-SQP here: A convex region $\mathbf{A}$ of area A contains m vehicles (servers) that travel at constant speed v between task locations. Tasks arrive according to a Poisson process with rate λ and have a location that is independent and identically distributed (i.i.d.) according to the pdf $f_X(x)$ within $\mathbf{A}$. Each task i requires an i.i.d. execution time B, with mean b, which is assumed to be finite. Define load $\rho = \lambda b$. The system time of task i, denoted T_i, is defined as the elapsed time between the arrival of task i and the time task i is completed. When the system is stable, T_i converges to some T in distribution. T is called the steady state system time. Similarly, we can define the steady state waiting time, W, service time, T_S, and travel time to serve a task T_D. Intuitively, $T_S = T_D + B$ and $T = W + T_S$. Our aim is to find some verifiable stability condition, derive the mean and variance of T for some reasonable scheduling policies, and seek the distribution of T where possible.

To the best of our knowledge, no analytical study has yet addressed the expressions of $E[T]$ and $Var[T]$ except some asymptotic results in [16], or the distribution of T for spatial queues. This is in sharp contrast to classical queues where both first and second moments are known for a wide variety of policies, e.g., First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS), Random-Order-of-Service (ROS), Shortest-Job-First (SJF) and Longest-Job-First (LJF) [24, 25]. In spatial queues $E[T]$ and $Var[T]$ are only known for the FCFS policy by substituting the task size B with $D + B$ in the expressions of $E[T]$ and $Var[T]$ of FCFS in classic queueing theory, where $D \equiv \|X_1 - X_2\|$, where X_1 and X_2 are two independent random points in region $\mathbf{A}$. The expected value of D and higher moments between two points uniformly distributed in a square of area A are given in Larson et al. [12], p.135 as:

$$E[D] \approx 0.52\sqrt{A}, E\left[D^2\right] = \frac{1}{3}A \tag{1}$$

1.2 Main Results

In [10, 11] we showed that $E - V$ analysis is possible for the m-SQP. In this paper we discuss the management of spatial queueing systems based on $E - V - \rho$ profile and distribution of T including predictions, administrative and resource control, and policy selection in Section 4. We predict in Section 4.1 the completion time of tasks based on four senarios depends on the arrival scenarios of the tasks. The effectiveness of increasing serving rate, vehicle speed and vehicle numbers to decrease $E[T]$ and $Var[T]$ is discussed in Section 4.2 and concluded in Properities *P1 - P4*. We also show in Section 4.3 that $E-V-\rho$ profile enables us to design good policies adaptive to the utility of each individual agency at different load levels ρ.

2. ALLOCATION-SEQUENCING POLICIES

This section is a review of our previous work in [10, 11]. We broke the m- SQP into m 1-SQP's through an m-center Voronoi tessellation and allocate a single vehicle to each Voronoi cell. We then decomposed the policy design problem into allocation and sequencing by defining a spatial equivalent of polling [23, 8] for the allocation phase. We called this the m-A-S class, which is broken into m 1-A-S class through Voronoi tessellation. For the convenience of the readers, we restate the formulation of the 1-A-S class here. The spatial polling system contains a single vehicle and r partitions with infinite capacities. Partition k has area $A^k, k=1,2,\ldots,r$, $A=\sum_{k=1}^{r} A^k$. The vehicle visits the partitions in cyclic order, $1,2,\ldots,r,1,2,\ldots$, and serves the queue in each partition. Without loss of generality, we assume that the vehicle is initially at partition 1. Thus, the n-th queue that the vehicle visits is in partition $I(n)=(n-1) \pmod r +1$, where $n \pmod r$ means the remainder of the division of n by r.

We denote by:
$G^k(N)$ the number of tasks that are served in partition k when the queue is of length N.
$T_S^k(N)$ the total service time of N tasks in partition k.
$S^k(N)$ the duration of the service in partition k when the queue is of length N.

$$S^k(N) = T_S^k\left(G^k(N)\right) \tag{2}$$

The *switch time* the vehicle takes from a random point in partition k to a random point in partition $k+1$ is denoted by $\Delta^k, k=1,\ldots,r-1$. The value from partition r to partition 1 is denoted by Δ^r. $\Delta^k, k=1,\ldots,r$ are bounded above by the diameter of the region $\mathbf{A}$ divided by the speed of the vehicle, v. The first moments of Δ^k is denoted by $\delta^k, k=1,\ldots,r$. Let $\Delta=\sum_{k=1}^{r}\Delta^k$ be the total switch time in a cycle and denote by δ the first moment of Δ.

