Abstract—Electronic trading venues (ETVs) must simultaneously exhibit high degrees of both fairness and performance, yet in diverse contexts it is recognized trade-offs often exists between these two things. Introduced in this paper is a metric that captures the minimum extent to which an ETV’s performance must be inhibited by buffering so as to rectify the temporal unfairness it would otherwise exhibit. In light of ‘tail latency’, and so as to minimize the value it takes, a refinement of the metric is further described using a finding from queue theory pertaining to skips and slips. The metric and its associated buffering mechanism have recently been put into actual use on a major ETV: Thomson Reuters Matching.

I. INTRODUCTION

Worldwide, electronic trading venues (ETVs) facilitate the exchange of literally trillions of dollars’ worth of financial instruments each day [1]. They may be classified as distributed event-based systems because responsive to order messages sent to an ETV by market participants seeking to buy or sell the financial instruments traded on it, the state of data structures that organize those orders on the ETV—the most widely-used being the limit order book [2]—are updated, and in turn, participants on the venue are sent market data updates which are a contemporaneous ‘snapshot’ of the new state of each such data structure. In what might be characterized as an ‘infinite feedback loop’, these market data updates, in turn, cause participants to send more orders to the ETV, oftentimes to revise the prices of their open orders on it, to levels that are consistent with the prevailing demand and supply conveyed in the market data update.

This ‘infinite feedback loop’ in combination with the rise of algorithmic trading—where it is computer implemented programs, and not human beings, that are responding to market data updates with order messages—has led to massive increases in requirements for the number of events ETVs should be capable of processing. To give an indication of the scale of event processing in real ETVs, if one counts an order message sent by a market participant as a single event, then in a ‘fast market’ it is conceivable a single ETV could be required to process 18,000 such events per instrument per second [3, ch.7]. Depending on asset class (e.g., equities, spot FX, futures, etc.) an ETV can host hundreds if not thousands of instruments, so for a venue as a whole, this rate is likely several orders of magnitude larger.

Given the above it is perhaps unsurprising to learn that considerable effort has been expended by operators of ETVs in understanding, measuring and improving their performance [3, ch.7] [4]–[6]. An area that has received less attention, however, is the relationship between an ETV’s performance and what the author has previously termed its temporal fairness1 [1]. In diverse contexts it has been noted that there is a trade-off between efficiency (which, of course, is related to performance when the resource in contention is ‘time’ e.g., number of events per second) and fairness [7]. It is certainly possible for a high performance ETV to exhibit temporal unfairness, and vice versa, but the question addressed in this paper is can one characterize an ETV’s performance from the perspective of its (temporal) fairness? Both are important because to attract and retain the business of participants an ETV must exhibit adequate performance in its event processing (otherwise participants will send their orders elsewhere), and simultaneously it must also meet its obligations to provide a fair market (otherwise it may face sanction from regulators) [1].

The remainder of this paper is organized as follows. In Section II the general problem being solved is described and initial, simple metric to capture it is proposed. In Section III ‘tail latency’ [8] is identified as a reason to improve upon the initial, simple metric, and a finding from queue theory on ‘skips’ and ‘slips’ [9] is applied in the derivation of an improved metric. Section IV concludes this work.

II. BACKGROUND

Quite a large number of metrics relating (exclusively) to the performance of ETVs have appeared in the literature (see e.g., [3, ch.7] [4]). The goal here, however, is not to review these performance-only metrics, but rather to view performance through the lens of temporal fairness. Most ETVs operate—in large part for reasons of performance—what are known as continuous markets, i.e., they process order messages as quickly as possible upon their receipt, without the deliberate imposition of delays on those messages by the ETV as would occur in what are known as batch markets [10]. What continuous markets have led to—some would argue disadvantageously [11], [12]—is a never-ending technology ‘arms race’ where each participant is incentivized to be faster than their peers.

