SIMULATION AT A MARITIME CONTAINER TERMINAL: MODELS AND COMPUTATIONAL FRAMEWORKS

Pasquale Legato  
Dipartimento di Elettronica, Informatica e Sistemistica  
University of Calabria  
Via P. Bucci, Cubo 41C, 87036, Rende (CS), Italy  
E-mail: legato@deis.unical.it

Daniel Gullì and Roberto Trunfio  
Centro di Supercalcolo per l’Ingegneria Computazionale  
CESIC – NEC Italia S.r.l.  
Via P. Bucci, Cubo 22B, 87036, Rende (CS), Italy  
E-mail: {daniel.gulli, roberto.trunfio}@eu.nec.com

Riccardo Simino  
High Performance Computing Division  
NEC Italy S.r.l.  
Via L. Da Vinci, 97, 20090, Trezzano sul Naviglio (MI), Italy  
E-mail: riccardo.simino@it.neceur.com

KEYWORDS
Maritime container terminals, holistic approach, discrete-event simulation, parallel computing.

ABSTRACT

NEC-Italy high-performance computing division recognises that computer simulation aided organisation and management is the true challenge for a new generation of advanced decision tools for supporting operations in modern logistic platforms. Collecting academic expertises, research skills, computing machinery and dynamic realities with a significant growth into the field of container terminal logistics is a new R&D project at the CESIC-NEC center at the University of Calabria. The starting idea of this project is that discrete-event simulation is the best modelling approach to manage the complexity of logistic processes at container terminals. For these time-based systems, operating in a stochastic environment, becomes crucial to highlight both congestion and starvation phenomena embedded into logistic processes, in order to achieve reasonable targets of a good management of resources and, therefore, stay on the market. Here we present some queuing network based representations that are at the basis of an integrated simulation model under development. Since large and complex models are affected by a high burden on execution, we also remark the benefits of parallel and/or distributed computational frameworks. Numerical results on parallel analysis of simulation output data are given.

INTRODUCTION

The increasing international division of labour in the course of liberalisation, the resulting trade movements and the consequent need for faster, just in time freight transportation and cargo handling in real time lead transportation companies to study and develop more efficient and effective transport systems able to follow customers on their demand. Computing system companies may have a crucial role in this business, by providing sophisticated decision support systems on powerful, advanced computational resources and platforms.

The common, basic idea in freight transportation is that of resorting to the aggregation of multicommodity flows to be transported, by large carriers, for long distance routes across a certain number of intermediate logistic platforms (terminals), where freight consolidation may occur together with store and forward logistic functions and other source/sink based functions. Standardised containerization appears as the most notable and consolidated practice for freight transportation, especially through long distance maritime routes. Containerization is a system of intermodal freight transport using standardized large bins known as containers that can be loaded and sealed intact using multiple modes of transportation: ship, rail, road and air. Containers are an ISO standardized metal box of 8-ft wide by 8-ft high; the most common lengths are 20-ft and 40-ft. The container capacity is often expressed in twenty-foot equivalent unit (TEU). Containers are made out of steel and can be stacked on top of each other.

Nowadays, approximately 90 percent of the non-bulk cargo worldwide moves by containers stacked on transport. On ships they are typically stacked up to seven units high and there are ships that can carry over 9,000 TEUs. The world container fleet amounts to about 23.2 million TEUs and the container throughput reached 440 million TEUs in 2006 (UNCTAD 2007).

Maritime container terminals are the most important crossroads for transshipment and intermodal container transfers. The maritime container shipment follows the spokes-hub distribution paradigm: containers are shipped from a port (spoke) to another one through a small number of maritime transshipment terminals (hubs). Both spokes and hubs could be connected with inland container terminals by road and rail (the so called intermodal transfer). In opposition with the high number of container ports in the world, there are a few
of transshipment terminals (e.g., the Gioia Tauro terminal). Transshipment container terminals are large facilities for intermodal transport, able to handle millions of containers per year and berth large container vessels. These hubs are linked with spokes by means of small vessels called feeders, through minor (or short sea trade) routes. There are a few major liner trade routes that link the hubs (or deep sea trade routes). There are also oceanic container vessels, known as mother vessels, which sail on these intercontinental routes.