The tasks arrive at partition k with a Poisson process of parameter $\lambda^k=\int_{A^k} f(x)\,dx\lambda$, then $\lambda=\sum_{k=1}^{r}\lambda^k$. Define $\rho^k=\lambda^k b$, then $\rho=\sum_{k=1}^{r}\rho^k$, $1\le k\le r$. Let $N^k(t_1,t_2]$ denote the number of Poisson arrivals to partition k, $1\le k\le r$, during a (random) time interval $(t_1,t_2]$. $N^k(t)\equiv N^k(0,t]$ is the number of arrivals in a time interval of length t.

The n-th value of the polling system is described by the random variables N_n^k, $1\le k\le r$, $n\ge 1$, where N_n^k represents the number of tasks in partition k when the vehicle arrives at the n-th queue. Let $N_n=\left(N_n^1,\ldots,N_n^r\right)$, taking values in $\mathbf{N}^r$, where $\mathbf{N}$ is the set of nonnegative integers. Denote by N_n^{Σ}, *the cycle number*, the number tasks served in region $\mathbf{A}$ when the vehicle arrives at the n-th queue. $N_n^{\Sigma}=\sum_{k=1}^{r}N_n^k$. N_n evolve according to the following evolution equations:

$$N_{n+1}^k = \begin{cases} N_n^k + N^k\left(S_n\right), & \text{if } I(n)\neq k \\ N_n^k - G^k\left(N_n^k\right) + N^k\left(S_n\right), & \text{if } I(n)=k \end{cases} \tag{3}$$

Denote by S_n, the *station time*, the time interval between the arrival times of the vehicle to the n-th queue and the $(n+1)$-st queue. Denote by C_n, the *cycle time*, the time interval between two successive arrivals of the vehicle to the same partition. $C_n=S_n+\ldots+S_{n+r-1}$.

$$S_n = S^{I(n)}(N_n^{I(n)}) + \Delta^{I(n)} \tag{4}$$

THEOREM 1. *In each Voronoi cell, the sequence $\{N_n\}_{n=0}^{\infty}$ is a Markov chain. Moreover, $\{N_{nr+k}\}_{n=0}^{\infty}$ is an homogeneous, irreducible and aperiodic Markov chain with state space $\mathbf{N}^r$.*

Definition 1. An allocation policy is called an Unlimited-Polling policy if it is a Polling policy and its $G^k(.)$ satisfies $G^k(N)\to\infty$, when $N\to\infty$. In particular, it is called a Gated-Polling policy if $G^k(N)=N$.

The Spatial-Polling policies give a finite set of tasks for the vehicle in a finite time horizon. The set of tasks can be sequenced based on their arrival time or location. Common policies include FCFS, SJF, Nearest Neighbor (NN) and Traveling Salesman Policy (TSP). For a set of N tasks, the expected travel time and expected service time to serve the N tasks under sequencing policy P are denote by $T_D^P(N)$ and $T_S^P(N)$, respectively. $T_D^k(N)\equiv T_D^P(N)$ and $T_S^k(N)\equiv T_S^P(N)$ when partition k is served under policy P. Rigorous definitions of $T_D^P(N)$ and $T_S^P(N)$ can be found in [10, 11].

$$T_S^P(N) = T_D^P(N) + Nb \tag{5}$$

Definition 2. A sequencing policy P is said to have economy of scale (EoS) if i) $T_D^P(n)$ is nondecreasing in n. ii) $T_D^P(n+1)-T_D^P(n)$ is nonincreasing in n.

Definition 3. A policy for an m-SQP is called an m-A-S policy if it allocates a single vehicle to each Voronoi cell of region $\mathbf{A}$ that is *equitable* with respect to $f(.)$ and $f^{\frac{1}{2}}(.)$, and runs the A-S policy independently in each Voronoi cell.

