A Collaborative Multiagent Taxi-Dispatch System

Kiam Tian Seow, Nam Hai Dang and Der-Horng Lee

Abstract—This paper presents a novel multiagent approach to automating taxi dispatch that services current bookings in a distributed fashion. The existing system in use by a taxi operator in Singapore and elsewhere attempts to increase customer satisfaction locally, by sequentially dispatching nearby taxis to service customers. The proposed dispatch system attempts to increase customer satisfaction more globally, by concurrently dispatching multiple taxis to the same number of customers in the same geographical region, and vis-à-vis human driver satisfaction. To realize the system, a multiagent architecture is proposed, populated with software collaborative agents that can actively negotiate on behalf of taxi drivers in groups of size \( N \) for available customer bookings. Theoretically, an analysis of the boundary and optimal multiagent taxi-dispatch situations is presented along with a discussion of their implications. Experimentally, the operational efficiency of the existing and proposed dispatch systems was evaluated through computer simulations. The empirical results, obtained for a 1000-strong taxi fleet over a discrete range of \( N \), show that the proposed system can dispatch taxis with reduction in customer waiting and empty taxi cruising times of up to 33.1% and 26.3%, respectively; and up to 41.8% and 41.2% reduction when a simple negotiation speedup heuristic was applied.

Note to Practitioners—With the liberalization of the taxi industry in Singapore and elsewhere, keener competition among taxi operators has emerged. We believe the taxi operator that leads the competition will be the one with the best automated taxi dispatch system, offering the highest cost productivity and customer satisfaction. Our research is motivated by the increasing need for better automated approaches to match customer service requests and taxis, whose arrival and availability, respectively, might be sporadic or not known a priori. In this paper, we propose a novel multiagent system, called \( NTuCab \) dispatch, to automate taxi dispatch in a distributed fashion. Our experiments for a 1000-strong taxi fleet show that \( NTuCab \) dispatch can outperform existing centralized dispatch in terms of reduction in customer waiting and empty taxi cruising times, when both leverage on real-time traffic information for shortest-time path computation over a road network as proposed in [2]. Additionally, \( NTuCab \) dispatch can be implemented on an existing technological infrastructure, providing the opportunities to harness the existing power of multiple intelligent transportation systems technologies. A more efficient dispatch system can help maintain a higher standard of customer service vis-à-vis human driver satisfaction, especially when the demand for taxi service is manageable for the fleet size. In future research, we will need to add more features and investigate their effectiveness towards achieving overall efficiency, including techniques to influence and better match the physical distributivity between service demand and available taxi supply in real-time.

I. INTRODUCTION

Taxis are a convenient means of paratransit in many countries, including Singapore. In providing quality customer service, fast and efficient fleet dispatching is essential. In Singapore, online dispatch of available taxis to current customer bookings is done with the aid of a satellite-based taxi dispatch system; the system utilizes a Global Positioning System (GPS) to automatically locate taxis in real-time [3], [4].

In handling current taxi bookings, the major focus of taxi dispatch systems has been primarily on reaching individual customers in the shortest time possible to enhance customer satisfaction [2]. This means reaching the customers via the shortest real-time paths possible. However, merely increasing individual customer satisfaction, as is the current practice, is a local endeavor, in that it entails assigning the nearest taxi to a customer prioritized in a first-come, first-served queue, without considering the effects of the assignment on other awaiting customers in the request queue.

In a real world scenario, there are often multiple taxi service requests (current demand) and multiple taxis available (current supply) in a given time window. To improve taxi fleet service performance, ideally, we should simultaneously and optimally assign taxis to service all customer bookings that are made within the time window. This is a challenging problem confronting current taxi dispatch systems. Practically, one feasible approach is to effectively group these customer bookings and then efficiently assign each group to the same number of available taxis. One method that we propose along this vein will be described later. The key purpose is to focus on group average customer satisfaction instead of a prioritized individual’s. This is a more global endeavor that will need to consider the mutual assignment exchange effects among the taxis for the concurrently awaiting customers, namely, ‘Would (group) total customer waiting time shorten if two taxis are allowed to exchange their currently assigned bookings?’

The motivation is that, by increasing group average customer satisfaction, overall, more customers can be satisfied.

To elaborate, consider a scenario of two available taxis in the vicinity of two taxi (service) requests, as depicted in Fig. 1. The shortest-time path to reach a request location can be computed using real-time traffic information [2], but for the convenience of illustration, we shall assume that a shorter distance path is also a shorter real-time path. Say, request 1 is initiated before request 2 within a small time window. Under the current practice, requests are allocated different taxis, one request at a time on a customer first-come, first-served basis. So the dispatcher would have to attend to request 1 first,
and assign taxi 1 to service it since taxi 1 is nearer to the request than taxi 2; this leaves taxi 2 to be assigned to the remaining request 2. However, one can see that if we could allow the taxis to exchange their assigned requests, the group average distance (or time) taken by a taxi, and hence the (group average) customer waiting time, would be shorter. This more balanced allocation is often possible if the two requests could be considered concurrently for taxi assignment, exploiting the mutual assignment effects in attempting to minimize the total real-time travel period of the taxis in picking up the customers. 

And the allocation is practically efficient only if it can be done without excessive communication between the taxi dispatch center and the taxis. To the best of our knowledge, current taxi dispatch systems do not support the kind of concurrent taxi assignment envisaged, and cannot therefore exploit this real world scenario to full advantage as needed for improving fleet performance.

