BASIS TRADING: MAXIMIZING OUR HIT RATES ACROSS U.S. TREASURIES CASH AND FUTURES

PURU P. SARATHY
APRIL 10, 2025

EXECUTIVE SUMMARY

This whitepaper examines the optimal wait time for executing basis trades in U.S. treasuries cash and futures markets. Due to network latency (~8ms) between New York (cash) and Chicago (futures), acting too quickly on an opportunity can reduce trade success.

We analyze two key metrics:

Opportunity Rate – The proportion of target opportunities that remain available after waiting a specified time period from when the opportunity first arose.

Hit Rate – The probability of executing at or better than the target price.

Using historical data, we find optimal wait times for different product pairs, finding 20–25 milliseconds as ideal.

INTRODUCTION

At QB, we regularly execute basis orders on behalf of our clients. The most common instance of this involves buying (or selling) a U.S. Treasury bond while simultaneously selling (or buying) a U.S. Treasury future with a similar duration.

For example, a typical basis trade might involve purchasing the seven-year cash treasury (CT7) on a cash venue while simultaneously selling the 10-year treasury future (ZN) on the CME. Each basis trade is defined by a set of parameters that determine its execution details. One of the most critical parameters is the target price for the basket, essentially, the price level at which the trade becomes favorable for execution.

QB continuously monitors the price of the basket, and once it reaches the target, our algorithm places orders in both the cash and futures markets simultaneously. However, while these markets are economically intertwined, they are physically located in different cities: cash treasuries trade in New York (NY), while futures trade in Chicago (CHI). Given the network transit time between these locations (~ 8 milliseconds), an immediate target price availability could be a result of network latency, either of the quotes could be stale depending upon whether the opportunity was being evaluated in NY or CHI. Acting on such an opportunity immediately may lead to suboptimal performance as the indicated prices may not be available when our orders get to the exchange.

This raises a key question: Once our client's target price becomes marketable, how long should the algorithm wait before sending out orders to the cash venue and the CME to maximize the likelihood of executing the order at the target price? This white paper outlines our approach to addressing this challenge.

OPPORTUNITY RATE AND HIT RATE

We define an opportunity as the event where the basket price drops below the target price (considering only buy trades, with the sell side being symmetric). The duration of an opportunity, denoted as D, is a random variable representing how long this condition persists. Let T be the time that the algorithm waits before sending the orders to exchanges.

The opportunity rate is the number of opportunities per unit time that remain available to the algorithm, given that it waits T milliseconds after detecting an opportunity. Naturally, as T approaches infinity, the opportunity rate drops to zero since no orders are executed. In contrast, at T = 0, the opportunity rate is 1, as all detected opportunities are immediately taken advantage of.

This leads to the concept of the hit rate, a key factor in defining this problem. Consider the case where the wait time is zero, and the opportunity rate is 1. Given the latency discussed previously, it is evident that a significant portion of trades will fail to execute at the desired price where the basket price remains below the target price. This measurement of the proportion of successful trades is called the hit rate.

We present the model that we use to trade-off opportunity rate with the hit rate to arrive at optimal wait times.

MODEL AND SOLUTION

The model's central concept lies in balancing the hit rate with the opportunity rate. Misses are expensive, as they result in order execution at less favorable prices. The goal is to minimize these misses while maximizing the number of actionable opportunities. However, prolonged waiting can reduce opportunities, as other participants might act on them sooner. When waiting surpasses a certain threshold, eliminating latency-driven opportunities, the improvement in hit rate eventually diminishes. This leads to a "sweet spot," beyond which further waiting offers no additional benefit.

Next we formalize the ideas stated above into a model:

- F is the cumulative distribution function of the duration of opportunity
- T is time the algorithm waits after detecting the opportunity
- L is the latency between the cash and futures venue

The Opportunity Rate (O) is obtained by measuring the probability that the opportunity is longer than the wait time, T:

$$O = 1 - F(T)$$

The Hit Rate (H) is calculated by considering the probability that an opportunity lasts more than $T + L$ milliseconds conditional on it lasting T milliseconds, where L as mentioned above is the latency between the cash and futures market venues.

$$H = Pr(t \geq T + L | t \geq T)$$

$$H = \frac{Pr(t \geq T + L, t \geq T)}{Pr(t \geq T)}$$

$$H = \frac{Pr(t \geq T + L)}{Pr(t \geq T)}$$

$$H = \frac{1 - F(T + L)}{1 - F(T)}$$

The objective function we use combines the opportunity rate and the hit rate in a linear fashion:

$$T^* = \underset{T}{argmax} \left[(1 - F(T)) + \lambda \frac{1 - F(T + L)}{1 - F(T)} \right] \quad (1)$$

We intuitively believe that the hit rate holds greater significance than the opportunity rate, as we are willing to sacrifice a small portion of the opportunity rate to achieve an improvement in the hit rate. Therefore, we adjust λ within the range of 1.1 to 1.5 to define the objective function. Using our internal historical data, we estimate F for each product pair. Once F is determined, we perform a numerical search to solve equation (1).

RESULTS

The numerical search procedure was applied to data for the CT2-ZT pair across CME and cash venues. The details of the orders and the target-price came from client orders that QB received in 2024. The distribution of the duration of the opportunity, the relationship between opportunity rate and wait time, and hit rate and wait time are shown in Figure 1.