THEOREM 2. *(Stability theorem): For any m-A-S policy, if each of its A-S policy satisfies Definition 1 (Unlimited-Polling) in the allocation phase and Definition 2 (EoS) in the sequencing phase, when $\rho+\lambda b_d<m$, for all $1\le k\le r$ the Markov chains in each cell $\{N_{nr+k}\}_{n=0}^{\infty}$ are ergodic, thus $\{N_{nr+k}\}_{n=0}^{\infty}$ together with the sequence of station times $\{S_{nr+k}\}_{n=0}^{\infty}$ and the cycle times $\{C_{nr+k}\}_{n=0}^{\infty}$ have stationary distributions.*

When the system is stable, $\{N_n\}_{n=0}^{\infty}$ has stationary distribution. Usually finding the stationary distribution of N_n is difficult, we obtained in [10, 11] the stationary distribution of N_n under *gated policies*. A one-partition A-S policy with allocation policy satisfying *Definition 1* (Gated-Polling) is called a gated policy. In a system with gated policies [24], an arriving task that finds the server idle causes a gate to close. When this task service is completed, the gate opens and admits into a waiting room all the tasks that arrived during the service time and then closes. When all the tasks in the waiting room have been served, the gate opens and admits into the waiting room all the tasks that have arrived during the collective service times of the preceding group of tasks, after which it closes. The process continues in this manner. Inside each gate, different sequencing policies apply, e.g. FCFS, TSP, thus we have a class of gated policies.

Now N_n becomes the number of tasks served in the n-th gate. $S_n = C_n = S^1\left(N_n\right) = T_S^P\left(N_n\right)$, where we include Δ^1 in $T_S^P\left(N_n\right)$. Since there is only one partition, Δ^1 is the time the vehicle travel in the first segment of the tour serving the N_n tasks, where the first segment connects the start position of the vehicle to the location of the first task served. When the system is stable, $S_n \rightarrow S$, where S is the steady state gate duration.

$\Pr\left(N(t) = j\right) = \frac{(\lambda t)^j}{j!} e^{-\lambda t}$ is the probability that j tasks arrive during time interval $(0, t]$ according to a Poisson process. The transition probability of $\{N_n\}_{n=1}^{\infty}$ is

$$p_{ij} = \begin{cases} 1, & i = 0, j = 1 \\ 0, & i = 0, j \neq 1 \\ \int_0^\infty \Pr\left(N(t) = j\right)\, dF_{T_S^P(i)}(t), & i > 0 \end{cases} \qquad (6)$$

In general, $\int_0^\infty \Pr\left(N(t) = j\right)\, dF_{T_S^P(i)}(t)$ cannot be integrated analytically because $F_{T_S^P(i)}(t)$ is usually complicated.
If $T_S^P(i)$ is Gamma distributed, i.e. $T_S^P(i) \sim \Gamma\left(\alpha, \beta\right)$,

$$f_{T_S^P(i)}(t) = \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)} \qquad (7)$$

where α and β are functions of i.
Then when $i > 0$, $p_{ij} = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)}\, dt$
$= \frac{\lambda^j \beta^k}{j!\Gamma(\alpha)} \int_0^\infty t^{j+\alpha-1} e^{-(\lambda+\beta)}\, dt = \frac{\Gamma(j+\alpha)\lambda^j \beta^\alpha}{j!\Gamma(\alpha)(\lambda+\beta)^{j+\alpha}}$,
the stationary distribution of N can be computed numerically in this case.

3. THE DISTRIBUTION OF T AND $E - V$ ANALYSIS

We used Gamma distribution to approximate the distribution of the service time $T_S^P(i)$ through the Kolmogorov-Smirnov (K-S) test for goodness of fit in [10, 11], and showed how to obtain the pdf of S and T by numerical integration under this assumption. Here we check the Gamma hypothesis for the TSP, NN, and DA sequencing policies by simulation. Tables 1, 2 and 3 summarize the simulation results.

The simulations are done in a square of size 1 × 1. The sample size for each simulation (row) is 10,000. #pt is the number of points in the region for each row. E and V are the expectation and variance of the tour length connecting #pt points. $\eta = \frac{E}{\sqrt{\#pt}}$. α and β are the two parameters in the Gamma distribution in (7). $h = 1(0)$ means the null hypothesis that the tour length is distributed according to $\Gamma\left(\alpha, \beta\right)$ is (not) rejected at the 5% level based on the K-S test. p is the asymptotic P-value. #s in Table 3 is the number of swaths in DA [4, 19]. We were unable to compute the TSP tours for 432 and 363 points on our 4 core MacBook pro used for the simulations.

The Gamma distribution hypothesis is not rejected at the 5% level based on the K-S test when the number of points is from 12 to 432 for the NN, from 65 to 243 for the exact TSP by Mathematica and from 27 to 432 for the DA.