A maritime container terminal is a large and complex logistic platform organised around a set of logistic processes. The logistic activities at a container terminal often belong to more logistic processes, which require a whole vision of the system to be properly organised. This fact is critical for a good management of the system and the choice of the system modelling approach: an efficient and effective management of logistic activities in a container terminal can decrease the operating costs and service times and increase the quality of services in order to achieve a better market position. In this paper we report on the development of a discrete-event simulation platform for the whole management of a maritime container terminal, as it is ongoing at the CESIC-NEC research laboratory in southern Italy. The availability of a simulation environment where decision models could be developed and tuned according to a final customisation phase is considered as a novel, challenging market option. Simulation modelling and analysis should become an ordinary planning and control tool, e.g., to evaluate the performance of the container terminal when changes occur in the system configuration (what-if analysis) or to use simulation-based optimization techniques to optimize the whole system (Law and Kelton 2000).

As matter of fact, the whole model of a large and complex system, e.g. a container terminal, is affected by a high burden on execution. This statement is true especially for the most commercial simulators that are generally based on the process interaction worldview (Derrick et al. 1989). As recently suggested (Taylor et al. 2002), it is preferable to integrate heavy simulation modelling into a computational framework oriented to high-performance computing (HPC).

In the following, we classify and describe the logistic problem at hand. Afterward, we depict the queuing network model used to represent the core logistic processes at a container terminal. At the end, we discuss why HPC computational frameworks are a key added value into the study of real systems through discrete-event simulation

**PROBLEM CLASSIFICATION**

Vis and De Koster (2003) produced an interesting overview paper in the area of container logistics that gives a classification of the logistic processes in modern container terminals: i) arrival of the ship, ii) discharging and loading of the ship, iii) transport of containers from ship to stack and vice versa, iv) stacking of containers, and v) inter-terminal transport and other modes of transportation. With respect to this classification, it is possible to identify a set of features that are common to many maritime terminals. The main difference between the greater part of container terminals located in Europe and North America and those in the Asia-Pacific region regards the logistic processes iii) and iv). The latter relies on the “Direct Transfer System” (ITS), in which process iii) and iv) are closely connected: a fleet of shuttle vehicles transports the containers from a vessel to beside the stack area (and vice versa) while dedicated cranes stack containers in compact regions. European and North-American container terminals are based upon the “Direct Transfer System” (DTS) in which process iii) and iv) are performed by the same actors: in this case a fleet of shuttle vehicles (called straddle carriers) moves the containers from a vessel to the storage area (and vice versa) for container stacking (retrieval) operations into (from) the slots assigned in the storage area (Cordeau et al. 2007). In our study we refer to a DTS maritime container terminal.

**A SIMULATION PLATFORM FOR A MARINE CONTAINER TERMINAL**

To fix ideas, we give a brief description of a real case that we are familiar with: the Gioia Tauro terminal. This terminal, located in southern Italy, since it opened in 1995 has become in just a few years the largest transshipment port on the Mediterranean Sea. Its management has been early characterised by significant efforts towards an increasing computer aided organisation and control of the various logistic processes. Recently a manager friendly simulator has been designed for studying the congestion phenomenon at the port-input channel, the port admission policy for the newly arrived vessels in the roadstead and the allocation policies of berthing slots (Canonaco et al. 2007). Now, the problem of achieving an integrated management based on the usage of simulators and other operations research tools becomes more and more important.

The terminal is a large facility composed by: a harbour entrance of 250m wide and 18m water depth followed by a large roadstead for incoming vessels; 2 pilot boats; a 3,100m quay length with along a channel of multiple water depth ranging from 13.5 to 15.5 m (the quay is discontinuous, thus it is usually decomposed into two sub-quays); 18 quay rail-mounted gantry cranes (RMGs); a fleet of 75 straddle carriers (SCs), handling vehicles used to transfer the containers between the quay cranes and the yard; and a yard surface of 1.1 million square meters that can store nearly 59,000 TEUs. The yard has 32 sectors parallel to the quay and organized in two lines. An average quay has 32 lines, each containing 16 slots. A slot can host two one-TEU containers stacked one on top of the
other. The layout of the container terminal is depicted in Figure 1.

