# Queueing networks for supporting container storage and retrieval

Pasquale Legato and Rina Mary Mazza
*Department of Informatics, Modeling, Electronics and System Engineering, University of Calabria, Rende, Italy*

Received 23 January 2023
Revised 10 March 2023
Accepted 20 March 2023

## Abstract

**Purpose** – An integrated queueing network focused on container storage/retrieval operations occurring on the yard of a transshipment hub is proposed. The purpose of the network is to support decisions related to the organization of the yard area, while also accounting for operations policies and times on the quay.

**Design/methodology/approach** – A discrete-event simulation model is used to reproduce container handling on both the quay and yard areas, along with the transfer operations between the two. The resulting times, properly estimated by the simulation output, are fed to a simpler queueing network amenable to solution via algorithms based on mean value analysis (MVA) for product-form networks.

**Findings** – Numerical results justify the proposed approach for getting a fast, yet accurate analytical solution that allows carrying out performance evaluation with respect to both organizational policies and operations management on the yard area.

**Practical implications** – Practically, the expected performance measures on the yard subsystem can be obtained avoiding additional time-expensive simulation experiments on the entire detailed model.

**Originality/value** – As a major takeaway, deepening the MVA for generally distributed service times has proven to produce reliable estimations on expected values for both user- and system-oriented performance metrics.

**Keywords** Queueing networks, Simulation, Mean value analysis, Container terminal

**Paper type** Research paper

## 1. Introduction

In a container terminal of pure transshipment, an optimal management of stacking and retrieval operations at the storage yard has to be pursued as a major goal of system performance. It provides the key to guarantee a timely coordination of different container handling equipment and policies involved in the vessel discharge/loading (D/L) process. Clearly, an inefficient management of yard capacity and internal container transfer can often lead port operators to face serious and complex operational challenges (Carlos *et al.*, 2014), especially when both import and export flows need to be handled concurrently.

The basic measure of performance for a well-performing D/L process relies on the timely and seamless flow of containers between the interacting terminal subsystems that operate at different time scales. The (average) quay crane productivity and/or the number of cranes allocated to an individual vessel is a target that is usually fixed by contractual conditions between the terminal operator and any given shipping line. The former party may have a degree of freedom in changing the number of assigned cranes and shuttle vehicles over the working shifts during which a specific vessel is berthed. This may be carried out by monitoring the status of the vessel's related D/L process in which activity progress is basically affected by both container transfer between the quay and yard areas and container stacking/retrieval operations on a specific part of the storage yard. For this reason, managing



This work was partially supported by Italy's Ministry of University and Research, through ministerial decree n°1062 of August 10, 2021, within the NOP on Research and Innovation 2014–2020 programme, in cooperation with the University of Calabria and PAC2000A Società Cooperativa (grant number: n°05-I-14893-1).

Maritime Business Review
Vol. 8 No. 4, 2023
pp. 301-317
Emerald Publishing Limited
2397-3757
DOI 10.1108/MABR-01-2023-0009

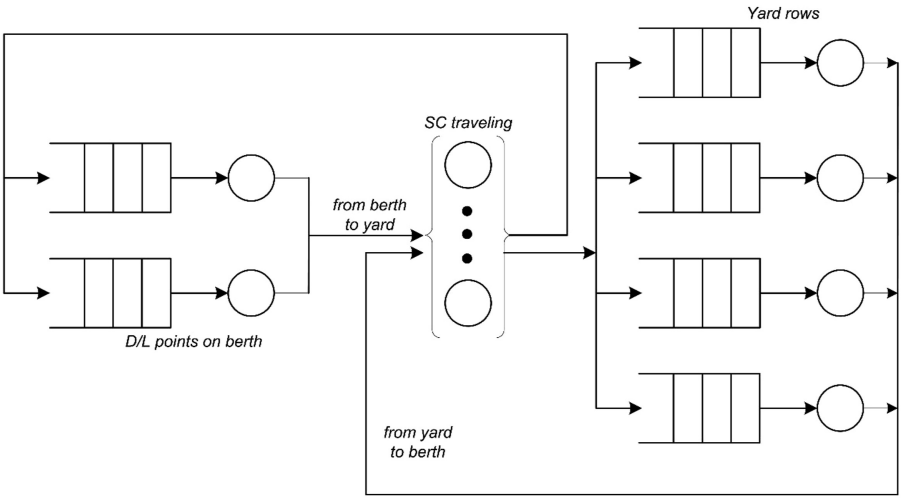
the path of shuttle vehicles through the transfer area and avoiding congestion during travel time as well as service blocking and resource starvation is mandatory.

Feeding, in a timely manner, both the quay and the container storage yard is crucial for achieving a time-effective configuration for the logistic process at hand. Queueing networks appear as the natural conceptual model at the basis of any modeling effort aimed to produce an effective time-oriented tool for performance evaluation. After some decades of research efforts (Dragović *et al.*, 2017; Legato and Mazza, 2020), the adoption of discrete-event simulation (DES) rather than analytical queueing approximations has become a common choice for practice-oriented studies. Unfortunately, it often occurs that simulation users are attracted by the possibility of providing very detailed representations of their systems. So, they are faced with time-consuming activities for model development, verification, execution and statistical analysis of the output from a number of simulation runs which, let alone, is not easy to decide. This stated, our research effort addresses the challenge of proposing a two-level queueing-based hierarchical model. Simulation is used for the one or more subsystems at the inner level, while analytical solutions can be pursued for the outer model. Specifically, the outer model acts as a simplified queueing network where some artificial pure delay stations are inserted to receive the defining parameters from the results of the simulation carried out for the inner model. This allows the user to repeatedly work with the whole detailed fine-grained simulator once it has been developed, verified and validated. What-if experiments on the subsystem of interest can be performed by quickly screening and selecting candidate alternatives of subsystem (re)design. Hence, expensive and detailed simulation of the integrated conceptual queueing network will be reconsidered only to refine policies and resource allocation schema for a few (re) configurations.