Table 1: Tour Distribution under TSP

#pt	E	V	η	α	β	h	p
243	12.1	.061	.775	2383	197.3	0	.15
192	10.8	.063	.778	1872	172.8	0	.26
147	9.5	.064	.781	1408	147.8	0	.32
108	8.2	.066	.785	1027	124.7	0	.09
75	7.0	.070	.792	684.5	98.5	0	.05
70	6.7	.070	.793	641.6	95.3	0	.13
65	6.5	.071	.794	591.3	90.9	0	.23
48	5.6	.064	.790	483.0	86.4	1	.00
27	4.4	.090	.811	209.7	47.9	1	.00
12	3.1	.111	.818	84.3	27.3	1	.00

Table 2: Tour Distribution under NN

#pt	E	V	η	α	β	h	p
432	18.8	.286	.903	1235	65.81	0	.13
363	17.2	.279	.904	1066	61.87	0	.25
300	15.7	.271	.905	908.4	57.95	0	.57
243	14.1	.274	.906	729.3	51.62	0	.76
192	12.6	.263	.907	600.2	47.77	0	.68
147	11.0	.257	.908	471.7	42.86	0	.80
108	9.4	.251	.907	355.0	37.65	0	.97
75	7.8	.240	.905	256.7	32.74	0	.92
70	7.6	.240	.905	239.0	31.56	0	.77
65	7.3	.239	.904	222.7	30.57	0	.27
48	6.2	.226	.899	172.0	27.61	0	.89
27	4.6	.213	.882	98.9	21.57	0	.99
12	2.9	.174	.829	47.1	16.40	0	.52

Table 3: Tour Distribution under DA

#pt	#s	E	V	η	α	β	h	p
432	12	19.4	.141	.931	2666	137.7	0	.86
363	11	17.7	.144	.931	2189	123.3	0	.85
300	10	16.1	.143	.932	1819	112.7	0	.95
243	9	14.5	.147	.933	1442	99.2	0	.31
192	8	12.9	.149	.933	1120	86.7	0	.87
147	7	11.3	.148	.933	865	76.4	0	.50
108	6	9.7	.153	.933	613	63.3	0	.99
75	5	8.1	.152	.931	426	52.8	0	.64
48	4	6.4	.162	.926	254.6	39.7	0	.76
27	3	4.7	.167	.909	133.3	28.2	0	.86
12	2	3.0	.173	.858	50.5	17.0	1	.04

Figure 2 shows the distribution of T for the exact TSP sequencing policy for different load levels. The operating region A is square with unit area. The task size B is $Unif\left[0, 1\right]$, i.e., uniformly distributed. The region is served by a vehicle with speed $v = 1$. These assumptions are made for all numerical results in this paper unless stated otherwise. We

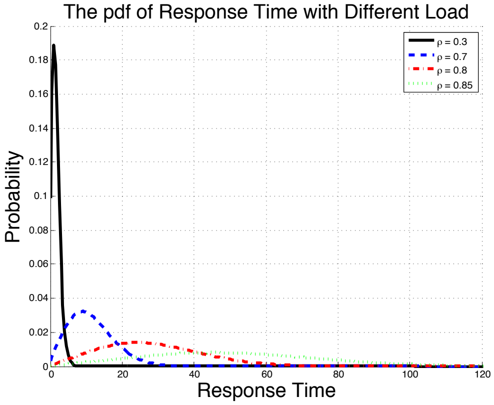

Figure 2: The pdf of T under Gated-TSP.

are now able to compute the $E - V - \rho$ profiles for different policies using the Gamma approximation to the pdf's in Figure 2. The results of computation are in Figure 3.

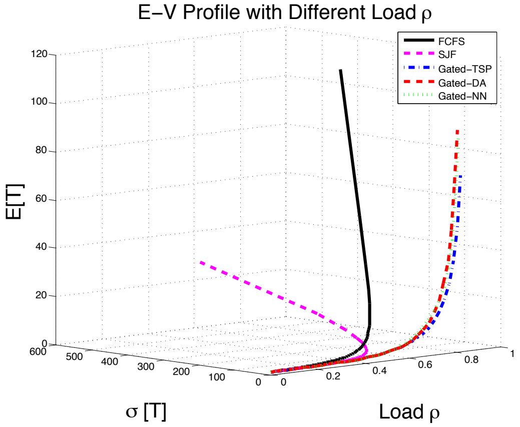

Figure 3: $E - V - \rho$ Profile for Different Policies.

