

WHITEPAPER

"SOR enhancements for cash treasuries"

JIUYUE ZHOU
ANDREW WESTHEAD
JANUARY 14, 2026

SUMMARY

In this paper, we propose a new model for passive order placement in smart order routing (SOR) for cash treasuries, designed to improve the estimation of order fill times and thereby enable more effective size allocation across venues by minimizing time to fill. While the previous model relied mainly on trade rates, the new model also incorporates cancellation rates to more accurately capture queue dynamics. By accounting for order cancellations ahead of the passive order in the queue, the new model provides a more realistic representation of execution processes. Empirical evaluation demonstrates a substantial improvement in time to fill prediction, with a 37% reduction in mean absolute error. These results highlight the effectiveness of our approach in enhancing execution efficiency in multi-venue trading environments.

PASSIVE PLACEMENT

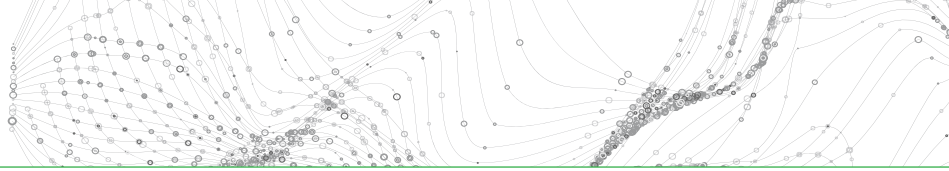
Upon receiving a passive order, the smart order router determines the allocation of order size across multiple venues by minimizing the overall expected time to fill. In the previous model, the estimated waiting time was based primarily on the trade rate, as trade activity is a critical driver of order execution. However, cancellations of orders ahead in the queue can also contribute significantly to faster fills. To capture this effect, our new model incorporates the cancellation rate alongside the trade rate. And it avoids the heavy computational demands of machine learning models and neural networks [XC18], thereby preserving computational efficiency.

Suppose we need to allocate N contracts to k venues. Denote the number of contracts allocated to venue i as n_i where $i = 1, 2, \dots, k$, $n_i \geq 0$ and $\sum_{i=1}^k n_i = N$. The waiting time to get n_i contracts fully filled on venue i is denoted as t_i . Then the total waiting time to get all of our orders fully filled is $T = \max_{1 \leq i \leq k} t_i$. Our goal of minimizing the total waiting time to get fully filled can be transformed into the optimization problem below:

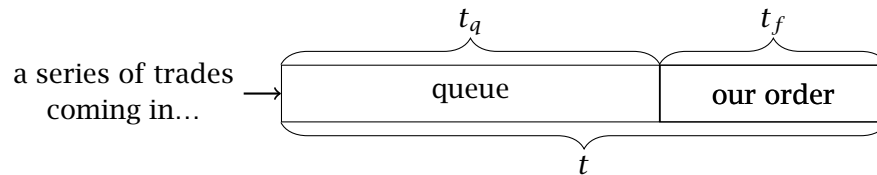
$$\min_{n_i} T = \min_{n_i} \max_{1 \leq i \leq k} t_i(n_i)$$

where

$$\sum_{i=1}^k n_i = N, n_i \geq 0$$



PAGE 2 On each venue, after placing our passive order at the inside quote, the total waiting time for our order to be fully filled t consists of two components: the time to reach the front of the queue (t_q) and the time to fully fill the allocated contracts t_f , as shown in the following diagram (for notational simplicity, we omit the venue subscript, as the following analysis applies identically to all venues):



As time progresses, the queue length ahead of us, denoted by Q , diminishes due to both trade and cancellation events. Assuming the cancellation of each order is independent, the canceled quantity in a given time Δt is $\lambda_c \frac{Q}{Q_0} \Delta t$ where λ_c is the real-time cancellation rate, Q_0 is the initial queue length when we place our order, and the ratio $\frac{Q}{Q_0}$ represents the percentage of the queue in front of us. Also assuming trade event happens at a constant rate λ_t on the near side, then the traded quantity in the given time Δt is $\lambda_t \Delta t$. Therefore, the change of queue length ΔQ follows the dynamics below:

$$\Delta Q = -\lambda_t \Delta t - \lambda_c \frac{Q}{Q_0} \Delta t$$

As Δt approaches 0, we have an ordinary differential equation of Q :

$$dQ = -\lambda_t dt - \lambda_c \frac{Q}{Q_0} dt$$

With the initial condition $Q(0) = Q_0$, Q has the following closed-form solution:

$$Q = Q_0 \left[\left(1 + \frac{\lambda_t}{\lambda_c} \right) e^{-\frac{\lambda_c t}{Q_0}} - \frac{\lambda_t}{\lambda_c} \right]$$

WANT THE FULL INSIGHT?

To access the full document, or request more, please get in touch

sales@quantitativebrokers.com