Turning our attention to the model, generally speaking, container operations on the yard area of a human-operated transshipment hub are triggered by the activities taking place in the bordering quay and transfer subsystems. The occurrence of container discharge on the quay first calls for container transfer and then storage on the yard; vice versa, container loading on the quay first requires container retrieval from the yard and transfer to the quay afterward. As a result, interaction among these areas is provided by the (hopefully) seamless flow of containers generated by straddle carrier vehicles (SCs). While performing container handling and transfer operations, SCs are viewed as a finite population of customers performing round trips in a closed queueing network. This network features a central pure delay station acting as a finite source of customers for both the quay subsystem and the yard subsystem and two subsystems each of which consists in its own subset of single-server stations under a first-come-first-served (FCFS) service discipline. A sample network integrating the quay and the yard subsystems is shown in Figure 1 (Legato and Mazza, 2022).

Given the above network, we wish to contribute in widening the comprehension of its potentialities for the performance analysis of the real system of interest when having to account for strategic decision-making on (re)designing and managing the yard area organized in parallel storage rows accessed by human-operated SCs. To do so, we investigate the practical accuracy of combining analytical mean value analysis (MVA) with a detailed event-based simulation for computing system throughput and time duration of round trips, as well as queue lengths and waiting times arising at internal congestion points. This is our contribution to the practice of applying queueing networks for time-effective decision-making in container terminals.

The rest of the paper is organized as follows. The literature on the use of analytically solved queueing networks for the yard area is reviewed in Section 2. An MVA-based approximation for solving the queueing network that models container stacking/retrieval operations on the yard of a human-operated maritime terminal is proposed in Section 3.




**Source(s):** Authors' work

**Figure 1.** Operations and queues in a container transshipment hub

How simulation can support the above MVA approach is described in Section 4 and illustrated with numerical experiments. Conclusions are drawn in Section 5.

## 2. Literature review

Any issue about efficient yard management in container terminals would be incomplete if it did not mention the support provided by the use of operations research methods and models, for some time now (Steenken *et al.*, 2004). In order to allow for more informed decision-making, timely problem-solving and improved efficiency in both operations planning and management, mathematical models and solution approaches have been adopted for describing, controlling and optimizing yard organization and practices. Among these, queueing-focused literature has often delivered analytical models that, if well tuned, can be pursued to produce an exact or, in most cases, a good approximation of the metrics of the terminal's (sub)system for a variety of sets of input parameters. To name a few, a semi-open queueing network model with bulk arrivals (by train), shared stack crane resources and multiclass containers is used in the study by Roy *et al.* (2022) to analyze the impact of prioritizing in stacking and internal transport handling the containers delivered by train. A closed-queueing network is proposed in the study by Xiang *et al.* (2022) to estimate the performance of an automated container terminal under traffic congestion, unbalanced task assignment, container batch arrival and different berth and yard layouts. An M/M/c queueing formulation is used in the study by Hasani Goodarzi *et al.* (2022) to model a cross-docking yard and focus attention on the receiving doors. A Jackson-type network approach is presented in the study by Ansorena (2020) to analyze port operations from vessel to yard. Focus is on both the berthing line and the service level at the storage yard. Batch arrivals of containers at a storage yard are considered in the study by Meštrović *et al.* (2018) with a multiserver queue $M^X$/M/c, under c yard cranes. Numerical experiments are returned for a different number of cranes which varies from 1 to 3.

On the other hand, complexity and many nonstandard operational features characterize the yard and the logistic processes carried out therein. As a result, an analytical-based solution approach of a yard-related queueing model is likely to be not accurate. This might

lead to using the queueing model as a mere conceptual representation, while the analytical solution might be pursued for computational convenience alone.

For the reasons mentioned beforehand, the analysis of nontractable queueing-based models is susceptible to the use of more cost-intensive and/or time-consuming solution approaches. For example, a simulation-based truck appointment mechanism is proposed in the study by Shao *et al.* (2022), in conjunction with a DES model to balance truck arrival peak and reduce truck turnaround time via a tristage queueing system, while considering the interference of yard occupation brought by vessel discharging/loading operations. An integrated queueing network is presented in the study by Legato and Mazza (2018) as the natural modeling paradigm for a decision support system aimed at highlighting and quantifying queueing-related phenomena experienced in real practice by container transfer and handling equipment on both the quay and yard. The traffic system in an automated container terminal is modeled in the study by Zhou *et al.* (2017) as a network of servers that represents both paths and junctions. Rather than using an analytical queueing model, they resort to simulation to assess how traffic is affected by the nature of the job sequence in the specific container terminal, the number of vehicles deployed and the respective yard planning strategies. An algorithm is proposed in the study by Singgih *et al.* (2016) as the combination of a modified Dijkstra's algorithm for finding the shortest quay–yard time paths and a queueing theory for calculating the waiting times during the travel.

Unlike the previous efforts in the literature, our contribution is based on the idea of leveraging the combination of queueing and simulation as complimentary approaches to provide quality and time-effective solutions in logistic platforms.

## 3. Mean value analysis of queueing networks

Modeling daily practices in a logistic platform requires organizing and controlling internal handling and transfer operations in order to pursue both system-oriented and customer-oriented performance targets. Mean values for both system throughput and customer waiting time, with reference to completed operations, are the two typical indicators. To this purpose, an approximate, yet fast evaluation of the above performance measures is well appreciated. This may be pursued by the analytical solution of a queueing-based conceptual network model because it allows the analysis of internal points of congestion (service stations) throughout the logistic process of interest.

Focusing on a generic queueing network, nowadays practitioners could even resort to the so-called multiclass population of customers, where all the customers are partitioned into multiple groups. Each group is characterized by its own routing probabilities among the system resources, as well as its own service duration and service discipline at any given service station. On the other hand, a single-class network, such as the one focused by the queueing network in Figure 1, is the straightforward first-step model to provide for a system-oriented performance analysis.

By adopting a single-class or a multiclass closed network of the Baskett-Chandy-Muntz-Palacios (BCMP) type (Baskett *et al.*, 1975), one could provide an exact solution to a model where performance measures may be computed exactly, but only once that several real features are relaxed. This is because the following set of simplifying assumptions is required in order for the exact solution to hold in the analytically tractable "product form" (Bolch *et al.*, 2006):

(1) Customers are routed among service stations according to a discrete-time Markov chain embedded at the instant of each customer departure from any given station;

(2) Customers belonging to different classes are routed among service stations according to different Markov chains;

(3) There are no restrictions on admitting customers to the queue upon their arrival, except for the case of adopting a rejection policy featuring random re-routing;

(4) Service stations must feature either single servers with a fixed or load-dependent service rate and/or (identical) multiple servers, but with fixed service rates and, in no case, service interruptions nor server vacations are allowed;

(5) Customers can be serviced in the FCFS order (besides random order and last-come-first-served), but the same exponential distribution is obliged for service times to customers belonging to different classes;

(6) No priority disciplines for preemptive or non-preemptive services are allowed among customer classes.