4. MANAGEMENT

4.1 Predictions

In this section we discuss the problem of predicting the completion time of a task as a way of managing spatial queuing systems to be predictable and transparent. A client presenting a task to a predictable system should be able to ask a question like - What is the probability my task will complete before time t? Thus if T_C is the completion time of a task, we consider a spatial queuing system to be predictable or transparent if the probability completion time is less than some t, conditioned on the state of the spatial queuing system is computable. If the $Pr(T_C < t|\text{system state})$ is computable then one can also compute the t that would correspond to this probability being equal to some high value such as 0.95 or 0.99. This is equivalent to the system providing its customer, the task owner, with a time, by which the task will complete with high probability. This we formulate the predictable or transparent management problem as the problem of computing $Pr(T_C < t|\text{system state})$.

Task completion time is the arrival time plus response time. The task arrival time should be known deterministically to task owner and system, meaning the problem of computing $Pr(T_C < t|\text{system state})$ is equivalent to the problem of computing $Pr(T < t|\text{system state})$, where T is the task response time. Section 3 described the computation of $Pr(T < t)$. Thus here we focus on the problem of defining and conditioning on the system state.

We think of the life of a task in the following phases.

1. Pre-arrival: The task does not yet have an arrival time. Without a specific time, there is no system state. Thus a client can only ask for $Pr(T < t)$ as determined by the stationary distribution of T.

2. Arrived and outside a Gate: In general some gate is in progress when a task arrives. If so the response time T is the sum of W_O, the waiting time outside the current gate, and W_I, the waiting time inside the next gate. Since the spatial queuing process is Markovian, the system state is the number of tasks remaining to be served in the current gate, the sizes of there remaining tasks, the current position of the server, and the sequence in which these tasks are to be served. If we neglect uncertainties in vehicle motion, W_O is deterministically a function of these quantities and hereafter denoted w_O. W_I would be uncertain because the set of tasks arriving in time w_O is uncertain. Thus to compute the distribution of $T = w_O + W_I$ we need to compute only the distribution of W_I. This is discussed later in this section.

3. Arrived and inside a Gate and waiting to be served: In this phase, the current gate has closed meaning the set of tasks in the gate together with their locations and sizes, the order in which they are to be served, and the current position of the server are all known. If we neglect uncertainties in vehicle motion then the task completion time would be deterministically a function of these quantities plus the current time.

4. Arrived and inside a Gate and being served: Here the task completion time would be deterministically a function of the task size and the current time since both quantities are known.

We are now left with the problem of computing the distribution of W_I conditioned on w_O. If the task of interest arrived at a time t_0, we would know the number of tasks arrived outside the current gate up until t_0, say n_0, deterministically. The number of tasks that will arrive during w_O, say N_0^+, will then have the distribution $\Pr\left(N_0^+ = j\right) = \frac{(\lambda w_O)^j}{j!} e^{-\lambda w_O}$. For the TSP sequencing policy, the duration of the next gate is then $T_S^P\left(n_0, N_0^+\right)$ which denotes the time corresponding to a TSP tour of the tasks n_0 and N_0^+. As per section 3 this is approximately Gamma distributed. We do not know whether the task of interest would be at beginning of the tour, its end, or at some intermediate position in the sequence. We propose that W_I be estimated at time t_0 by assuming all of these cases are equally likely. Since the positions of task are also uniformly distributed over the operating area, we conjecture that assuming W_I to be uniformly distributed over the time interval $T_S^P\left(n_0, N_0^+\right)$ is an adequate approximation to reality. We propose to check this by simulation in the future.

4.2 Administrative and Resource Control

In this section we use the response time T distributions derived in section 3 to understand the best kind of resources (vehicles) for an m-vehicle spatial queuing system at different load levels ρ. We assume the management of a spatial queuing system might be interested in choosing the number of vehicles m, the speed of the the vehcies v, or the service rate of the vehicles μ. Figure 3 shows that E or V can change rapidly and significantly at load levels beyond 0.4. Both could be reduced by increasing m, v or μ.

We try to understand which of these parameters work better in different circumstances. The results discussed here set $\mu = 1, v = 1, m = 1$ as a base case and then double μ, v, or m and quantify the reduction in the expectation and variance of response time T. This analysis is all in figure 4. The vertical axis shows the ratio of E or V to the base case. For example, when the service rate μ is doubled and the load is 0.5, $E[T]$ is about 0.47 times the $E[T]$ in the base case. We read this off the solid blue curve. In other words, doubling the service rate produces a 53% reduction in $E[T]$.

Each plot corresponds to a set of values for μ, v, and m as indicated by the legend. While spatial queuing systems may use many values of μ, v, and m not covered in the figure, we restrict discussion to figure 4 because it properly captures and illustrates the trends we have discovered. The operating area of the spatial queuing system is 1 without loss of generality. The task sizes B are $Unif\,[0, 1]$. Figure 4 analyzes only the TSP. Trends for the two other location based sequencing policies, i.e., NN and DA, are similar. We have four finds itemized as P1, P2, P3, and P4.