1Temporal unfairness as the term is used here roughly corresponds to what economists more generally term information asymmetry.
To draw an analogy, an ETV implementing a continuous market is much like a race track. Intuitively, a race on the track is ‘fair’ if no participant in it receives a ‘head start’, or gets to run a shorter distance. Further, under two these conditions, the fastest participant in such a race will win it. Absent one (or both) of these two conditions, the race will not be ‘fair’ because the fastest participant may not win it. To tie the analogy back to properties of an ETV: a ‘head start’ corresponds to a market participant being sent a market data update earlier than another; a ‘shorter distance’ corresponds to a market participant’s order being subject to a shorter processing time by the ETV than another’s. ‘Winning’ the race—and under certain circumstances placing earlier in it—corresponds to allocation of a scarce resource on the ETV to the participant. Examples of such resources include favorable queue position in the instrument’s limit order book when participant compete to make a price, or being able to buy or sell the instrument at a favorable price when participants compete to take a price.

Perhaps unfortunately from the perspective of fairness, and for the reasons discussed in more detail in the author’s recent work, it is likely impossible to ensure that any single ‘race’ on an ETV is ‘fair’. Due to the serialization inherent in network communication and in processing messages against a limit order book, in any single race, some participants will receive head starts over others, and similarly some will have orders subject to shorter processing delays than others. It seems to follow from this that a metric relating to both the performance and fairness of an ETV could be derived by examining, for each market data update (i.e., ‘head starts’ in the analogy), the amount of time that elapses between the first participant being sent that update, and the last being sent it. A different form of the same metric can be then also be derived by examining differences in order processing times (i.e., ‘track distance’ in the analogy) when those orders are competing for the same resource. Let $T_{mktd}$ and $T_{ords}$ denote these two forms of the metric, respectively.

On the extent to which the metric captures performance characteristics of the ETV consider, ceteris paribus, increasing the bandwidth of the network connection over which market data updates are sent. What this should result in, if the bandwidth is the limiting factor in sending messages, is a reduction in the elapsed time between sending the first message in an update and the last. In many instances then the $T_{mktd}$ value should decrease as a result of this performance-related upgrade to the ETV. A similar argument can be made for $T_{ords}$ if, ceteris paribus, the ETV’s hardware were upgraded—order processing times should decrease and therefore differences in those times should also decrease. It does seem to follow then that smaller values in the two forms of this metric indicate a higher degree of performance of the ETV.

On the extent to which the metric captures the (temporal) fairness of the ETV, the less elapsed time between the sending of the first and last message in a market data update, the smaller the ‘head start’. Similarly, the smaller the differences in order processing times, the smaller the differences in ‘track distance’. Thus, smaller values may indicate a higher degree of temporal fairness because although ‘head starts’ and ‘shorter distances’ might inherently be unfair, it is their magnitude that affects race outcomes. Put another way, the fastest participant in a race may still win if the magnitude to which others receive head starts and run shorter distances is less than the winning margin (in units time) that participant would exhibited had the race been fair.

If one accepts the arguments set forth above for why the metric reflects aspects of both performance and fairness on a venue, it may seem that the goal of this paper has effectively been achieved. Trivially, over some chosen time horizon, one could find the maximum values of $T_{mktd}$ and $T_{ords}$ on the venue, and sum the two values to yield a value for this ‘fairness-oriented performance metric’ for the venue. The metric could be collected again over other time period, after say the venue was subjected to a change (e.g., adding new market participants or instruments, upgrading its hardware or modifying its software implementation, etc), to determine whether that change resulted in an increase or decrease in temporal fairness on the venue.

Finally, and perhaps most importantly, the value taken by the metric when these two maximum values are summed can be used to inform the length of a buffer that may be retrofit to an ETV to rectify the temporal unfairness it would otherwise exhibit absent that buffer. Buffering, of course, is a general concept that has existing in computer science for at least 60 years, and many specific buffering mechanisms have recently been proposed to this end (even if their authors do not explicitly identify them as such). While it has been shown that the actual length of the delay a buffer must impose to ensure or approach temporal fairness depends on its specific implementation, on a single race basis this metric captures the lower bound of a delay any buffer would have to impose.

To summarize: operators of ETVs want to simultaneously ensure their venues have high performance and exhibit temporal fairness. The metric described in this section may help them to do this by quantifying the deliberate degradation in performance that would result on it if, to rectify temporal unfairness, a buffer were deployed. The goal of venue operators should thus be to design and operate their ETVs to minimize the value of this metric.