**Description of the Logistic Processes**

The logistic processes are described in the following. Planners of the terminal company construct a “berth schedule” on a weekly basis, which shows the berthing time and position for each incoming vessel according to the ETA (Expected Time of Arrival) and DTD (Due Time of Departure) of the vessel and the preferred sub-quay. This is the so called Berth Planning Problem (Park and Kim 2003; Cordeau et al. 2005). The aim is to find the optimal berth position for each vessel, i.e., the berth position that minimizes container handling cost from the vessel to location in the marshalling yard where outbound containers for the corresponding vessel are stacked. In Figure 2 there is an example of a berth schedule. We assume that the berth schedule is an exogenous input to the part of the simulation model discussed here. Actually, vessel entrance at port and berth schedule may be affected by the specific features and physical characteristics of the real port of interest, but, here we focus on common logistics connected with
other basic operations to be performed on berthed (standard) vessels by (standard) rail mounted gantry cranes. These operations are at the kernel of the simulator platform under development at CESIC-NEC Italy.

The arrival of the ship process proposed in (Legato and Mazza 2001) is the starting point for a new generalised model and is described in the following. Once the vessel has arrived outside the port (the arrival time is an uncertain time close to the vessel ETA), its mooring along the assigned berth position depends on the following requirements: i) formal conditions (e.g. contractual agreements between the vessel’s shipping line and the port of call for the use of port facilities), which means a priority policy for the port entrance queue; ii) operational settings (i.e. pilot mariner and pilot boat availability, berth space assignment and the least number of free quay cranes). If requirements are met, the ship is manoeuvred down the navigation channel and into its berth slots by one or two pilot boats; otherwise it must wait in the roadstead. In our model we have a finite vessel population of 97 ships of which 12 are mother vessels.

Once a vessel is at berth, container discharge/loading can be initiated only if mechanical (and human) resources are allocated; if not, the ship waits in its berth position until resource assignment. Discharge/loading operations are performed by RMGs placed along the berth: one or multiple cranes move containers between the ship and the quay area. The maximum number of quay cranes that is possible to assign to each vessel is restricted by i) the total number of cranes in the quay and ii) the maximum number of allowable cranes to each vessel due to physical (i.e. the length of the vessel) and logical constraints (i.e. interference between crane booms). Considering the span of the cranes (30 m) and the horizontal space necessary to stack and transfer away the incoming/outgoing containers of a vessel, the maximum number of cranes assignable to the longest vessel is 5 (this number is proportionally decreased for shortest vessel). The problem of assigning RMGs to the vessels for each time-slot is called “Quay Crane Deployment Problem” (QCDP) (Park and Kim 2003).

When multiple cranes are assigned to the same ship, crane interference has to be avoided and a complex scheduling problem arises to manage the relationships (precedence and mutual exclusion) existing among the holds of the same ship. This is the “Quay Crane Scheduling Problem” (QCSP) (Canonaco et al. 2008).

We manage both problems together in a two-phase approach: the first phase concerns the QCDP, while the second one involves the QCSP. Assuming that the berth schedule is deterministic for the model, in the first phase we use CPLEX to solve a mathematical model that is able to assign for every vessel at each one-hour time-slot a sub-set of RMGs (Legato et al. 2008a). Afterwards, in the second phase we use a metaheuristics to schedule the ship holds to the assigned RMGs (Legato et al. 2008b). Thus, the simulation model initialization is schematized in Figure 3.

The performance of the discharge/loading process highly depends on the availability of this type of cranes and their turnover speed. RMG performances are summarized in Table 1. Again, this problem has been deeply studied within the real context of Gioia Tauro Terminal (Canonaco et al. 2008), but without being integrated with the problem of simulating yard organisation, SC travel back and forth from the yard and analysing the effects of crane shifting along the berth, under the condition that multiple vessels are at berth.