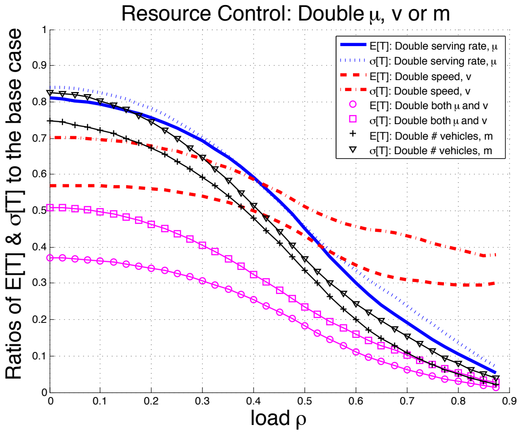

Figure 4: Resource Control under Gated-TSP.

P1: Doubling μ, v or m does more to reduce $E[T]$ than $V[T]$. This is because in each case, the $E[T]$ line is below its companion $V[T]$ line in figure 4. This finding is weak for the first (blue) pair which corresponds to the doubling of the service rate μ.

P2: The lines all slope down. This means the doubling of μ, v or m has more impact at higher loads ρ. Superior resources produce larger decreases in E and V at higher loads. Moreover, we observe an Economy of Scale (EoS) affect when ρ is greater than a certain value in the sense that doubling μ, v or m decreases $E[T]$ or $V[T]$ by more than 50%. When $\rho > 0.5$ all the ratio values are below 0.5. Below a ρ of 0.5, doubling service rate, speed, or the number of vehicles, may yield reductions in E or V that are less than 50% because some ratio values are greater than 0.5. This is because when load ρ is low, the idle time of the vehicle is long, and increasing a resource (μ,v or m) mainly increases the idle time, which does not properly utilize the increased resource. When ρ is high, the idle time of the vehicle is short, and the increasing resource decreases $E[T]$ and $Var[T]$ more strongly.

P3: We compare the value of doubling the service rate μ with doubling the speed v. Doubling v more effectively decreases both $E[T]$ and $V[T]$ than doubling μ when the load ρ is low, but is less effective when ρ is high. This is because of the EoS property of Gated-TSP policy. Task response time is a function of vehicle travel time and task service time. At higher loads service time contributes more to response time than travel time with the converse being true at lower loads. Increasing μ reduces service time but does not impact travel time. Increasing v reduces travel time but has no impact on service time. Hence the greater effectiveness of increasing v at high loads and increasing μ at low loads.

P4: We compare the value of having one “large” vehicle with double the μ and v with having two “small” vehicles. Doubling μ and v more effectively decreases $E[T]$ and $V[T]$ than doubling m at all loads ρ. We understand this as follows. Consider doubling μ while at the same time assuming $v \to \infty$ or $A \to 0$. This eliminates the value of a moving vehicle and its speed v, thereby collapsing the spatial queue under the Gated-TSP policy to a classic queue under the Gated-ROS policy. We know in classic queue that double μ is more effective in decreasing $E[T]$ and $V[T]$ than halve λ or double m [24]. Next consider doubling v while assuming that the task size B is 0. This eliminates the effect of changing μ. In general,the number of tasks inside each gate of the “large” vehicle will be greater than that of any one of the “small” vehicles. The larger vehicle will enjoy a more efficient TSP tour in the sense of traveling less distance per task than the smaller vehicles, i.e., the “large” vehicle enjoys a stronger EoS effect. Thus double the service time μ and the speed v does more to decrease $E[T]$ and $V[T]$ than doubling the number of vehicles at all load levels.

4.3 Policy Selection

The $E - V - \rho$ profile obtained from the pdf of T for different policies is given in Figure 3. Table 4 shows some values from Figure 3. Note that TSP, DA and NN have almost identical $E - V$ values at low values of ρ. However, the difference between policies is changes with ρ. For example, FCFS has higer $E[T]$ (lower reward) but lower $Var[T]$ (lower risk) than SJF, TSP, DA and NN when $\rho = 0.1$. As ρ becomes greater, the $E - V$ of FCFS and SJF increase much faster than of TSP, DA and NN. The system becomes unstable under FCFS and SJF when ρ approaches 0.5, while the $E - V$ of TSP, DA and NN are still defined. When ρ approaches 1, the system becomes unstable for all policies, which agrees with Theorem 2. TSP has lower $E - V$ than

NN and DA when ρ is high, since NN and DA produce suboptimal solutions to the TSP. DA has lower $E - V$ than NN when $\rho = 0.675$, but higher $E - V$ than NN when $\rho = 0.85$.