Expected performance measures, such as throughputs and waiting times in queue, can be obtained for each service station starting by the MVA (Reiser and Lavenberg, 1980) algorithm. MVA is very easy to implement, in the version covering just a single-class population under a single fixed-rate server at any given station and one or more pure delay (infinite-servers) stations. Despite its popularity among the original community of computer scientists among which it was developed, MVA still needs renewed research efforts to deeper understand the extent to which the underlying "product-form solution" and the so called "arrival theorem" (Lavenberg and Reiser, 1980; Sevcick and Mitrani, 1981) are robust in practice and unavoidable for the theory underlying the exact solution of Markovian queueing networks. After 40 years, the extent to which a service station within the network can be treated as an external one and analyzed in isolation as subjected to a specific stochastic arrival process (state-dependent or not) is still a rather open research topic. The characterization of the stochastic flow of customers among the stations of a closed queueing network with an arbitrary topology and one or more return paths to the station just exited or previously exited by customers during the round trip between their (fixed) origin-destination stations is still an open problem. Despite this, scarcely justified decomposition-based approaches oriented to get an approximated solution of a non-product-form queueing network appear as the unique methods available for practitioners interested in fast, analytical solutions for practical performance evaluation studies in logistics.

### *3.1 Deeper into mean value analysis*

Our contribution to comprehending the potentialities and limitations of the MVA algorithm for the queueing network at hand (Figure 1) is reported here, after the preliminary list of notations used:

(1) $M$: the fixed number of service stations, including one or more pure-delay (infinite-server) stations;

(2) $N$: the fixed number of circulating customers, as the network population;

(3) $\overline{S}_j$: load-independent expected service time for single-server stations under FCFS policy;

(4) $C^2_{S_j}$: coefficient of variation of the random variable modeling the service duration at station $j$;

(5) $\overline{V}_j$: the expected number of visits by a customer to station $j$, between two consecutive visits to any of the D/L point stations;

(6) $\overline{R}_j$: expected residual service time of the customer being serviced upon the arrival instant of a new customer;

(7) $n$: population index, used in iterative computation as the current number of customers, $n = 1, \ldots, N$;

(8) $n_j$: the current number of customers, out of $N$, present at station $j$, $n_j = 1, \ldots, N$;

(9) $n_1, \ldots, n_j, \ldots, n_M$: network state as the joint number of customers found at each station $j$, $j = 1, \ldots, M$;

(10) $P(n_1, \ldots, n_j, \ldots, n_M)$: stationary joint probability of the current network state;

(11) $G(N, M)$: normalization constant for the joint probabilities of network states, under fixed $N$ and $M$;

(12) $\overline{D}_j(N)$: expected delay (waiting + service) at station $j$, under $N$ customers;

(13) $\overline{L}_j(N)$: the expected number of customers found in waiting status upon the arrival instant of a new customer;

(14) $\overline{Q}_j(N)$: the expected number of customers found at station $j$, both in waiting status and being serviced;

(15) $\overline{U}_j(N)$: server utilization, as the probability of the server being busy due to at least one customer in station $j$;

(16) $\overline{X}_j(N)$: throughput of station $j$, as the expected instantaneous rate of customer departures after service;

(17) $\overline{X}(N)$: throughput of the network, as the expected instantaneous rate of customers completing a network round trip

The recursive core computation of the MVA algorithm restricted to queueing stations with a single-server station operating under a fixed-service rate, according to an FCFS service discipline and a single-class customer population ($N$), can be resumed here:

*RECURSIVE_SINGLE-SERVER_MVA (MVA-R)*

$$\overline{D}_j(1) = \overline{S}_j, \ \overline{Q}_j(0) = 0, \quad j = 1, \ldots, M$$

*FOR* $n = 1$ *to* $N$

$$\overline{D}_j(n) = \overline{S}_j\left[1 + \overline{Q}_j(n-1)\right], \ j = 1, \ldots, M \qquad \text{[key delay equation for product-form networks]}$$

$$\overline{X}(n) = n \Big/ \sum_{i=1}^{M} \overline{V}_i \overline{D}_i(n) \qquad \text{[Little's law for network throughput]}$$

$$\overline{Q}_j(n) = \overline{X}(n) \cdot \overline{V}_j \cdot \overline{D}_j(n), \ j = 1, \ldots, M \qquad \text{[Little's law for n° of customers in station]}$$

*END_FOR*

$$\overline{X}_j(N) = \overline{X}(N)\,\overline{V}_j \qquad \text{[forced flow law for station throughput]}$$

$$\overline{U}_j(N) = \overline{X}_j(N)\,\overline{S}_j \qquad \text{[Little's law for single-server utilization]}$$

As one may easily recognize, this implementation returns for each station the expected values of delay (waiting plus service time), the number of customers (waiting and under service), throughput (rate) and server utilization (factor).

This stated, a recursive formula for computing the expected number of customers in a single-server fixed-rate station of a product-form network under a unique class of customers is derived as follows. This formula allows us to obtain the key delay equation in a straightforward manner with respect to Theorem 1 in the study by Reiser and Lavenberg (1980), where the MVA algorithm for closed BCMP queueing networks was originally presented.

From the study by Bolch *et al.* (2006), we rewrite the product-form solution of the network state probability (chapter 8, section 8.1, page 313) according to our notation as follows:

$$P(n_1, ..., n_j, ..., n_M) = \frac{1}{G(N,M)} \prod_{j=1}^{M} \left(\overline{V}_j \overline{S}_j\right)^{n_j}.$$

Focusing on the marginal state-probability formula,

$$P(n_j \geq k|N) = \frac{G(N-k,M)}{G(N,M)} \left(\overline{V}_j \overline{S}_j\right)^k, k = 1, ..., N$$

we particularize the previous one in the following two:

$$P(n_j \geq k|N-1) = \frac{G(N-k-1,M)}{G(N-1,M)} \left(\overline{V}_j \overline{S}_j\right)^k$$

and

$$P(n_j \geq 1|N) = \frac{G(N-1,M)}{G(N,M)} \left(\overline{V}_j \overline{S}_j\right)^1.$$

Multiplying the above two equalities, we find that

$$P(n_j \geq k|N-1)\, P(n_j \geq 1|N) = \frac{G(N-k-1,M)}{G(N,M)} \left(\overline{V}_j \overline{S}_j\right)^{k+1} = P(n_j \geq k+1|N)$$

i.e.

$$P(n_j \geq k+1|N) = P(n_j \geq k|N-1)\, P(n_j \geq 1|N).$$

Besides enabling a new way to get the recursive formula for computing the expected number of customers in a fixed-rate single-server station for a product-form network, this result is interesting due to the similarity with the birth–death process–based formula for isolated queueing stations under Poissonian arrivals.