In our opinion, to support such concurrent taxi assignment for current taxi bookings necessitates a novel approach to fleet dispatch operation. The new approach should not only aim at increasing average customer satisfaction, but also vis-à-vis average driver satisfaction. And it must be implementable on an existing technological infrastructure. With this philosophy in mind, we propose a multiagent approach deploying collaborative taxi agents that can, on behalf of taxi drivers, cooperatively negotiate to decide among themselves their different assignments, from among the multiple taxi requests initiated within a time window. A taxi agent is an active software entity residing in an in-vehicle computing unit of a taxi. By cooperative negotiation [5], several taxi agents can collaboratively search for an assignment solution that they jointly agree.\footnote{At this juncture, a general notion of negotiation suffices; it will become clearer in Section II-B, where a specific automated negotiation mechanism for taxi agents is introduced and elaborated.}

Intelligent agents and multiagent approaches are gaining increasing prominence in automation research, and have been successfully formulated or applied in the domains of manufacturing (e.g., [6]), warehousing (e.g., [7]), product family design (e.g., [8]) and hierarchical decision-making protocols (e.g., [9]), to name a few. Although using multiple automated agents is also not new in intelligent service transportation, as seen in transport logistics [10] and route guidance [11], we realize that it might be a radically new ideology to deploy them for the specific purpose of taxi dispatch. For, the new idea entails the software (‘agents’) localized in the in-vehicle computing units to collaborate and play a more active role in consensus decision-making, rather than just passively presenting a new request from, and relaying the human driver’s request acceptance or refusal decision to the taxi dispatch center. However, in our opinion, investigating this idea is timely, since a multiagent approach will invariably provide a set-up to harness the existing power of multiple intelligent transportation systems (ITS) technologies in, for example, vehicle routing [12], automatic vehicle location [13], mobile phone location determination [14] and palmtop-based navigation [15], exploiting the huge investments already made in the internet, wireless communication and mobile devices, as well as GPS-based location, geographic and traffic information systems.

The rest of this paper is organized as follows. Section II describes the basic architectures of both the existing and proposed taxi dispatch systems, with emphasis on the core problem and solution of the latter. Section III presents and discusses a microscopic simulation study comparing the proposed and current dispatch approaches. Section IV concludes the paper and points to some future work.

II. TAXI DISPATCH SYSTEMS

A. Current Taxi Dispatch System

Fig. 2 depicts the basic architecture of a current taxi dispatch system in use by a taxi operator in Singapore and elsewhere. Incoming taxi service requests are queued on a first-come, first-served basis at the dispatch center. For each customer
Proposed NTuCab dispatch system: Non-centralized

In general, it is not feasible to compute the (locally) shortest-time path for each of a possibly large number of available taxis nearby a customer location, since determining such a path requires a considerably significant amount of computation time. Thus, it has been assumed [2] that a taxi with the shortest-time path is found among a limited number (say, 20) in the vicinity that are the nearest in terms of their shortest straight-line distances to the customer location.

As a remark, besides [2] and a related but different effort that empirically examines the performance improvement of using GPS in taxi dispatch operations [3], there is apparently little work in the published literature on automated taxi dispatch. This is despite its importance and commercial interest. Henceforth, for comparison purposes, we shall assume that the existing (state-of-the-art) taxi dispatch system available employs the centralized architecture of Fig. 2 and applies the dispatch method proposed in [2]. In our opinion, this system should compare favorably with the centralized dispatch system provided by a market leader. According to the website (www.cordic.com), the commercial system associates every taxi service request with a taxi queue according to its customer pick-up location, and then offers it to the first taxi in that queue.

B. Proposed Taxi Dispatch System

As suggested in the introduction, multiagent collaboration might leverage on the improved dispatch method [2] by exploiting the mutual assignment exchange effects among multiple taxis awaiting multiple requests - a scenario that is not uncommon. Towards this end, an infrastructure to support a taxi agent capable of collaboration is proposed, as shown in Fig. 3(a). And, to effectively utilize such software agents for taxi dispatch, we deploy them in a multiagent architecture, proposed as depicted in Fig. 3(b).

1) Multiagent Taxi Dispatch Architecture: The architecture assumes that a geographical road network is partitioned into \( m \) logical areas of taxi operation, \( m \geq 1 \). This partition is made known to all taxi agents. At the dispatch center, available taxis and booking requests in a logical area are recorded in a taxi queue and a request queue, respectively. The area and corresponding queues are identified by the same index \( i \), \( 1 \leq i \leq m \).

Initially, all participating taxis are in one of the designated areas of operation, and the taxi queues at the dispatch center are updated accordingly. The ‘housekeeping’ communication protocol supporting the dispatch operations can then be prescribed as follows:

1) A taxi agent [see Fig. 3(a)] performs the following mandatory tasks:
   a) Announce the availability of its taxi in a new area of operation to the dispatch center when (i) it is not in negotiation, (ii) the taxi has entered (and is currently in) the new area, and (iii) the passenger

2) Examples of commercial companies delivering automated taxi dispatch solutions include Cordic, DDS Digital Dispatch and Mobisoft.

3) We recommend JADE [16] as the agent technology development package for MA3-LM agent software implementation. In principle, however, any other agent development package that supports communication may be used.

4) That an area \( i \) is new means that, operationally, the taxi agent has not notified the dispatch center to insert its taxi record into the taxi queue \( i \) (that otherwise does not contain the record).
Under the proposed architecture, the dispatch center carries out the following operations per cycle:

1) Do housekeeping and await taxi assignment decisions.

2) The dispatch center performs the following mandatory housekeeping task:

a) DMT-A: Update the availability of taxis in the respective taxi queues, taking the information from a common taxi queue that is continually updated in response to notification by the taxi agents. When a taxi has left an area $x$ and entered a new area $y$, update by deleting the taxi record from the taxi queue $x$, and inserting the record, in a first in first out fashion, into the taxi queue $y$.

b) DMT-B: Insert the records of taxi bookings into the respective request queues, taking the information from a common request queue that is continually accepting customer taxi bookings. If a taxi request is from an area $i$, insert the record, on a first-come, first-served basis, into the request queue $i$.

c) DMT-C: Delete the records of a taxi and the negotiated assignment that the taxi has accepted; and communicate the service information of the taxi to the assigned customer. Penalize by placing the record of a taxi that has refused to accept its negotiated assignment at the tail of the same taxi queue at the time of update; and compensate by placing that of the refused assignment at the head of the respective request queue.