---

3A rough proof for this is as follows. If two participants have the same exact response time (or in the analogy, the same exact speed), and one receives a head start of $H$ units time and gets to run a distance $D$ units time shorter, then the participants orders will be received $H+D$ units of time apart. But both orders have to be in the buffer simultaneously for the buffer’s implementation to reorder them as output. For this to happen, the buffer must thus delay the first order by at least $H+D$, at which time the second order will arrive as input to the buffer.

4Examples of such design features that minimize the metric’s value appear in [15], [16].
III. IMPROVING THE METRIC

It has been said that besides finding new ways to measure things in the real-world, scientists should also seek to improve upon measures that already exist [17, ch.1]. This naturally seems to raise the question of ‘how’—and perhaps more fundamentally of ‘why’—the metric described in the previous section should be subject to improvement? As for the ‘why’, like most distributed systems [8], ETVs oftentimes exhibit long or fat tails in their times taken to process requests [3, ch.7]. What this means, per the description of the metric in the previous section, is that it will take the value of the most extreme difference in market data update transmission time, and the most extreme difference in order processing time by the venue. The author’s own experience has been that these most extreme differences (or ‘tail values’) can be two orders of magnitude larger than the ‘normal’ values on an ETV. Further, what might seem the ‘easy fix’ for this—particularly, filtering out the ‘large’ values—is fraught with subjectivity, and will effectively ignore whatever unfairness may manifest in the tail. It does seem then that given a venue operator’s desire to minimize the value taken by such a fairness-oriented measurement of performance that a well-thought-out improvement on the previous section’s metric is warranted.

As for the ‘how’ of this improvement: a finding from the field of queue theory is that the benefit of receiving a ‘skip’ (i.e., overtaking someone) might be statistically offset by the penalty of later receiving a ‘slip’ (i.e., being overtaken by someone) when a customer is repeatedly accessing a resource by queueing for it [9]. If one accepts that despite a venue operator’s desire to minimize the value taken by such a fairness-oriented measurement of performance that a well-thought-out improvement on the previous section’s metric is warranted.

In light these three things: (1) serialization, (2) repeatedness, and (3) the finding from queue theory, it seems now that the problem could be framed in terms of multiple races instead of in terms of single races as was its treatment in the previous section. In particular, in the previous section, the formulation of the problem was such that the buffer was ‘sized’ (by way of the two forms of the metric) to nullify the effect of a ‘head start’ or ‘shorter distance’ in any race, the implication of this being that the metric would take the most extreme values observed for a ‘head start’ and for difference in track distance. The formulation of the problem in this improved approach involves accepting that ‘head starts’ and ‘shorter distances’ are inevitable in individual races, and crucially tolerating them to the extent that no participant is disadvantaged over any other participant by habitually receiving ‘head starts’ over the other, or habitually running shorter distances than the other.

In this improved approach then, the buffer is ‘sized’ not so as to eliminate ‘head starts’ and shorter distances, but rather to equalize the number of ‘head starts’ a participant gets on another when they mutually compete for a resource type, and similarly to equalize the number of times one gets to run a shorter distance than the other. It is the above insight on equalizing ‘head starts’ and shorter distances that led to the following stochastic definition of temporal fairness provided the author’s recent work [1]. Specifically, an ETV may be considered temporally fair if for all pairs of participants, and for each instrument traded on the ETV, the following two conditions, C1 and C2, are met:

C1 The number of times a participant P is sent market data updates before P′ is approximately equal to the number of times P′ is sent them before P.

C2 When P and P′ are competing for a certain type of resource by way of sending orders, the number of times the orders of P overtake those of P′ is approximately equal to the number of times those of P′ overtake those of P.