The recursive formula is derived starting from the definition of the expected number of customers at station j:

$$\overline{Q}_j(N) \hat{=} \sum_{k=1}^{N} n_j P(n_j = k|N) = \sum_{k=1}^{N} P(n_j \geq k|N)$$

$$= P(n_j \geq 1|N) + \sum_{k=2}^{N} P(n_j \geq k|N)$$

$$= P(n_j \geq 1|N) + \sum_{k=1}^{N-1} P(n_j \geq k+1|N)$$

$$= P(n_j \geq 1|N) + \sum_{k=1}^{N-1} P(n_j \geq k|N-1) P(n_j \geq 1|N)$$

$$= P(n_j \geq 1|N) + P(n_j \geq 1|N) \sum_{k=1}^{N-1} P(n_j \geq k|N-1)$$

$$= \overline{U}_j(N) + \overline{U}_j(N)\overline{Q}_j(N-1)$$
$$= \overline{U}_j(N)\left[1 + \overline{Q}_j(N-1)\right]$$

where

$$P(n_j \geq 1|N) = \overline{U}_j(N)$$

is valid under the assumption of no server interruptions/vacations.

In conclusion.

$$\overline{Q}_j(N) = \overline{U}_j(N)\left[1 + \overline{Q}_j(N-1)\right].$$

At this point, the key delay equation used in the MVA algorithm is easily obtained by resorting to Little's law applied to both the time in service (utilization law) and the delay time in station (congestion law):

$$\overline{U}_j(N) = \overline{X}_j(N)\,\overline{S}_j \text{ [utilization law]}$$
$$\overline{Q}_j(N) = \overline{X}_j(N)\overline{D}_j(N) \text{ [congestion law]}$$

$$\Rightarrow \overline{D}_j(N) = \frac{\overline{Q}_j(N)}{\overline{X}_j(N)} = \frac{\overline{U}_j(N)[1 + Q_j(N-1)]}{X_j(N)} = \overline{S}_j(N)[1 + Q_j(N-1)].$$

*3.2 Fixed-point approximation for MVA*

The key delay equation of MVA may be interpreted by saying that the queue length sampled at any given station on customer arrival instants corresponds to the one sampled by a random observer in the network with one less customer. Nevertheless, in a product-form closed queueing network, the stochastic flow of customers circulating through any couple of stations, viewed as a producer–consumer couple, does not correspond to a completely random process with uniformly distributed arrival time instants. Recall that the customers of a BCMP network are routed according to a Markov chain embedded at the instant of each customer departure from any given station. So, here the importance and related impact on numerical results of the dependencies within the (nonrenewal) flow of customers circulating in a closed network can be highlighted by starting from the following relationship:

$$\overline{Q}_j(N) = \overline{Q}_j(N-1) + \Pr\{\text{find a customer at station j}\}.$$

Using a semi-Markov assumption on the customer circulation among the stations of the closed network leads to the following relationship:

$$\Rightarrow \overline{Q}_j(N) = \overline{Q}_j(N-1) + \left[\overline{V}_j\overline{D}_j(N) \Big/ \sum_{i=1}^{M} \overline{V}_i\,\overline{D}_i(N)\right]$$

from which one may easily go from the previous recursive version to a fixed-point version of the core computation for the single-server FCFS MVA.

*FIXED-POINT_SINGLE-SERVER_MVA (MVA-F)*
$\bar{D}_j(N) = \bar{S}_j,\ \bar{\bar{Q}}_j(N) = N/M,\quad j = 1,\ldots,M$
*REPEAT*
$$\bar{D}_j(N) = \bar{S}_j\left[1 + \frac{N-1}{N}\bar{Q}_j(N)\right],\ j = 1,\ldots,M$$
$$\bar{Q}_j(N) = N \cdot \left[\bar{V}_j\bar{D}_j(N) \Big/ \sum_{i=1}^{M} \bar{V}_i\,\bar{D}_i(N)\right],\ j = 1,\ldots,M$$
*UNTIL convg on* $\bar{Q}_j(N), j = 1,\ldots,M$

Observe that the above fixed-point computation procedure does not require the assumption that service times (at FCFS stations) are exponentially distributed. Hence, one may argue that the approximation error on the sojourn time equation arises from the correlation (neglected by the semi-Markov assumption) among the sequence of sojourn times per visit experienced by whatever customer circulating among the service stations. Our numerical experience indicates that the weight of the correlation among sojourn times per visit translates into an error on the expected sojourn time returned by the (approximate) fixed-point core MVA. This error can be quantified between 7 and 17% for expected queue lengths, with a typical value of 10%, while it is only a few percentage points for expected station throughputs.

### *3.3 MVA with nonexponential services*

A point of strength of the (exact) MVA recursive procedure relies in the possibility of extending the key delay equation for $\overline{D}_j(n)$ to the case of nonexponential service times by resorting to the classical paradox of residual life (Heyman and Sobel, 1982) to quantify the residual duration of any given service. Hence, the extended key delay equation for single-server service stations under FCFS policy becomes the following:

$$\overline{D}_j(n) = \overline{S}_j \cdot \overline{L}_j(n-1) + \overline{R}_j \cdot \overline{U}_j(n-1) + \overline{S}_j \, , j = 1, ..., M$$

where

$$R_j \widehat{=} \left(\overline{S}_j/2\right) \cdot \left(1 + C_{S_j}^2\right) \, , j = 1, ..., M.$$

Note that the residual service time at any station j ($\overline{R}_j$) refers to the random variable service duration ($\overline{S}_j$) that defines the renewal process (op. cit.) for an uninterrupted sequence of independent and identical distributed service times. Furthermore, in the paradox, by assumption, sampling instants of residual life follow a uniform distribution, whereas sampling instants corresponding to customer arrivals to a station with a busy server do not. Clearly, the accuracy of the paradox-based relationship for $\overline{D}_j(n)$ still relies on the assumption that the true sampling pattern to a service station has little influence on the expected delay in station, despite the random observer point of view inherent to the paradox of residual life.