Under the proposed architecture, the dispatch center carries out the following operations per cycle:

1) Distribute pending requests to available taxis for each logical area $i$.

Take a group of $N$ requests from the head of the request queue $i$ and send it to an ad hoc group of taxis assembled from the head of the taxi queue $i$. The size of $N$ of the last group assembled may vary depending on the number of pending requests and available taxis. Do so until the taxi or request queue is empty. (As soon as a group of bookings is received, the taxi agents concerned will asynchronously carry out intra-group negotiation over the $N$ requests.)

2) Do housekeeping and await taxi assignment decisions.

a) Do housekeeping tasks DMT-A and DMT-B as needed till notification is received about completion of a group collaboration; following which there is a pre-specified period when every agent in the group submits the human driver’s decision of either accepting or refusing its negotiated booking.

b) Do housekeeping task DMT-C.

3) Continue with Step (2) if not all the ad hoc taxi groups across all the logical areas have completed negotiation, else do tasks DMT-A and DMT-B as needed and proceed to next cycle of dispatch operations.

We call the resulting system the N-Taxi GroUp CollABorative (NTuCab) dispatch system.

2) Core Problem & Solution: Within this architecture, the core issue in multiagent taxi dispatch is a linear assignment problem (LAP) [17]. The problem is concerned with efficiently assigning every taxi agent with a different taxi request. The efficiency (or optimality) of the concurrent allocation is measured either in terms of minimizing total cost or maximizing total service quality.

To elaborate formally, consider a group of taxi agents $A = \{a_0, a_1, \cdots, a_{N-1}\}$ of size $N \geq 2$, and a group of different taxi-service requests $O = \{r_0, r_1, \cdots, r_{N-1}\}$ of size $N$. The A-QoS (application quality-of-service) that an agent $a \in A$ can offer for each request is defined by $d[a, r]$ for all $r \in O$. Then our core objective of taxi dispatch is to find, for an $N \times N$ LAP, the particular (total) assignment solution

$$\Pi : A \rightarrow O$$

such that for $a_i, a_j \in A$, $i \neq j$ implies $\Pi(a_i) \neq \Pi(a_j)$

a one-to-one matching of agents to requests that (ideally) maximizes the total A-QoS $S_{tot}$.

$$S_{tot} = \sum_{i=0}^{N-1} d[a_i, \Pi(a_i)]$$

$\Pi(a) \in O$ refers to a request selection by agent $a \in A$ (under an arbitrary permutation of $\Pi$); and $\max\{\forall a \in A, \Pi(a)\}$ defines the (ideal) optimal social goal of the agents. An assignment or allocation set corresponds to one permutation of $\Pi (1)$, and can also be equivalently represented as containing elements of the form $(a, \Pi(a)) \in A \times O$.

The generic LAP has been extensively researched in the Operations Research literature [18]. Using agent modeling ideas, recent research has developed agent mechanisms [19], [20] for a multiagent version, termed collaborative LAP (CLAP), where knowledge about the LAP is distributed among the agents, such that every agent $a \in A$ initially only has its own local information, $d[a, r]$ for all $r \in O$. The automated mechanisms developed enable collaborative taxi agents to cooperatively negotiate for different requests by themselves, in contrast to a centralized algorithm deciding for them as in [18]. In essence, these taxi agents can compute and leverage on the possible overall A-QoS increments, achievable through properly reassigning requests among themselves.

One important development for CLAP is a decentralized agent algorithm called MA$^3$-LM [20], which is well-suited and adapted for use with the proposed taxi dispatch architecture. The reasons for considering MA$^3$-LM are: (i) the core problem in collaborative taxi dispatch - taxi online assignment or allocation - can clearly be treated as a CLAP; (ii) MA$^3$-LM is decentralized and so maps directly onto our proposed multiagent architecture, and (iii) being computationally simple and easy to understand, one of our research contributions is to propose the first potential real-world service-automation
application of MA$^3$-LM, along with an investigation of its applicability and performance.

The MA$^3$-LM agents divide their collaborative reasoning process into negotiation rounds. All the agents begin negotiation with an initial selection made under an arbitrary permutation of $\Pi$ (1) (the one-to-one mapping). For taxi dispatch, every taxi agent in group $A$ negotiates with the other agents in the same group over the requests in $O$. The purpose is to decide which two agents are to exchange their current request selections in every consecutive round. This is done by collaborative reasoning per round, during which every agent determines its individual request exchange desires and resulting exchange intention (its best desire), based on its computed beliefs of exchange alternatives that may increase the group’s total A-QoS, $S_{\text{tot}}(2)$.

**Definition 1 (Belief Set $B_i$):** Given that an agent $a_i \in A$’s current request selection is $r^i \in O$. Then its (current) belief set $B_i$ is given by

$$B_i = \{ r \in O \mid d[a_i, r] > d[a_i, r^i]\}$$

(3) If $B_i \neq \emptyset$, this means that agent $a_i \in A$ has at least one alternative request selection $r \in B_i$ that may lead to an increase in total A-QoS $S_{\text{tot}}(2)$ when made in exchange with an agent whose current selection is $r \in O$.