<table>
<thead>
<tr>
<th>Performance</th>
<th>(m/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container hoisting speed rated with load/light-load</td>
<td>50/120</td>
</tr>
<tr>
<td>Trolley travel speed</td>
<td>210</td>
</tr>
<tr>
<td>Gantry travel speed</td>
<td>45</td>
</tr>
</tbody>
</table>

In a DTS container terminal the transport of containers from ship to yard (and vice versa) and the container stacking processes are performed by a fleet of straddle carriers (SCs). SCs take in charge containers and cycle between the berth area and the assigned storage positions within the yard. Straddle carriers are capable with a laden container of relatively low speeds (up to 20–26 km/h). The yard is a passive resource: it consists of a matrix of 2 lines and 16
columns, in which each matrix element is a three-dimensional matrix that stands for a yard sector. SCs are able to select a slot within the yard structure to load/unload or stack containers. They are also able to compute the distance from the assigned RMG to the yard slot and back.

Each RMG has four SCs assigned. We have not developed any optimization code to dynamically assign SCs to RMGs (and we only use 72 of the 75 available SCs).

In the present case, no considerable inter-modal TEU transportation will be considered.

Modelling approach

The model has been developed by using a discrete-event simulation architecture based on finite state automata. The approach used to model the whole container terminal is based on previous promising works (Pidd and Castro 1998; Legato and Trunfio 2007): it represents the interaction between different logistic processes through special model objects called resource managers. Resource managers are gifted of a high-level view of the whole system and are able to operate on the system resources. They use exogenous data (i.e. berth schedule, RMG deployment and scheduling) to dynamically assign the container terminal resources to the jobs.

A Queuing Network Model Description

The logistic processes described above have been depicted using queuing networks. The arrival of the ship process is described in Figure 4. The model is based on the hierarchical representation proposed by Legato and Mazza (2001). Incoming vessels wait in a priority queue that represents the roadstead; outgoing vessels wait in another priority queue that stands for served vessels waiting in the assigned berth slot. A hidden model object called “Berth Manager” assigns pilot boats to incoming and outgoing vessels according to previously defined conditions (i.e., active and passive resource availability).

A vessel entering the rightmost (black) box in the figure means that it has been berthed and another queuing network model takes charge of it.

![Figure 4: Arrival of the Ship Process](image)

Figure 4: Arrival of the Ship Process

Figure 5 shows the queue network model for both the i) loading/unloading of the ship process and ii) container transport from the ship to the assigned yard-slot (and vice versa) and the consequent storage process. The simulation model uses a list of berthed vessels. Each vessel is assigned to a berth slot (a discrete part of the quay assigned in the berth schedule). For each vessel in list, a model object called “Operations Manager” assigns cranes to the vessels and starts the operations on the vessels. When discharging and/or loading operations have been performed, the Operations Manager pushes the served vessel out of the model and updates the system performance measures (e.g., overall completion times).

In Figure 5, both the discharging and loading process of the ship are depicted. The Operations Manager tasks are: i) for each vessel in the berth, to assign the vessel holds to be served to each RMG; and ii) to route SCs within the network in order to execute the optimal handling operations.

COMPUTATIONAL FRAMEWORKS

As part of the industry, our belief is that large and complex simulation models of logistic systems are not only an interesting academic research field of interest, but that they can be applied to real systems to provide effective scenario analysis and decision support to operational management in a reasonable time. Here is the key point: detailed models of large and complex systems are affected by a high burden on execution,
but, on the other hand, we do not believe as a good idea, when modelling logistic systems, that of using simplified system description and interdependencies among subsystems to get a lower computational burden. Modern simulation methods and technologies, such as simulation-based optimization are applied more and more in the organisation and management of logistic systems. Frequently these methods require evaluating thousands of alternative system configurations with the goal of searching for better system policies and configurations. Hence, only computational frameworks based on high-performance computing will be able to be candidate as effective tools for the tactical and operative managers.

**Integration of Queuing-Network Models in a Parallel Computational Framework**

In this context, we aim to integrate the model previously described into a HPC computational framework devoted to the statistical analysis of the output data from simulation runs.