Table 4: $E - V - \rho$ Profile (US = Unstable)

	$\rho = 0.1$		$\rho = 0.49$		$\rho = 0.675$		$\rho = 0.85$	
Policies	E	V	E	V	E	V	E	V
FCFS	1.32	0.59	2967	2966	US	US	US	US
SJF	1.15	0.71	729	64173	US	US	US	US
G-TSP	1.19	0.87	2.95	2.14	9.5	5.4	51	24
G-DA	1.19	0.87	2.96	2.16	10.9	6.3	69	31
G-NN	1.19	0.87	2.97	2.19	12.1	6.7	64	30

The $E - V - \rho$ results in Figure 3 enable one to recommend policies adapted to the utility of an agency. The recommendation can also be adapted to the load ρ. Figure 5 shows two agencies, A and B, modeled by the Cobb-Douglas utility functions $U = \left(\frac{1}{\sigma[T]}\right)^{\epsilon} \left(\frac{1}{E[T]}\right)^{(1-\epsilon)}$ with $\epsilon^A = 0.4$ and $\epsilon^B = 0.8$. The two ϵ values make agency A more risk averse than Agency B.

At load level $\rho = 0.1$, agency A should choose FCFS while agency B should choose SJF. At $\rho = 0.45$, both agencies choose Gated-TSP. For agency A, while FCFS is preferred when $\rho = 0.1$, Gated TSP is preferred when $\rho = 0.45$. For agency B, while SJF is preferred when $\rho = 0.1$, Gated TSP is preferred when $\rho = 0.45$. In this manner a utility function can be combined with an $E - V$ computation as done in Section 3. to select the best spatial queueing policy for an agency at different load levels.

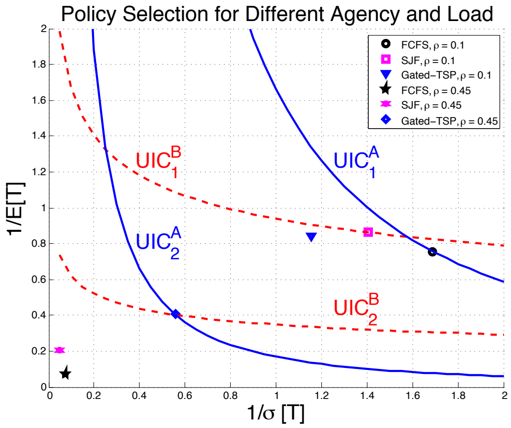

Figure 5: Policy Selection.

5. CONCLUSION

This paper has sought to provide insight useful for the management of multi-vehicle spatial queuing systems. Tasks for such a system arrive distributed in space. The vehicles must travel to the task locations and service the tasks. Tasks are uniformly distributed in size, arrive according to a Poisson process, and are uniformly distributed in space. Our management objectives include the reduction of the expected value of task response time, its variance, and the estimation of the probability a task will complete by a certain time conditioned on system state. We consider the ability to provide this probability at any time to a customer as a way of making the system transparent and predictable to the customer.

We restrict attention to gated policies with Voronoi tessellations. The operating region of the system is tessellated into Voronoi cells so that all cells are equally loaded. Each Voronoi cell is served by one vehicle or server, meaning the number of cells is equal to the number of vehicles. An arriving task is allocated to a Voronoi cell based on its location. The service policy in each cell is gated meaning, a task typically arrives when a gate is closed. Arriving tasks queue up when a gate is closed and are all admitted when the gate is opened. The gate is closed and the queue is empty. New tasks then arrive to fill the queue and wait for the next gate. The tasks in a gate are then sequenced according to policies such as FCFS, TSP, NN, and DA. We analyze only one non-gated policy, i.e., SJF.

We compute the response time distributions for each of these policies. We use numerical integration for FCFS and SJF. For TSP, NN, and DA we use Monte-Carlo simulation to justify a Gamma distribution approximation using the K-S test. Using these response time distributions we then compute $E - V - \rho$ profiles for all these policies. These profiles then become the basis for the insights into system management.

We explore three aspects of system management. The first is providing customers with the probability a task will complete by a certain time conditioned on the current system state. This is to make the system predictable and transparent to the customer. We next show how E and V change as one increases the number of vehicles, their speed, or task service rate. The changes depend on load levels. Finally, we model agencies by Cobb-Douglas utility functions and show how to use these models to select the operating policy appropriate for an agency. Once again, the best choice depends on the load level as well.