**Definition 2 (Desire Set $D_i$):** Given that an agent $a_i \in A$’s current request selection is $r^i \in O$ and its belief set is $B_i$, $B_i \neq \emptyset$. An arbitrary agent $a_j \in A$ whose current request selection is $r^j \in O$ is said to accept agent $a_i \in A$’s beliefs $B_i$ if $r^j \in B_i$. To generate the desired exchange options or desires $D_i$, agent $a_i \in A$ broadcasts its beliefs $B_i$ and current selection $r^i \in O$, and an arbitrary agent $a_j \in A$ who accepts the beliefs would respond with a pair of A-QoS values $d[a_j, r^i]$ and $d[a_j, r^j]$, so that for each of the $|B_i|$ responses received, the corresponding request exchange option $[(a_i, r^j), (a_j, r^j), \rho] \in D_i$ (i.e., is agent $a_i \in A$’s desire) if $\rho > 0$, whereas $\rho$ is defined by

$$\rho = -d[a_i, r^j] + d[a_i, r^j] - d[a_j, r^i] + d[a_j, r^i]$$

(4) $\rho > 0$, gives up its current selection $r^i \in O$ and selects $r^j \in O$, and in exchange, agent $a_j \in A$ gives up its current selection $r^j \in O$ and selects $r^i \in O$. Thus, any desire $\rho \neq D_i$, when carried out, will definitely lead to an increase in total A-QoS without violating $\Pi$. Quite naturally, it provides the motivation for agent $a_j \in A$ to want to exchange its current request selection.

**Definition 3 (Intention $I_j$):** Given that an agent $a_i \in A$’s desire set is $D_i$, $D_i \neq \emptyset$. Then, agent $a_i \in A$’s intention $I_j$ is given by

$$I_j = [(a_i, r^j), (a_j, r^j), \rho] \in D_i$$

(5) for which

$$\rho = \max \{ \rho' \mid [-, -, \rho'] \in D_i \}$$

Agent $a_i \in A$’s decisive stance or intention has to be $I_j$ since it is viewed as the best exchange option (in terms of net exchange gain from the agent’s perspective) that the agent can propose. It is said to have no intention if either $B_i = \emptyset$ or $D_i = \emptyset$, in which case $I_j = \text{nil}$, where $\text{nil} = [-, -, 0]$.

All the agents’ exchange intentions (or the lack thereof communicated as a nil intention) $I_j \in I$ would need to undergo arbitration to decide which two agents to proceed with the request exchange, before a negotiation round is concluded, and the next round begins. Essentially, an intention $I = [-, -, \rho]$ with the highest exchange gain $\rho > 0$, i.e., one that contributes to the highest increase (in total A-QoS) if carried out, would need to be selected. In so doing, all the agents perform local mediation to arbitrate their intentions. This involves incremental arbitration of intention as the better intention (i.e., the one with higher net exchange gain) - between a taxi agent’s own and the intention it received - is passed by the agent to another in a circular fashion, before finally reaching a taxi agent that is concurrently and dynamically entrusted the role of a request selection-exchange initiator for a particular negotiation round [20]. This process differs from that in the original mechanism called MA$^3$ [19], which uses the same BDI reasoning model but has an extra agent dedicated to receiving and arbitrating all agent intentions. The negotiation process will terminate following a negotiation round when all agents have no (more) intention to exchange request selections and so submit nil intentions, discovered through arbitration by local mediation.

The relevant algorithmic details for MA$^3$-LM agents [20] based on the formal description above is generically summarized in [1, p. 1050]. MA$^3$-LM agents always terminate with an often highly efficient but not necessarily optimal solution after a finite number of negotiation rounds [19, 20].

To reduce the number of negotiation rounds in the negotiation process, an assignment initialization heuristic (labelled H-Max) can be applied prior to negotiation, the details of which are documented in the appendix. And to significantly reduce the average communication time per negotiation round in practice, a faster version called non-redundant BDI reasoning discussed in [21, p. 685] for MA$^3$ can be easily adapted to MA$^3$-LM agents.

In our current work, the A-QoS data entry $d[a_i, r_j] < 0$ denotes the negation of the shortest (planned) real-time for a taxi (that the MA$^3$-LM agent $a_i \in A$ represents) to travel from its designated or current location to the pick-up location of a customer (who initiated the request $r_j \in O$); the negation is so that the MA$^3$-LM agents, in maximizing the total A-QoS $S_{\text{tot}}(2)$, are minimizing the total travel time of their taxis. This real-time travel period is that computed over a directed road network with real-time traffic information [2]. The smaller this time value is, the nearer to the customer the taxi is said to be.

3) Decentralized versus Centralized - A Discussion: The NTuCab dispatch system architecture deliberately avoids using a centralized (single-agent) algorithm [18] for assigning requests to taxi agents, thus removing a processing bottleneck and a crucial ‘single point of failure’. The A-QoS information is already distributed among the taxi agents, so it is wasteful to have to replicate by centralized information gathering or computation, required if single-agent assignment processing is used instead. In decentralized MA$^3$-LM processing, multiple taxi agents share their local A-QoS information only as needed when they negotiate for a request. In practice, the proposed architecture can even allow these MA$^3$-LM taxi agents to flexibly respond on-the-fly (i.e., during negotiation) to A-QoS data invalidation, by their independently correcting their local
A-QoS values as necessary before the start of each negotiation round. Such data invalidation can occur due to, say, sudden changes detected in road traffic conditions. In this practical situation, the centralized algorithm, on the contrary, would have to recompute from scratch the assignment solution, after data validation through recomputing or re-gathering the A-QoS values from the agents involved; and this could possibly incur even higher processing and communication overheads. Finally, the proposed system can also be readily extended to degrade more gracefully when some of the taxis or taxi agents are out of service temporarily.

4) Theoretical Boundary & Optimal Situations: The physical locations of taxis and customers are significant information for taxi dispatch. They are in general not known a priori. However, two theoretical boundary situations about their relative locations can be identified:

1) DT-CC: Geographically distributed taxis for concentrated customers, as when dispatching taxis to a taxi stand.

2) CT-DC: Geographically concentrated taxis for distributed customers, as when dispatching taxis from a depot.