Actually, we have developed a simple test case based on a single server, queue-model with the aim of experimenting some different methods for confidence interval estimation upon average performance measures that can be produced from simulation output data. Using a PC, wall-clock time experienced by us to get a satisfactory simulation output analysis for a single configuration of the queuing-network model that we proposed is of the order of dozens of minutes.

The methods to be compared have been selected between the following approaches (Katzaros and Lazos 2001): the well known independent replications (IR) and batch means (BM) and a mixed recent approach known as replicated batch means (RBM) (Andradóttir and Argon 2001).

The IR approach is a pure parallel methodology: the independent replications refer to simulation runs of the same fixed length; each run starts from the same initial state, but uses a different seed in the pseudo random number generator. Despite the fact that each run starts from the same system state, there is no correlation between each replication: in fact, considering that for each run there is the presence of an initial transient that must be removed, using different seeds we guarantee that the remaining part of the run is really independent. Moreover, samples from the different replications are identically distributed. This approach is perfect whenever a parallel machine, with a set of $m$ CPUs large enough ($m=20,..,30$) to submit one run per CPU, is available. In this case, to obtain $m$ independent replications, we need the same number of CPUs and we commit $m$-times the effort to execute the initial transient.

The BM approach is a merely sequential method that cannot be used for variance estimation purposes in transient simulation. It is a really effective variance-estimation methodology for steady state simulation, in fact it uses only one long run, purged by the initial transient and then divided into equal batches. All the batches registered on the same run are used for variance estimation purposes. Using a sequential machine, the BM is better than the IR methodology, because it allows collecting $m$ independent replications with the effort of estimating the initial transient only once. The drawback of this method lies in the correlation between samples used for variance estimation. Nevertheless, when simulating complex systems (e.g., the queuing network that we proposed) too much computational effort is needed using the BM method, while the IR methodology with the use of parallel machine seems to better suit our needs and overcome the BM.

The RBM variance-estimation methodology combines good characteristics of IR and BM: as a matter of fact, the number of initial transients that are estimated is reduced by simulating a system on $m$ CPUs using the BM methodology. As for the IR, each CPU uses a different seed for the pseudo random number generation and simulates the system from the same initial state. Recently, it has been empirically proved that the RBM methodology is better than BM in terms of overall coverage of confidence intervals (Andradóttir and Argon 2001; Alexopoulos et al. 2006).

We believe that the RBM is the perfect trade-off between the IR and BM methodology and that it will be used as a basis for simulation output analysis in high-performance computing frameworks.

Here, we test the best methodology for simulation output analysis that we used for our complex queueing-network model, via a basic and repeatable test case consisting of a single queue model that may be considered as extracted from one of the queue network models presented in the previous section. By fixing the arrival rate, $\lambda = 9$, from a Poisson distribution and the service rate, $\mu = 10$, from an exponential distribution we may use the well known analytical (exact) result on the mean waiting time $\left(1/(\mu - \lambda)\right)$ to evaluate the coverage and the quality of the confidence intervals produced by different computational choices. In Tables 2, 3 and 4, we present numerical results on mean, variance and interval estimate of the waiting time, as obtained after having deleted 5.000 observations to avoid the bias effect of the initial transient behaviour. In particular, results in Table 4 are referred to a shorter batch length – i.e. 1.000 observations instead of 5.000 – because we were exploiting the effect of a quick stop of the run length by each processor. As one may see, numerical results are still quite goods whenever one decides to restrict the level of parallelism from 30 parallel processors, each performing 1 batch, to 5 parallel processors, each performing 6 batches. But we may not recommend a sort of saving on the computational power of each processor, by reducing the batch length: related interval estimates become unsatisfactory.
Another feature of the computational framework we are using is the master/slave approach proposed in (Legato et al. 2006). To ensure fault tolerant computing, we resort to the common practice of checkpoints, a policy to manage two different types of crashes: a crash of the master node, and a crash of a slave node. The idea is described in the following. In our master/slave approach, the master keeps track of the pool of independent replications dispatched to the slaves. Thus, whenever a slave crashes, the master can reassign the same configuration to another slave. On the contrary, whenever the master node crashes there is no way to recover the instantaneous state of the computational procedure. To react at the occurrence of a master crash, we have implemented a periodic checkpoint. In particular, at a sequence of suitable time points, we take a copy of (i) the model object state (e.g., resource state, queued jobs), (ii) the current seed value and (iii) the list of scheduled events; therefore, the list of the ‘‘latest’’ parallel replications that have been dispatched to the slaves is saved in a file system. In such a way the simulation can be restarted from the latest point of each replication. This consolidated technique is based on using the so-called inter-communicators. The framework uses machine built-in error handlers as well as other ones defined by us: both allow us to detect a slave failure due to a non-success return code from a communication operation (i.e. a send or receiver) on one of the inter-communicators.