For the sake of completeness, the extension of the *RECURSIVE_SINGLE-SERVER MVA* algorithm obtained by incorporating the paradox of residual life is given here:

*NON-EXPONENTIAL_SINGLE-SERVER_MVA (MVA-N)*

$\overline{D}_j(1) = \overline{S}_j, \ \overline{L}_j(0) = 0, \ \overline{U}_j(0) = 0, \quad j = 1, ..., M$

*FOR* $n = 1$ *to* $N$

$\quad \overline{D}_j(n) = \overline{S}_j \cdot \overline{L}_j(n-1) + \overline{R}_j \cdot \overline{U}_j(n-1) + \overline{S}_j, \quad j = 1, ..., M$

$\quad \overline{X}(n) = n \Big/ \sum_{i=1}^{M} \overline{V}_i \, \overline{D}_i(n)$

$\quad \overline{L}_j(n) = \overline{X}(n) \cdot \overline{V}_j \cdot (\overline{D}_j(n) - \overline{S}_j), \quad j = 1, ..., M$

$\quad \overline{U}_j(n) = \overline{X}(n) \cdot \overline{V}_j \cdot \overline{S}_j, \quad j = 1, ..., M$

*END_FOR*

### *3.4 Numerical comparisons*

Our extensive experience on closed queueing networks like the one in Figure 1 suggests that the above approximate MVA-N for single-server FCFS queueing stations under nonexponential service times performs satisfactorily for practical applications. Percentage errors on throughput and queue length are deemed acceptable in real practice, especially when related to lack of input data, poor modeling and/or insufficient

precision on parameters of the queueing network model of interest. For the sake of completeness and illustrative purposes, let us consider an example. The numerical results returned from the recursive MVA algorithm (MVA-R), the fixed-point MVA (MVA-F) and the MVA for the nonexponential FCFS services (MVA_N) are in Tables 1–6, together with 95% confidence intervals obtained by DES. The example is centered on assuming a Cox-2 distribution bearing a coefficient of variation of service duration set to 2 at both the D/L points in Figure 1. Notice that, based on the typical real values observed on the field for a D/L point, this setting can be considered a worst case. The average time to travel from the quay to the yard or vice versa (i.e. the delay for each customer at the pure-delay station) is set equal to 5 min, while the expected container handling time is the same for each of the 4 yard rows in Figure 1 and set equal to 2 min. Assuming now 12 circulating customers (straddle carriers), we set the expected service times at the two D/L points equal to 2 min in the first run of the example and to 4 min in the second run. In so doing, we are able to show

**Table 1.**
Results for server utilization in the sample instance with D/L mean = 2

| | Server_Utilization | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 0.62–0.67 | 0.57 | 0.67 | 0.68 |
| Pure delay | – | – | – | – |
| Yard row 1 | 0.30–0.33 | 0.28 | 0.33 | 0.34 |

**Source(s):** Authors' work

**Table 2.**
Results for queue length in the sample instance with D/L mean = 2

| | Number_In | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 1.94–2.14 | 2.18 | 1.71 | 1.53 |
| Pure delay | 6.04–6.47 | 5.69 | 6.66 | 6.77 |
| Yard row 1 | 0.43–0.46 | 0.39 | 0.48 | 0.49 |

**Source(s):** Authors' work

**Table 3.**
Results for queueing time in the sample instance with D/L mean = 2

| | Time_In_Que | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 4.40–4.79 | 5.68 | 3.14 | 2.82 |
| Pure delay | – | – | – | – |
| Yard row 1 | 0.83–0.90 | 0.72 | 0.88 | 0.90 |

**Source(s):** Authors' work

**Table 4.**
Results for server utilization in the sample instance with D/L mean = 4

| | Server_Utilization | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 0.75–0.81 | 0.68 | 0.83 | 0.85 |
| Pure delay | – | – | – | – |
| Yard row 1 | 0.19–0.21 | 0.17 | 0.21 | 0.21 |

**Source(s):** Authors' work

typical numerical outputs associated to significant and realistic utilization levels (i.e. 60–80%) for the D/L points.

Besides appreciating the reasonable accuracy of the MVA-N algorithm when dealing with a nonexponential service distribution, our major observation on the above tables comes from the combination of the MVA-N, MVA-F and MVA-R results. As a matter of fact, by combining these results, we often obtain an interval that covers the simulation results. At moderate to high congestion levels (i.e. 60–80% realistically pursued at D/L points by the operations manager to avoid crane blocking/starvation), MVA-N tends to overestimate the waiting times and, thus, queue lengths. Therefore, under a fixed population of customers, the station throughput and, thus, utilization will result lower than the exact values at the D/L points because of Little's law. Vice versa, at low congestion levels (i.e. 20–30% realistically pursued at yard rows to avoid locking phenomena), the roles of the analytical algorithms are reversed: MVA-N overestimates small waiting times in the related stations, thus returning underestimated throughput and utilization measures, while MVA-F and MVA-R behave in the opposite direction.

## 4. Supporting MVA-based approximation by simulation

Performance evaluation studies on queueing stations and networks under several difficult features have been addressed in the studies by Dragović *et al.* (2006, 2012) and Legato and Mazza (2020) by approximations and simulation. Here we focus on the storage/retrieval operations occurring in the yard subsystem and embrace a deeper modeling prospective to capture some real rules of container handling and SC movements. To this purpose, the yard-related subnetwork of the queueing network model in Figure 1 is now analyzed by means of a closed network featuring two artificial pure delay stations, $M$ peripheral service stations and a population of $N$ SCs, as shown in Figure 2.

