Interaction delay in M/M/C/N and the impact of buffers on harbor quay-crane operations

Container terminals are essential intermodal interfaces in the global transportation network. Efficient container handling at terminals is important in reducing transportation costs and keeping shipping schedules (Zhang, 2003). Reducing the average processing time (APT) of loading and unloading operations is one of the most important objectives in a container terminal. One of the factors influencing total process time is the interaction between quay crane (QC) operation and the yard trucks (YTs) operation. The gap between the APT and its lower bound is a result of the interaction between the arrival process and the service process, due to the variance in both processes. Our interest is in the delay caused by the interaction, and therefore we define the *Relative Interaction Delay* (RID) as the difference between the APT per unit and its lower bound, divided by the lower bound. We quantify the benefit of the use of a buffer referring to those operations.

Buffers in the context of harbor operations can be implemented in several ways, and their effectiveness may be sensitive to the overall implementation scheme, as well as the additional time required for buffer operations. We analyze an implementation in which static driverless YTs are standing in front of the QC:

1) Unloading process: When the QC picks a container from a vessel and there is no available YT with a driver, the QC crane releases its load on an empty driverless truck and continues unloading the next container from the vessel. When a truck with a driver arrives at berth but the QC is not ready with another container, the driver parks the empty truck in a buffer area for later use and takes the full truck to the storage yard. This operation of driver changing trucks may require non-negligible additional time. The QC waits when all YTs are unavailable and all buffer trucks are full. Drivers wait when the QC is unavailable and all buffer trucks are empty.
2) Loading process: When YT with a driver arrives at berth and the QC is unavailable, the driver leaves the truck in the buffer area and takes an empty truck to the storage yard. As in (1), this operation of driver changing trucks may require additional time. When the QC is available for loading but no truck with a driver is there, the loading is carried out from the buffer. The QC waits when all YTs are unavailable and all buffer trucks are empty. Drivers wait when the QC is unavailable and all buffer trucks are full.

For a given number of servers (C), APT increases as a function of the supply-demand ratio (θ), with increasing slope. Two lower bounds can be identified. The first lower bound is dictated by the QC rate, as $\mathrm{APT} \geq \frac{1}{\lambda}$. The second lower bound is dictated by the YT rate μ, as $APT \geq \frac{1}{C \cdot \mu}$. When $\theta = \frac{\lambda}{C \cdot \mu} \ll 1$, performance is dominated by the QC rate. When $\theta \gg 1$, performance is dominated by the YT rate. In all cases the more limiting bound governs. The ideal case of zero additional time can be considered as an upper bound for the benefit from the buffer.

When buffer operations involve no further delay, the system can be represented by the analytic M/M/C/N queueing model. In the more general case, an event-based simulation model is used. The maximal difference in RID values between the simulation and the analytic models is 0.0436 (4.36%).

Several main patterns about RID behavior can be observed using the analytic model: The values of RID range from 0 to 0.5. For every combination of number of buffer spaces, Q=N-C and C, RID reaches its peak when θ is one, and approaches zero when θ gets away from one. For each C, interaction delay decreases when Q increases. For each Q, interaction delay decreases when C increases.

The most significant improvement in interaction delay as a function of available queue space is in the transition from no buffer to one buffer space. This improvement is particularly significant for small number of servers, and reaches its peak when C=1 and $\theta = 1$. Increasing available queue space by one in this case decreases interaction delay nominally by 17%, from 50% to 33.3%.

For one server, RID values are symmetrical on both sides of $\theta = 1$. As C increases this symmetry no longer holds, and values of RID are closer to zero for θ smaller than one compared with their reciprocals. For example, $RID\left(\theta = \frac{1}{2}, C = 6, Q = 0\right) = 2.6\%$ while $RID(\theta = 2, C = 6, Q = 0) = 5.6\%$.

In practice buffers are not likely to be ideal, and their operation may involve some slowdown. If the buffer slowdown (τ) is minor (up to 10%), results remain rather similar to the ideal case. Different values of τ produce markedly different results. When τ is small (0.1) or zero, RID decreases as the buffer size increase. On the other hand, when τ is more substantial, the pattern is much more complex. For example, with τ =100% (in the case of

C=1), increasing buffer size is counterproductive whenever $\theta > 0.75$. Even if τ is only 50%, increasing buffer size may also be counterproductive, but only if $\theta > 1.5$. More generally, increasing Q is helpful only until a specific value of θ, above which a reverse impact occurs, as the harm of buffer slowdown becomes higher than the benefit of the buffer. The specific θ's threshold value decreases as C increases.

A possible explanation for the impact of the combinations of θ and τ on RID values, which is well supported by our results, is that the YTs work time increases when θ increases. For large values of θ, the YTs have rarely idle time, and any additional effort they make immediately affects the overall system performance. With smaller values of θ the additional service time of the drivers influences drivers' idle time, and thus has smaller impact on overall performance.