<table>
<thead>
<tr>
<th></th>
<th>r₀</th>
<th>r₁</th>
<th>r₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₀</td>
<td>x₀</td>
<td>x₁</td>
<td>x₂</td>
</tr>
<tr>
<td>a₁</td>
<td>y₀</td>
<td>y₁</td>
<td>y₂</td>
</tr>
<tr>
<td>a₂</td>
<td>z₀</td>
<td>z₁</td>
<td>z₂</td>
</tr>
</tbody>
</table>

(a) For DT-CC | | | |

<table>
<thead>
<tr>
<th></th>
<th>r₀</th>
<th>r₁</th>
<th>r₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₀</td>
<td>x₀</td>
<td>y₀</td>
<td>z₀</td>
</tr>
<tr>
<td>a₁</td>
<td>x₁</td>
<td>y₁</td>
<td>z₁</td>
</tr>
<tr>
<td>a₂</td>
<td>x₂</td>
<td>y₂</td>
<td>z₂</td>
</tr>
</tbody>
</table>

(b) For CT-DC

Fig. 4. Types of CLAP for the boundary situations, shown for $N = 3$

Being ‘geographically concentrated’ means that the taxis or customers are in close physical proximity of one another. Theoretically, we shall assume that they are at the same location point. Following, the DT-CC and CT-DC situations result in taxi agents facing the types of CLAP as depicted in Fig. 4, where each row of matrix entries are the respective A-QoS values $d(a_i, r_j)$ as computed by an agent $a_i \in A$ for every request $r_j \in O$.

In between the two boundary situations, we also identify a situation, labelled OM-TC, where the geographical distributions between the taxis and the service requests are optimally matched, such that every taxi agent $a \in A$ has a different request $r \in O$ that is the nearest to its taxi location; all its other requests are not as near. An example of situation OM-TC is given in Fig. 5, in which the values shown each indicates the largest A-QoS (or the least negative travel time) of an agent $a_i \in A$ for a different request.

<table>
<thead>
<tr>
<th></th>
<th>r₀</th>
<th>r₁</th>
<th>r₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₀</td>
<td>-</td>
<td>-</td>
<td>x₀</td>
</tr>
<tr>
<td>a₁</td>
<td>-</td>
<td>x₁</td>
<td>-</td>
</tr>
<tr>
<td>a₂</td>
<td>x₀</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 5. A CLAP instance of OM-TC, with $N = 3$

Following, we present two propositions on MA₃-LM in the situations identified. The proofs entail an understanding of the formal definitions of Belief (B), Desire (D) and Intention (I), and the termination condition of MA₃-LM. For the proof of Proposition 2, the reader is also referred to the appendix for details of the speedup heuristic, H-Max.

Proposition 1: The MA₃-LM taxi-agents in situations DT-CC and CT-DC will incur one negotiation round to reach an optimal (or a maximal) total A-QoS assignment which is arbitrary.

Proof: We sketch the proof as follows: In either situation, in the first round of negotiation, every MA₃-LM taxi-agent will yield a nil intention (Definition 3) - the termination condition for MA₃-LM. In the former situation, it is due always to an empty belief set (Definition 1) computed by every agent. In the latter, it is due always to an empty desire set (Definition 2) computed by every agent except one, which yields a nil intention due always to an empty belief set. The first round is thus the last negotiation round. Finally, the total A-QoS assignment solution reached must be optimal when negotiation terminates, since, in either situation, any collaborative assignment reached is associated with the same set of A-QoS values, $\{d(a, II(a)) | a \in A\}$, and thus it has the same total value regardless of the request selections [under II (1)]. Hence the proposition.

In other words, in situations DT-CC and CT-DC, MA₃-LM taxi-agents will incur the minimum number of negotiation rounds [19], [20]. Despite the quick decision, negotiation is not needed in such situations since every assignment is optimal.

Proposition 2: The MA₃-LM taxi-agents using H-Max in situation OM-TC will incur a total of two negotiation rounds to reach an optimal total A-QoS assignment.

Proof: We sketch the proof as follows: In situation OM-TC, immediately after assignment initialization through H-Max, every MA₃-LM taxi-agent [under II (1)] will have selected or reselected a different request that is the nearest to its taxi location. This initialization process is considered to incur the first negotiation round. In the second round of negotiation that follows, every MA₃-LM taxi-agent will yield a nil intention - the termination condition for MA₃-LM - due always to an empty belief set computed by the agent. The second round is thus the last negotiation round. Finally, since every agent will have selected its nearest request [under II (1)] when negotiation terminates, it follows that the total A-QoS assignment solution reached is optimal (or maximal). Hence the proposition.

Proposition 2 suggests that situation OM-TC characterizes the ideal condition to be in, where a relatively quick negotiation will result in an assignment that is both locally and globally optimal. By local optimality, every taxi agent $a \in A$ selects a request $r \in O$ for which the A-QoS $d[a, r]$ among all requests in $O$ is the largest; and by global optimality, we mean that the total A-QoS of the assignment is maximized.

From the above analysis and discussion, we infer that if the working scenarios are exclusively DT-CC or CT-DC, as could be reasonably assumed in the past, the proposed taxi dispatch approach (see Fig. 3) might only perform as well as the current approach (see Fig. 2). However, in a situation which is in between DT-CC or CT-DC and OM-TC, we envisage that the proposed distributed approach can often significantly
The use of mobile phones and the associated technologies in cities outperform the existing centralized one. With the pervasive use of mobile phones and the associated technologies in cities like Singapore, where one can conveniently book a taxi from almost anywhere, the demand by individual customers for taxi service has set to become more geographically distributed. This emerging real-world scenario, along with the induced necessity for empty taxis to roam about more frequently, means that such in-between situations could become a common occurrence, and it is no longer acceptable to simply assume situation DT-CC or CT-DC. All these are suggestive of a better performance using our proposed approach, and this is confirmed by simulations.