The SCs circulate alternating between the delay stations and the yard subnetwork. Here we do not pursue the classical flow-equivalent server representation of the berth and transfer areas subnetwork, under exponential service times, resulting from a Norton theorem–based reduction (Balsamo and Iazeolla, 1982) of any product-form subnetwork. In the current

| | Number_In | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 3.38–3.72 | 3.71 | 3.42 | 3.33 |
| Pure delay | 3.64–4.01 | 3.34 | 4.14 | 4.26 |
| Yard row 1 | 0.24–0.26 | 0.20 | 0.26 | 0.27 |

**Source(s):** Authors' work

**Table 5.** Results for queue length in the sample instance with D/L mean = 5

| | Time_In_Queue | | | |
|---|---|---|---|---|
| Stations | DES | MVA-N | MVA-F | MVA-R |
| D/L point 1 | 14.1–15.2 | 17.9 | 12.5 | 11.6 |
| Pure delay | – | – | – | |
| Yard row 1 | 0.51–0.56 | 0.39 | 0.47 | 0.52 |

**Source(s):** Authors' work

**Table 6.** Results for queueing time in the sample instance with D/L mean = 4

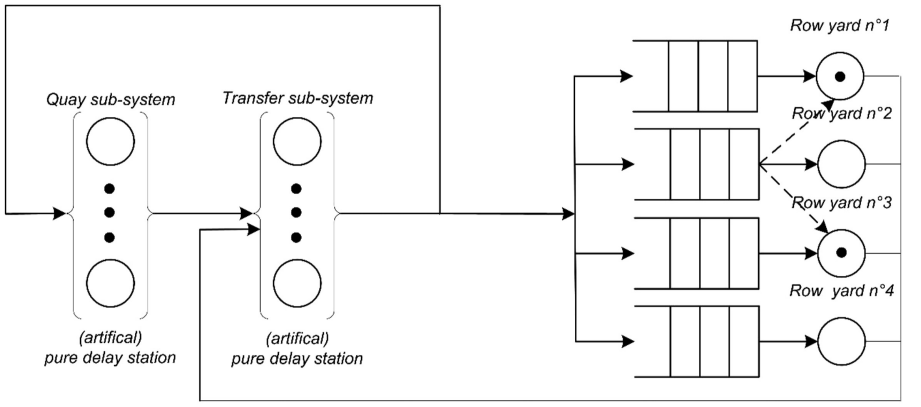



**Figure 2.**
The closed queueing network with two artificial pure delay stations and the yard subsystem

**Source(s):** Authors' work

network representation portrayed by Figure 2, we adopt two artificial pure delay stations. In the first pure delay station (i.e. *quay subsystem*), in principle, each delay server should account for the duration of the activity under the quay crane which requires synchronization between the arrival of a vehicle and the availability of a container or buffer space for vessel discharge or loading, respectively. The second delay station (i.e. *transfer subsystem*) serves the purpose of modeling the number of vehicles simultaneously moving to/from the yard, while not disregarding the mutual dependencies arising among these vehicles (i.e. the active delay servers in the pure delay station). The average time duration for delay servers in both delay stations is provided by means of a model-driven simulator previously verified and validated (Legato and Mazza, 2018).

As for the yard subsystem and the operational rules applied therein, to perform handling and transfer operations, a vehicle must access a row in compliance with security measures that depend on both yard organization and technology. An SC may access a row if no other vehicle is already performing handling operations in that row or in either of the adjacent rows. As shown by the dotted arrows in Figure 2, by entering the central row to stack/retrieve a container, an SC issues a busy condition for that row and a so-called locking condition on the two adjacent rows. The locking condition yields dependency relations among the neighboring network stations, due to service prevention at a given station if an adjacent station has already started to deliver service. This stated, the resulting service time in the yard row may be represented by the stage-type diagram in Figure 3. Once again, simulation is used to provide the probability of locking and the average duration of the locking and container stacking/retrieval service (cit. op.).

As a result, this simulation-fed modeling approach first leads to restoring independency among service stations during service operations, a condition which is at the basis of the analytical tractability of any queueing network. It then paves the way for using analytical solutions, such as MVA-N, to carry out parameter-based analysis by systematically returning quantitative bound computation-oriented guidelines for (sub)system design, whether the (sub)system be existing or soon-to-be. As for the time complexity of the single-class MVA-based analytical solution for the overall network, observe that it is proportional to the product of $M*N$: it loops over the number of stations (inner level) and the number of circulating customers (outer level). So, when going from the four rows in Figure 2 to the actual number of rows in a real-life model, the computational burden remains acceptable.

*4.1 Illustrative numerical example*

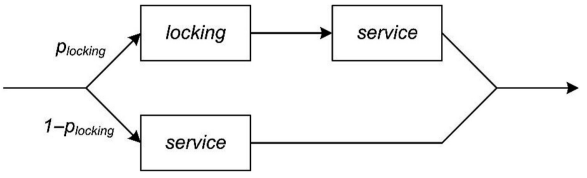
In this section, numerical experiments have been carried out on the network model in Figure 2 with the objective of showing the potentiality of the two approximate MVA-based algorithms (MVA-N and MVA-F) when evaluating the queueing phenomena at the yard row stations. We remark that the user has only one degree of freedom when pursuing the approximate analytical solution of a conceptual queueing model for subsystems and/or stations where nonstandard service mechanisms occur in real practice. In our case, one can only rely on the capability of capturing these nonstandard service features within an ad hoc stage diagram (i.e. the one in Figure 3).

Let us consider two scenarios: in the first one, the number of circulating customers (SCs) has been set equal to 12 units, while in the second one, it has been set equal to 6. The rationale of this choice is to highlight at the row stations, on the one side, the major (intolerable) impact on queue length and time due to the row locking phenomenon under 12 units and, on the other side, the minor (acceptable) impact of row locking under 6 units. Clearly, the delay times at the two pure delay stations as well as the service time, probability of locking occurrence and related duration have been previously estimated by a fine-grained DES. Considering that a simulation is a computer-based statistical sampling experiment, a proper analysis has been carried out on the simulation output data (Law and Kelton, 2000). In particular, to end up with an estimator with a small prespecified (relative) error $\epsilon$, for instance, within 5% of the correct value with $100(1-\delta)\%$ confidence interval, the number of simulation samples $n$ to be taken has been determined by the following formula:

$$n = \left\lceil \left( \frac{z \cdot \overline{S}(n)}{\overline{X}(n) \cdot \epsilon} \right) \right\rceil^2$$

where $\overline{S}(n)$ is the sample standard deviation, $\overline{X}(n)$ is the sample mean and $z$ is the desired quantile value of the standard normal distribution (Nakayama, 2003). With respect to the value of the input settings used for the simulation experiments as specified in Table 7, according to the above formula, $n = 10$ runs have been deemed sufficient to return nonfloating credible values for our simulation experiments. The values of the output performance measures returned by the simulation experiments are reported in Table 8 in terms of estimates for the mean and standard deviation (the latter is not required for the time at pure delay stations). These results serve as input for the MVA-based computations.