Figure 1 (middle chart) illustrates how the number of opportunities decreases as the wait time increases. Notably, the opportunities to the left of the vertical line (at 8 ms) are challenging to act upon due to latency effects. Figure 1 (bottom chart) highlights the probability ratio that forms the hit rate, which initially increases with increasing wait time but eventually levels off.

Figure 2 demonstrates the changes in the objective function as the wait time increases. The key takeaway is that waiting beyond the latency threshold, approximately 8 ms, improves the hit rate while sacrificing some opportunities. However, this trade-off only holds to a certain extent, beyond which additional waiting yields diminishing returns. The optimal wait time is expected to lie within this plateau region of the objective function.

Using the estimate of the maximum value of the objective function with a λ of 1.5, we arrive at an optimal wait time of 23 milliseconds (solid vertical line). The results are robust to different values of λ.

It's worth mentioning that Figure 1 provides estimates of the opportunity rate and hit rate within a range of 0 to 50 milliseconds, as the optimal wait time is unlikely to exceed this upper limit.

FIGURE 1
Top: Histogram for Duration of Opportunity, Middle: Opportunity Rate, Bottom: Hit Rate

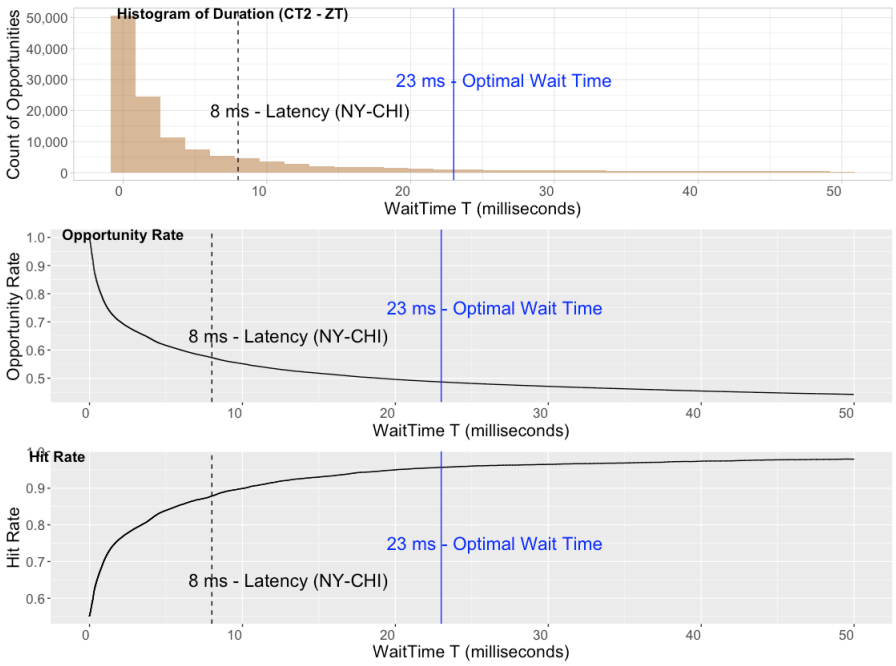
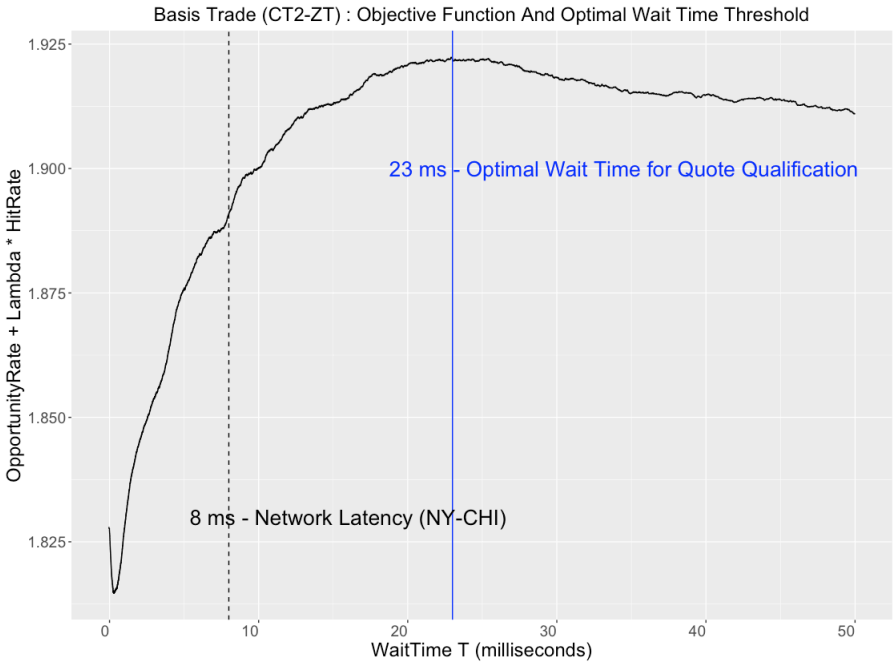


FIGURE 2
A numerical search yields 23 milliseconds as the optimal wait time for the CT2-ZT basis trade



A similar procedure applied to all pairs yielded optimal wait times for all products, as shown in Table 1. QB uses 22 milliseconds in production for all product pairs.

TABLE 1
Optimal Wait Times for Various Product Pairs

Cash Product	Futures Product	Optimal Wait Time (milliseconds)
CT2	ZT	23
CT5	ZF	25
CT7	ZN	25
CT10	TN	24
CT20	ZB	22
CT30	UB	22