III. SIMULATIONS & PERFORMANCE EVALUATION

To study the comparative performance of the proposed NTuCab dispatch system and an existing centralized system (see Section II-A), we conducted microscopic computer simulations, simulating taxi operations in a selected ITS-managed urban road network of reasonable complexity. We focused on operational efficiency. For both the systems, this was investigated in terms of customer waiting time versus empty cruising time. In our simulations, customer waiting time is measured from the moment a customer raises a request to the moment an assigned taxi arrives to pick up the customer; empty taxi cruising time is measured from the moment it is available to the moment it accepts (or commits to service) a negotiated assignment.

A. Experimental Scope & Investigation

The experiments were performed on MITSIMLab [22], [23], [24], through a Taxi Management Microscopic Simulator (TM2S) developed for this research study. The overall software architecture of the simulator is shown in Fig. 6.

The MITSIMLab simulation software (available from web.mit.edu/its/mitsimlab.html) is an existing simulation-based laboratory originally developed for evaluating traffic management system designs at the operational level. It was adapted as the environment for all the experiments, along with a urban road network model built by MITSIMLab developers using a road network editor [23]. The urban road network model used is shown in Fig. 7; it covers a physical area of about 15km × 10km. In providing the required real-time traffic information for dispatch operation, an abstracted road network was also built over the MITSIMLab-based urban road network model.

The TM2S module wraps around the MITSIMLab to simulate the real time activities of a dispatch operator and the associated MA₃-LM negotiation processes of a network of collaborative taxi agents, in accordance to either the centralized or the proposed NTuCab dispatch system (see Section II), but confined to one logical area of operation. The TM2S module assumes that the taxi agents negotiate over a high-speed wireless communication network. In calculating the multiagent negotiation time, the module abstracts away the underlying detailed communication costs and instead estimates the total negotiation time based on the number of negotiation rounds taken and a conservative estimate of 0.2N seconds per round for an N-group of taxi agents. The estimate is the longest average time per negotiation round that we found at N = 20, when we experimented separately for groups of N ∈ {5, 10, 15, 20}, by running a Java program of MA₃-LM implemented on a multiagent platform called JADE [16] in a local area network (emulating the high-speed communication infrastructure), and without applying the speedup heuristic, H-Max.

In the MITSIMLab environment, the taxis move randomly in the road network. The human drivers of available taxis are assumed to always accept and service any taxi request negotiated by their agent. Upon the taxi agents in TM2S accepting their negotiated assignment, each corresponding taxi moves from its current location to the assigned customer’s pick-up location, and then to the customer’s destination.

For both the centralized and proposed systems, we computed the travel times using real-time traffic information as proposed in [2]. As defined, the A-QoS d[a,r] < 0 denotes

![Fig. 6. The TM2S-MITSIMLab simulation model](image-url)

![Fig. 7. The urban road network model used](image-url)
the negation of the expected shortest travel time for a taxi (represented by an agent $a \in A$) to move from its current location to the pick-up location of a customer (who initiated the pending request $r \in O$); since taking the shortest-time path is a time-efficient way to service taxi requests [2]. The A-QoS formula used is

$$d[a, r] = -\min_{i \in O} \{T_i\}$$  \hspace{1cm} (6)

$T_i$ denotes the expected real-time for the taxi to travel on an arbitrary path $i$ through the road network connecting the two locations. The expected shortest travel time $\min_{i \in O} \{T_i\}$ was calculated using the route choice model provided by the traffic flow simulator module of MITSIMLab. The time computation considered (i) current traffic conditions, (ii) delay and regulation of turning movements at road intersections and (iii) possible penalties for certain designated links (e.g., freeway bias). Along with the determination of the associated shortest-time path for a taxi, it is essentially a vehicle route planning problem, solved using the techniques developed for MITSIMLab [24, Appendix B].

In principle, both the existing centralized and NTuCab dispatch systems should operate for any taxi fleet size. For our simulations over the TM2S-MITSIMLab model, a taxi fleet size of 1000 was simulated for a one-hour duration. The taxi fleet is about 30% of the traffic volume that can be simulated on MITSIMLab, and constitutes a reasonable traffic composition in the urban setting considered.

We carried out experiments for a range of hourly demand rates (defined by taxi bookings per taxi per hour). For each demand rate, simulated with incoming requests generated by the request manager, the customer waiting and empty taxi cruising times were recorded for centralized dispatch, and collaborative agent dispatch for several group sizes $N \in \{5, 10, 15, 20\}$. The experiments were repeated for collaborative agent dispatch with speed-up initialization heuristic incorporated (see Appendix).

The simulation data (raw customer waiting and empty cruising times) gathered from the dispatch operator and taxi agents were saved into a text file for offline performance analysis.

### B. Analysis of Numerical Results

Table I shows that under NTuCab dispatch, at each demand rate and for a small $N = 5$, the customer waiting time is longer. This is due to less efficient assignments as some better positioned taxis for the requests might not be in the taxi groups that negotiated for them. As $N$ increases, initially, the customer waiting time becomes shorter due to increase in grouping efficiency, but for a bigger $N = 20$, especially at higher demand rates, it starts increasing due to offset to the gains in grouping efficiency, namely, longer negotiation time for a bigger $N$, and service demand outstripping taxi supply at higher demand rates.