The computational framework is based on a NEC TX-7 CC-NUMA parallel machine with 32 Intel 64-bit Itanium-II CPUs with 64Gb of RAM: the machine is located at the CESIC research centre (CESIC 2006). To overcome the problem of the pseudo random number generator to be used, we use the Mersenne Twister algorithm (Matsumoto and Nishimura 1998) as suggested by Katzaros and Lazos (op cit).
Through this framework, we will conduct practical investigation using “what if” scenario analysis or using simulation-based optimization to evaluate the impact of the possibility of purchasing new resources (RMGs and SCs) in the container terminal. Hence, considering that in a real organisational framework, this kind of application is expected to run on PCs, one may recognize that an HPC-based computational framework appears to be the unique cost effective tool which can provide proper solutions to decision problems in a reasonable amount of time for the terminal manager.

CONCLUSION

Current practice with the usage of simulation for the optimal management of logistic processes reveals that decisions are pursued by decomposing and separately modelling the whole process in a set of sub-process. In this paper, we report on the CESIC-NEC project for developing an overall simulation model based on previous experiences at the largest container terminal on the Mediterranean Sea. The simulation model has been depicted using queuing-networks presented here and model development has been done using finite state automata. The model uses mathematical models or heuristics approaches to evaluate such feasible solution for the berth planning, quay crane deployment and quay crane scheduling problems. Currently, we are involved in a simulation-based optimization model to optimize the core terminal operations. Finally, we are attempting to address what the implications of high-performance computing in simulation-based computational frameworks are for this industry. In particular, we stated that the RBM variance-estimation methodology for simulation output analysis is the best, using modern low cost multi-core machines for simulation output analysis.

REFERENCES


AUTHOR BIOGRAPHIES

PASQUALE LEGATO is an Assistant Professor of Operations Research at the Faculty of Engineering (University of Calabria), where he teaches courses on simulation for system performance evaluation. He has published on queuing network models for job shop and logistic systems, as well as on integer programming models. He has been involved in several national and international applied research projects and is serving as reviewer for some international journals. His current research activities are on the development and analysis of queuing network models for logistic systems, discrete-event simulation and the integration of simulation output analysis techniques with combinatorial optimization algorithms for real life applications in Transportation and Logistics. His home-page is http://www.deis.unical.it/legato.

DANIEL GULLÌ received a Laurea degree (cum laude) in Automotive Engineering from the University of Calabria, Rende, Italy, in 2005. He is devoted to research in numerical simulation at NEC Italy Center for High-Performance Computing and Computational Engineering (CESIC). His current research interests include discrete-event simulation models for logistic systems and parallel simulation.

ROBERTO TRUNFIO received a Laurea degree (cum laude) in Management Engineering from the University of Calabria, Rende, Italy, in 2005, and is currently pursuing the Ph.D. degree in Operations Research from the same university. Moreover, he is a logistics engineer at NEC Italy Center for High-Performance Computing and Computational Engineering (CESIC). His current research interests include modelling languages for discrete-event simulation, simulation-based optimisation of logistics systems, and parallel simulation.

RICCARDO SIMINO Graduated in Electronics Engineering at the Politecnico of Milano in 1984 presenting a research on 3D vision systems for industrial robots based on Artificial Intelligence Technologies. He achieved a master in Marketing at the University of Bocconi, Milan, Italy. After having worked for several years as employee for Digital Equipment Corporation and as consultant for International Government Agencies, from April 2001 he is director of the High Performance Computing Division of NEC Italy and from 2003 General Manager of CESIC.