6. REFERENCES

[1] K. Baker. *Introduction to sequencing and scheduling.* Wiley, 1974.

[2] D. Bertsimas and G. Van Ryzin. A stochastic and dynamic vehicle routing problem in the euclidean plane. Working papers 3286-91., Massachusetts Institute of Technology (MIT), Sloan School of Management, 1991.

[3] B. Binger and E. Hoffman. *Microeconomics With Calculus.* The Addison-Wesley series in economics. Addison-Wesley, 1998.

[4] C. F. Daganzo. The length of tours in zones of different shapes. *Transportation Research Part B: Methodological*, 18(2):135–145, April 1984.

[5] P. De, J. B. Ghosh, and C. E. Wells. Expectation-variance analyss of job sequences under processing time uncertainty. *International Journal of Production Economics*, 28(3):289 – 297, 1992.

[6] P. H. Douglas. The cobb-douglas production function once again: Its history, its testing, and some new empirical values. *Journal of Political Economy*, 84(5):pp. 903–916, 1976.

[7] E. J. Elton and M. J. Gruber. Modern portfolio

theory, 1950 to date. *Journal of Banking & Finance*, 21:1743–1759, 1997.

[8] C. Fricker, C. Fricker, M. R. Jaibi, and M. R. Jaibi. Monotonicity and stability of periodic polling models, 1994.

[9] C. Haley and L. Schall. *The theory of financial decisions*. McGraw-Hill series in finance. McGraw-Hill, 1979.

[10] J. Huang and R. Sengupta. Risk and reward in spatial queuing theory. Working papers, UC Berkeley, Institute of Transportation Studies, 2012.

[11] J. Huang and R. Sengupta. Risk and reward in spatial queuing theory. In *Submitted to the 51st IEEE Conference on Decision and Control*, 2012.

[12] R. C. Larson and A. R. Odoni. *Urban operations research*. Prentice Hall, 1981.

[13] H. M. Markowitz and E. L. Dijk. Single-period mean-vaiance analysis in a changing world. In G. Infanger, editor, *Stochastic Programming*, volume 150 of *International Series in Operations Research & Management Science*, pages 213–237. Springer New York, 2011.

[14] J. Mossin. *Theory of financial markets*. Prentice-Hall international series in management. Prentice-Hall, Englewood Cliffs, NJ., 1973.

[15] G. Newell. *Applications of queueing theory*. Monographs on applied probability and statistics. Chapman and Hall, 1971.

[16] M. Pavone, E. Frazzoli, and F. Bullo. Adaptive and distributed algorithms for vehicle routing in a stochastic and dynamic environment. *IEEE Trans. on Automatic Control*, 2010. To appear.

[17] J. Perloff. *Microeconomics: Theory & Applications with Calculus*. The Addison-Wesley Series in Economics. Pearson Addison Wesley, 2008.

[18] S. Rathinam, Z. Kim, A. Soghikian, and R. Sengupta. Vision Based Following of Locally Linear Structures using an Unmanned Aerial Vehicle. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on*, pages 6085–6090, 2005.

[19] F. Robuste, C. F. Daganzo, and R. R. Souleyrette. Implementing vehicle routing models. *Transportation Research Part B: Methodological*, 24(4):263–286, August 1990.

[20] S. Sarin, B. Nagarajan, S. Jain, and L. Liao. Analytic evaluation of the expectation and variance of different performance measures of a schedule on a single machine under processing time variability. *Journal of Combinatorial Optimization*, 17:400–416, 2009. 10.1007/s10878-007-9122-0.

[21] S. Sarin, B. Nagarajan, and L. Liao. *Stochastic Scheduling*. Stochastic Scheduling. Cambridge University Press, 2010.

[22] B. Srirangacharyulu and G. Srinivasan. Completion time variance minimization in single machine and multi-machine systems. *Computers &amp; Operations Research*, 37(1):62 – 71, 2010.

[23] H. Takagi. *Analysis of polling systems*. MIT Press series in computer systems. MIT Press, 1986.

[24] H. Takagi. *Queueing Analysis: Discrete-time systems*. Queueing Analysis: A Foundation of Performance Evaluation. North-Holland, 1993.

[25] A. Wierman, J. Lafferty, A. Scheller-wolf, and W. Whitt. Scheduling for today's computer systems: Bridging theory and practice. Technical report, 2007.