### Table I

**Operational efficiency of NTuCab and centralized dispatch:**

**Customer waiting time (in s)**

<table>
<thead>
<tr>
<th>Demand rate</th>
<th>$N$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>128.1</td>
<td>127.0</td>
<td>135.9</td>
<td>165.7</td>
<td>189.0</td>
<td>281.2</td>
<td>479.9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>109.9</td>
<td>111.4</td>
<td>116.1</td>
<td>141.0</td>
<td>161.3</td>
<td>225.6</td>
<td>371.8</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>110.1</td>
<td>106.6</td>
<td>113.5</td>
<td>136.6</td>
<td>155.5</td>
<td>222.1</td>
<td>352.8</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>112.0</td>
<td>112.9</td>
<td>120.8</td>
<td>144.1</td>
<td>186.6</td>
<td>255.9</td>
<td>422.9</td>
<td></td>
</tr>
<tr>
<td>(b) With H-Max</td>
<td>$N$</td>
<td>127.0</td>
<td>126.6</td>
<td>135.0</td>
<td>166.5</td>
<td>178.3</td>
<td>205.4</td>
<td>312.4</td>
</tr>
</tbody>
</table>

Table III shows that, regardless of the group size $N$, under NTuCab dispatch, as the demand rate increases, empty cruising time shortens. This implies that taxis roam less frequently without customers onboard. The empty cruising time converges approximately to the $N$-group negotiation time, with the roaming time without request negotiation approaching zero. Comparing Tables III(a) and III(b), we observe that with the speedup heuristic, group negotiation, and hence empty cruising time, shortens for all demand rates and $N$-group sizes considered.

At each demand rate (except when it is 1 for $N = 5$, and without applying speedup heuristic), NTuCab dispatch outperforms centralized dispatch, with good reductions in customer waiting time of up to 33.1% at demand rate 4 for $N = 15$ [see Table II(a)], and up to 41.8% at demand rate 2.5 for $N = 20$, when the speedup heuristic is used [see Table II(b)]; and with good reductions in empty cruising time of up to 26.3% at demand rate 1.5 for $N = 20$ [see Table IV(a)], and up to 41.2% at demand rate 4 for $N = 20$, when the speedup heuristic is used [see Table IV(b)].

#### Table III

**Customer waiting time:**

<table>
<thead>
<tr>
<th>Demand rate</th>
<th>$N$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>128.1</td>
<td>127.0</td>
<td>135.9</td>
<td>165.7</td>
<td>189.0</td>
<td>281.2</td>
<td>479.9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>109.9</td>
<td>111.4</td>
<td>116.1</td>
<td>141.0</td>
<td>161.3</td>
<td>225.6</td>
<td>371.8</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>110.1</td>
<td>106.6</td>
<td>113.5</td>
<td>136.6</td>
<td>155.5</td>
<td>222.1</td>
<td>352.8</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>112.0</td>
<td>112.9</td>
<td>120.8</td>
<td>144.1</td>
<td>186.6</td>
<td>255.9</td>
<td>422.9</td>
<td></td>
</tr>
<tr>
<td>(b) With H-Max</td>
<td>$N$</td>
<td>127.0</td>
<td>126.6</td>
<td>135.0</td>
<td>166.5</td>
<td>178.3</td>
<td>205.4</td>
<td>312.4</td>
</tr>
</tbody>
</table>

#### Table IV

<table>
<thead>
<tr>
<th>Demand rate</th>
<th>$N$</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>128.1</td>
<td>127.0</td>
<td>135.9</td>
<td>165.7</td>
<td>189.0</td>
<td>281.2</td>
<td>479.9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>109.9</td>
<td>111.4</td>
<td>116.1</td>
<td>141.0</td>
<td>161.3</td>
<td>225.6</td>
<td>371.8</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>110.1</td>
<td>106.6</td>
<td>113.5</td>
<td>136.6</td>
<td>155.5</td>
<td>222.1</td>
<td>352.8</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>112.0</td>
<td>112.9</td>
<td>120.8</td>
<td>144.1</td>
<td>186.6</td>
<td>255.9</td>
<td>422.9</td>
<td></td>
</tr>
<tr>
<td>(b) With H-Max</td>
<td>$N$</td>
<td>127.0</td>
<td>126.6</td>
<td>135.0</td>
<td>166.5</td>
<td>178.3</td>
<td>205.4</td>
<td>312.4</td>
</tr>
</tbody>
</table>
TABLE III
OPERATIONAL EFFICIENCY OF NTuCab AND CENTRALIZED DISPATCH:
EMPTY CRUSING TIME (IN S)

(a) Without H-Max

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2780.7</td>
<td>1973.9</td>
<td>1572.4</td>
<td>855.2</td>
<td>495.1</td>
<td>408.9</td>
<td>303.9</td>
</tr>
<tr>
<td>10</td>
<td>2305.5</td>
<td>1544.0</td>
<td>1215.9</td>
<td>897.2</td>
<td>389.2</td>
<td>333.2</td>
<td>247.1</td>
</tr>
<tr>
<td>15</td>
<td>2221.3</td>
<td>1519.8</td>
<td>1184.5</td>
<td>845.4</td>
<td>389.2</td>
<td>342.8</td>
<td>239.3</td>
</tr>
<tr>
<td>20</td>
<td>2332.6</td>
<td>1487.5</td>
<td>1227.6</td>
<td>799.8</td>
<td>412.7</td>
<td>325.3</td>
<td>250.9</td>
</tr>
<tr>
<td>(c)</td>
<td>2715.5</td>
<td>2018.3</td>
<td>1596.3</td>
<td>904.2</td>
<td>511.4</td>
<td>420.2</td>
<td>315.2</td>
</tr>
</tbody>
</table>

(b) With H-Max

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2685.6</td>
<td>1947.7</td>
<td>1529.3</td>
<td>869.0</td>
<td>495.1</td>
<td>403.0</td>
<td>303.9</td>
</tr>
<tr>
<td>10</td>
<td>2142.5</td>
<td>1463.3</td>
<td>1211.6</td>
<td>622.1</td>
<td>367.7</td>
<td>279.5</td>
<td>201.1</td>
</tr>
<tr>
<td>15</td>
<td>2169.7</td>
<td>1519.8</td>
<td>1168.5</td>
<td>641.1</td>
<td>346.0</td>
<td>279.9</td>
<td>200.8</td>
</tr>
<tr>
<td>20</td>
<td>2251.1</td>
<td>1408.8</td>
<td>1096.7</td>
<td>602.2</td>
<td>340.6</td>
<td>270.2</td>
<td>185.4</td>
</tr>
</tbody>
</table>