The definition refers to all pairs of participants because this is implied by the requirement that no participant be advantaged any other while simultaneously recognizing ordinality generally matters when participants race for prices on an ETV.5 The definition refers to ‘certain types of resources’ because the benefit received from receiving a ‘skip’ on one resource type (e.g., joining the top of the book as price-maker) is unlikely to be statistically offset by the penalty of receiving a ‘slip’ on another distinct type of resource (e.g., price-taking from the limit order book, or joining ‘deep’ in the limit order book). For the same reason, the definition refers to individual instruments because the benefit of receiving a ‘skip’ for a given resource type on one instrument (e.g., eur/usd, which is very liquid and heavily traded) is unlikely to be statistically offset by the penalty of a slip for that same resource type on a different instrument (e.g., on usd/zar, which is much less liquid). The definition specifies approximate (and not exact) equality because of non-determinism that manifests in an ETV’s implementation [1]. Put another way, exact equality is not specified in the definition for the exact same reasons that flipping a fair coin N times will not necessarily result in exactly N/2 heads (or tails).

To complete the description of this improved metric then6:

5It has been said that continuous markets are “winner-takes-all” markets. This, however, is not strictly true. In price-taking, if three participants race for a low-priced offer of size 2 units, and the fastest participant’s buy order is of size 1 unit, then the second fastest participant’s buy order will also be matched for 1 unit. An even stronger argument can be made for the decreasing desirability of being first, then second, then third in queue at a price level in the limit order book when price-making.

6For reasons of brevity and clarity certain details of the metric have been omitted here, but may be found in [18].
two minimum buffer lengths are once again found, such that conditions C1 and C2 from above may be met. In particular, if two messages in the same market data update were sent more than $T_{mktd}$ units of time apart, then any buffer having length less than $T_{mktd}$ will not be able to reorder them, and it will be as if the message sent earlier skipped that sent later. The goal is to find the minimum value for $T_{mktd}$ that approximately equalizes the number skips each participant has over the other, on a per instrument basis. To actually implement this notion of approximate equality, one can use a statistical test ‘in reverse’—e.g., if the significance level is chosen to be 0.05 then the minimum difference between the expected values and actual values is what is targeted for this notion of ‘approximate equality’. For instance, among 100 updates sent greater than $T_{mktd}$ units of time apart to two participants, to not have significance at the 0.05 level using the chi-squared test, ‘approximate equality’ would constitute 41 and 59 or closer since $p=0.07186$ (and since at 40 and 60 $p=0.0455$). A similar procedure can be applied to find a minimum value for $T_{ord}$ with the caveat is it applied per resource type per instrument and that an order overtakes another if it was received at the venue’s ‘edge’ later than another, but was processed for matching against the limit order book earlier than the other.

Ongoing collection of this improved metric and its associated buffer [2] has been implemented on a major spot FX trading venue: Thomson Reuters Matching (TRM). For illustrative purposes, some minimum values for the $T_{mktd}$ form of the metric calculated on an hourly basis over the course of a single trading week are shown for three instruments (red, blue and green) in Fig.1. Each instrument on the graph has strong seasonality (of period one day) and advantageously appears ‘stable’ day-to-day in the absence of structural changes to the venue (e.g., hardware and software upgrades, etc). Larger values in the metric’s time-series correspond to times where the market for each instrument is busier.

IV. CONCLUSIONS

This paper has introduced and described the refinement of a fairness-oriented performance metric for use on ETVs. The metric seeks to formally capture a trade-off between a specific form of performance relating to the deliberate imposition of delays in processing order messages and form of fairness relating to what economists term information asymmetry. The novelty of the refinement is in its use of ‘skips’ and ‘slips’ from the field of queue theory to minimize the metric’s value—and therefore detriment on performance if it is used to inform the length of a buffer on the ETV—without controversially having to exclude from its computation extreme values that manifest in ‘tail latency’. A further novelty of the metric, especially when it is compared to other widely-used fairness metrics such as that of Jain et al. [19], is that it is constructive—not only does its value indicate the degree of temporal unfairness exhibited by ETV, but its value can also be used directly to ‘fix’ the temporal unfairness it quantifies by informing the length of a buffer that may be retrofit to the ETV

Fig. 1. One week of hourly minimum $T_{mktd}$ values for three instruments on TRM. (Scale on y-axis is redacted; each tick on x-axis is one hour.)

[2] Finally, the metric provides a very different perspective on empirical buffer sizing to that which appears in the recent economics literature (see e.g., [20]–[23]).

REFERENCES