Observe that different settings have been provided for the blocking phenomena depending on whether the yard rows are located on border or internal positions. In the former case (i.e. rows 1 and 4), rows are affected by blocking on only one side, whereas in the latter case (i.e. rows 2 and 3), they are affected by blocking on both sides. Thus, the probability of an SC getting blocked as well as the duration of this blocking are considerably different in the two above cases. Note that, for the sake of simplicity, we have assumed a balanced workload on the four yard rows. Thus, we have set equal probabilities




**Source(s):** Authors' work

**Figure 3.**
The stage-type diagram for service time corrupted by row locking time

(0.25) for each row targeted by an SC. So, we can just show numerical results for one of the two border rows (i.e. row 1) and again one of the two internal rows (i.e. row 2) in Tables 9 and 10. One may recognize that more than 6 out of 12 (6.57) SCs are cumulatively expected at the two internal rows, while only two SCs (2.164) are cumulatively expected at the two border rows. This is sufficient to conclude that opting for 12 SCs is an unfeasible decision due to the lack of physical space for vehicle queueing at the front of and near the two internal rows. From the complete analytical results, we have further registered that (on average).

(1) 0.5 SCs at each D/L point (i.e. first pure delay station) represent an intolerable blocking/starvation phenomenon for the quay cranes;

**Table 7.**
System settings for simulation

| Subsystem | Resource | N° | Service time |
|---|---|---|---|
| Quay | D/L points | 2 | Erl(16,100) [s] |
| Transfer | straddle carriers | 12 | 4 [m/s] (loaded)<br>5 [m/s] (unloaded) |
| Yard | yard rows | 4 | p1*Erl(16,30) + p2*Erl(16,90) + p3*Erl(16,150) [s] |

**Source(s):** Authors' work

**Table 8.**
System settings returned by simulation for the MVA algorithm

| | 12 SC units | | 6 SC units | |
|---|---|---|---|---|
| Feature | Mean | St. Dev | Mean | St. Dev |
| Delay in quay subsystem | 30.3 | – | 30.3 | – |
| Delay in transfer subsystem | 98.2 | – | 98.2 | – |
| Service time on yard rows | 60.2 | 44 | 60.2 | 44 |
| P (SC locked at row 1 ∨ 4) | 0.35 | – | 0.21 | – |
| P (SC locked at row 2 ∨ 3) | 0.49 | – | 0.39 | – |
| Locking time at row 1 ∨ 4 | 289 | 256 | 108 | 117 |
| Locking time at row 2 ∨ 3 | 379 | 308 | 126 | 141 |

**Source(s):** Authors' work

**Table 9.**
Two-way comparison of the results for the queue length

| | Queue length (12 SCs) | | Queue length (6 SCs) | |
|---|---|---|---|---|
| Stations | MVA-N | MVA-F | MVA-N | MVA-F |
| Border row | 1.08 | 1.03 | 0.43 | 0.44 |
| Internal row | 3.28 | 3.44 | 0.65 | 0.70 |

**Source(s):** Authors' work

**Table 10.**
Two-way comparison of the results for the queue time (s)

| | Queue time (12 SCs) | | Queue time (6 SCs) | |
|---|---|---|---|---|
| Stations | MVA-N | MVA-F | MVA-N | MVA-F |
| Border row | 154 | 151 | 26.3 | 29.4 |
| Internal row | 687 | 789 | 53.2 | 64.5 |

**Source(s):** Authors' work

(2) roughly 3 (2.84) out of 12 SCs are traveling (i.e. second pure delay station) between their D/L point and target yard row, but more traffic could be allowed causing particular congestion along the way.

Due to the above findings, the user may decide to see what happens when reducing the SCs from 12 to 6. To this purpose, a second set of values has been reported in Tables 9 and 10. The reduction of the queue lengths at the rows and an even more significant reduction of the blocking/starvation at the D/L point represent well-appreciated results for the operations manager. This appreciation lies in the need to properly manage the number and utilization time of equipment and human gangs supporting the (expensive) quay cranes on the given D/L points.

Thinking of how the yard row is fed by the completion rate of the D/L operations in the quay subsystem (roughly an SC departure every 70 and 60 s for the scenario featuring 12 and 6 SCs, respectively), a third comparison on an additional performance indicator may be found in Table 11. This index is the busy factor for both border and internal yard rows. It is meant to measure the (expected) percentage of time during which a row is either (*1*) locked by the service occurring in an adjacent row or (*2*) busy providing service to an SC in that same row. Clearly, the greater the busy factor of a specific row, the greater the concentration of storage/retrieval operations on that row or on the adjacent ones.

For the sake of completeness, note that a different evaluation of the time in queue has been provided by the MVA-N and MVA-F algorithms for the internal rows. This difference is due to the weight of the coefficient of variation for the "inflated" service time (i.e. locking and service time) which is explicitly taken into account by the paradox of residual life formula in the MVA-N procedure.

## 5. Conclusions

A fast MVA of queueing networks is required for a first-order approximate performance analysis of complex logistic systems in order to overcome time-consuming simulation experiments in real environments. Usually, scheduling policies and allocation schema are evaluated and (re)organized to pursue a time-effective system management. After more than 40 years of modeling efforts, one has to recognize that congestion arising from resource sharing under the occurrence of resource locking, starvation and blocking phenomena cannot be captured by heuristic adjustments on analytical formulas derived for product-form queueing networks. So, a revisitation of some analytical results for product-form queueing networks and related approximations covering nonexponential service times have been investigated. Then, a hierarchical combination of simulation with analytical solution methods may offer satisfactory achievements. Partial results from DES have been fed into a simplified analytical model for storage and retrieval operations of containers in a maritime transshipment hub. Numerical results encourage more research efforts along this avenue to the aim of specializing the classical simulation metamodeling approach through the adoption of fast-to-solve, yet accurate analytical queueing networks.

| Stations | Busy factor (12 SCs) | | Busy factor (6 SCs) | |
|---|---|---|---|---|
| | MVA-N | MVA-F | MVA-N | MVA-F |
| Border row | 0.54 | 0.53 | 0.31 | 0.32 |
| Internal row | 0.85 | 0.83 | 0.43 | 0.44 |

**Source(s):** Authors' work

**Table 11.** Two-way comparison of the results for the busy factor

## References

Ansorena, I.L. (2020), "Operational strategies for managing container terminals. An approach based on closed queuing networks", *International Journal of Industrial and Systems Engineering*, Vol. 35 No. 1, pp. 13-27.