(c): Centralized dispatch

TABLE IV
EMPTY CRUSING TIME: REDUCTIONS (IN %) OF NTuCab OVER CENTRALIZED DISPATCH

(a) Without H-Max

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-2.4</td>
<td>2.2</td>
<td>1.5</td>
<td>2.1</td>
<td>3.2</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td>10</td>
<td>15.1</td>
<td>23.5</td>
<td>22.6</td>
<td>22.9</td>
<td>23.9</td>
<td>20.7</td>
<td>21.6</td>
</tr>
<tr>
<td>15</td>
<td>18.2</td>
<td>24.7</td>
<td>25.8</td>
<td>23.9</td>
<td>23.9</td>
<td>22.7</td>
<td>24.1</td>
</tr>
<tr>
<td>20</td>
<td>22.3</td>
<td>24.7</td>
<td>23.3</td>
<td>21.5</td>
<td>19.3</td>
<td>22.6</td>
<td>20.4</td>
</tr>
</tbody>
</table>

(b) With H-Max

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.1</td>
<td>3.5</td>
<td>4.2</td>
<td>3.9</td>
<td>3.2</td>
<td>3.4</td>
<td>4.1</td>
</tr>
<tr>
<td>10</td>
<td>21.1</td>
<td>27.5</td>
<td>24.1</td>
<td>31.2</td>
<td>28.1</td>
<td>33.5</td>
<td>36.2</td>
</tr>
<tr>
<td>15</td>
<td>20.1</td>
<td>24.7</td>
<td>26.8</td>
<td>29.1</td>
<td>33.4</td>
<td>33.4</td>
<td>36.3</td>
</tr>
<tr>
<td>20</td>
<td>17.1</td>
<td>30.2</td>
<td>31.3</td>
<td>33.4</td>
<td>33.4</td>
<td>35.7</td>
<td>41.2</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

For automating distributed taxi dispatch, this paper has introduced the new idea of multiagent collaborative assignment of current bookings. The proposed NTuCab dispatch system realizes the idea using MA^3-LM through a proposed multiagent architecture (see Section II-B). Theoretical boundary (DT-CC and CT-DC) and optimal (OM-TC) situations are analyzed (see Propositions 1 and 2, respectively) and discussed. Using TM2S-MITSIMLab simulations on an urban road network model (Fig. 7), we evaluated the performance of the NTuCab dispatch system, and showed that, even on a basic infrastructure [Fig. 3(b)], the distributed multiagent system approach is promising in terms of significant improvements in operational efficiency over an existing centralized approach (see Section II-A). We also showed that incorporating a simple negotiation speedup heuristic (see Appendix) could raise the efficiency further.

To approximate more towards the ideal situation OM-TC in real-time so as to better manage the overall efficiency of NTuCab dispatch, future work includes investigating the following issues:

1) Intelligent taxi roaming: Using historical service demand data might lead to better advisory for available taxi drivers to roam more intelligently to match the distributivity of service requests.

2) NTuCab dispatch queue pre-processing: Re-grouping based on physical proximity between requests and taxis in the respective queues, prior to the start of every dispatch cycle, might better match the geographic distributivity of service requests to that of the available taxis.

The impetus is to influence and better match the physical distributivity between service demand and available taxi supply in real-time.

In conclusion, leveraging on the shortest-time paths computed using real-time traffic information [2], the proposed NTuCab dispatch system can potentially achieve high efficiency, particularly in limiting customer waiting time provided the demand for taxi service is manageable for a fleet size. Explicated in this paper are, in our opinion, important theoretical and empirical insights about our proposed multiagent approach. These fundamental insights would serve as a base reference in further research and development on automating distributed taxi dispatch, including the formulation and investigation of new A-QoS formulas to incorporate human drivers’ preferences.

APPENDIX

Assignment Initialization Heuristic: The negotiation speed of MA^3-LM depends on the initial assignment. In attempting to hasten the negotiation, a simple heuristic (labelled H-Max) is presented herein. Intuitively, the heuristic attempts to set the initial (assignment) solution, with each agent \( a \in \mathcal{A} \) selecting, as far as it is possible, a self-optimal but different request \( r \in \mathcal{O} \) to begin negotiation with. Logically, it can be said to contribute one negotiation round but the overall process can become faster. The details, with \( |\mathcal{A}| = |\mathcal{O}| = N \geq 2 \), are as follows:

**Heuristic H-Max:** One of the agents in group \( \mathcal{A} \) is designated as the speedup initialization (SI) agent, and this fact is made known to all agents at the outset. Every agent \( a \in \mathcal{A} \) selects and proposes to the SI agent a request \( p \in \mathcal{O} \) that satisfies the following:

\[
d[a,p] = \max\{d[a,r] | r \in \mathcal{O} \}
\]  

(7)

The SI agent, upon receiving all such proposals (including its own), will evaluate them as follows. For each request \( r \in \mathcal{O} \):

- It will approve an agent’s proposal (i.e., request selection) if it is the only agent selecting request \( r \in \mathcal{O} \).
- If two or more agents select the same request \( r \in \mathcal{O} \) in their proposal, it will approve the proposal by an agent that offers the highest A-QoS value among them.

Assume that \( F \geq 1 \) agents have had their proposals approved. Then following the proposal evaluation, if \( (N - F) \geq 1 \), the SI agent will arbitrarily allocate each of the remaining \( (N - F) \) agents with a different request taken from the \( (N - F) \) unselected requests.
ACKNOWLEDGEMENT

The authors acknowledge the technical support of Mrs Siew Lai Ng-Goh, Irene from the Parallel and Distributed Computing Center (PDCC), School of Computer Engineering, Nanyang Technological University (NTU), Singapore, including the installation of the MITSIMLab simulator software on a Unix workstation.

REFERENCES