Balsamo, S. and Iazeolla, G. (1982), "An extension of Norton's theorem for queueing networks", *IEEE Transactions on Software Engineering*, Vol. SE-8 No. 4, pp. 298-305.

Baskett, F., Chandy, K.M., Muntz, R.R. and Palacios, F. (1975), "Open, closed, and mixed networks of queues with different classes of customers", *ACM Journal*, Vol. 22 No. 2, pp. 248-260.

Bolch, G., Greiner, S., de Meer, H. and Trivedi, K.S. (2006), *Queueing Networks and Markov Chains*, John Wiley & Sons, New Jersey, NJ.

Carlos, H.J., Vis, I.F. and Roodbergen, K.J. (2014), "Storage yard operations in container terminals: literature overview, trends, and research directions", *European Journal of Operational Research*, Vol. 235 No. 2, pp. 412-430.

Dragović, B., Park, N.-K. and Radmilović, Z. (2006), "Ship-berth link performance evaluation: simulation and analytical approaches", *Maritime Policy and Management*, Vol. 33 No. 3, pp. 281-299.

Dragović, B., Park, N.-K., Zrnić, N.Đ. and Meštrović, R. (2012), "Mathematical models of multiserver queuing system for dynamic performance evaluation in port", *Mathematical Problems in Engineering*, Vol. 2012, pp. 1-19, 710834.

Dragović, B., Tzannatos, E. and Park, N. (2017), "Simulation modelling in ports and con- tainer terminals: literature overview and analysis by research field, application area and tool", *Flexible Services and Manufacturing Journal*, Vol. 29 No. 1, pp. 4-34.

Hasani Goodarzi, A., Diabat, E., Jabbarzadeh, A. and Paquet, M. (2022), "An M/M/c queue model for vehicle routing problem in multi-door cross-docking environments", *Computers and Operations Research*, Vol. 138, 105513.

Heyman, D.P. and Sobel, M.J. (1982), *Stochastic Models in Operations Research, Volume I: Stochastic Processes and Operating Characteristics*, McGraw-Hill, New York, NY.

Lavenberg, S.S. and Reiser, M. (1980), "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers", *Journal of Applied Probability*, Vol. 17 No. 4, pp. 1048-1061.

Law, A.M. and Kelton, W.D. (2000), *Simulation Modeling and Analysis*, 3rd ed., McGraw-Hill, New York, NY.

Legato, P. and Mazza, R.M. (2018), "A decision support system for integrated container handling in a transshipment hub", *Decision Support Systems*, Vol. 108, pp. 45-56.

Legato, P. and Mazza, R.M. (2020), "Queueing analysis for operations modeling in port logistics", *Maritime Business Review*, Vol. 5 No. 1, pp. 67-83.

Legato, P. and Mazza, R.M. (2022), "Performance evaluation of container handling in a transshipment hub", in Dragović, B., Zrnić, N., Chen, G. and Papadimitriou, S. (Eds), *Proceedings of the Maritime and Port Logistics of the XXIV International Conference on Material Handling, Constructions and Logistics (MHCL 2022), Bar, Montenegro*, SaTCIP Publisher, Vrnjačka Banja, Serbia, pp. 13-17.

Meštrović, R., Dragović, B., Zrnić, N. and Dragojević, D. (2018), "A relationship between different costs of container yard modelling in port using queuing approach", *FME Transactions*, Vol. 46 No. 3, pp. 367-373.

Nakayama, M.K. (2003), "Analysis of simulation output", in Chick, S., Sánchez, P.J., Ferrin, D. and Morrice, D.J. (Eds), *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, IEEE, pp. 49-58.

Reiser, M. and Lavenberg, S.S. (1980), "Mean value analysis of closed multichain queueing networks", *ACM Journal*, Vol. 27 No. 2, pp. 313-322.

Roy, D., van Ommeren, J.-K., de Koster, R. and Gharehgozli, A. (2022), "Modeling landside container terminal queues: exact analysis and approximations", *Transportation Research Part B: Methodological*, Vol. 162, pp. 73-102.

Sevcick, K.C. and Mitrani, I. (1981), "The distribution of queueing network states at input and output instants", *Journal of the Association for Computing Machinery*, Vol. 28 No. 2, pp. 358-371.

Shao, Q., Huang, M., Zhang, S. and Zhang, Y. (2022), "Simulation of truck arrivals at container terminal based on the interactive truck appointment system", *International Journal of Shipping and Transport Logistics*, Vol. 14 Nos 1-2, pp. 141-171.

Singgih, I.K., Hong, S. and Kim, K.H. (2016), "Flow path design for automated transport systems in container terminals considering traffic congestion", *Industrial Engineering and Management Systems*, Vol. 15 No. 1, pp. 19-31.

Steenken, D., Voß, S. and Stahlbock, R. (2004), "Container terminal operation and operations research — a classification and literature review", *OR Spectrum*, Vol. 26, pp. 3-49.

Xiang, X., Liu, C., Lee, L.H. and Chew, E.P. (2022), "Performance estimation and design optimization of a congested automated container terminal", *IEEE Transactions on Automation Science and Engineering*, Vol. 19 No. 3, pp. 2437-2449.

Zhou, C., Lee, L.H., Chew, E.P. and Li, H. (2017), "A modularized simulation for traffic network in container terminals via network of servers with dynamic rates", in Chan, W.K.V., D'Ambrogio, A., Zacharewicz, G., Mustafee, N., Wainer, G. and Page, E. (Eds), *Proceedings of the 2017 Winter Simulation Conference*, Las Vegas, IEEE, pp. 3150-3161.

**Corresponding author**
Pasquale Legato can be contacted at: legato@dimes.unical.it