Communication between Brokers/Proprietary Traders and Trading Venues

There is a lot of information that flows back and forth between the trading venues and their members (the brokers and proprietary traders who are trading on their platform). There is the flow of orders from members into the venue, but there is also a lot of information that must travel the opposite way. The members want to receive timely information from the venue about the status of their own orders and trades. Members also want timely information on all quotes and trades happening on the venue, so they can base their own decisions on an up-to-date view of the market.

At a technical level, information passing from one party to another involves several different layers of physical and logical infrastructure. There is the physical equipment that carries data packets between machines, a protocol specification for the format and meaning of each individual packet, and a high level specification of what data will be communicated, when it will be communicated, and how it is organized. Below we'll give a brief overview of these components for the communications between trading venues and their members. It is also worth noting that these components in aggregate account for the latency of information between brokers and trading venues and between proprietary traders and trading venues. Though these brokers and proprietary traders often use similar means of communicating with venues, their actual needs in terms of latency might vary considerably. We discuss this further in our blog.

Microwave/laser networks

The fastest way to send packets of information across miles of geographic distance is via microwave towers. As a result, market participants seeking the lowest latency in their connections to and from exchanges use microwave as their primary communication medium.

However, microwave can be unreliable (particularly in bad weather), so these participants typically also have fiber optic connections to fall back on.

Extranet providers/dark fiber

Participants who are slightly less latency senstive might use fiber optic connections as their primary way of connecting to trading venues. There are many "extranet" providers who are in the business of leasing their extablished fiber optic connections to market participants.

Data centers

Whether they are traveling via microwave or fiber, data packets orginate and terminate at machines that are housed in data centers. The time it takes a data packet to be formulated at its origin, travel from its origin to its destination, and be processed at its destination is referred to as "latency," and it depends upon the medium of transport (microwave vs. fiber), the distance traversed, any intermediary routing technology (switches etc.), congestion, and the hardware and software that handles the low-level details of packet processing.

The spaces in data centers nearest to where the machines belonging to a NYSE, Nasdaq, or CBOE family stock exchange reside are controlled by those same companies, and they allow market participants to purchase "colocation" - the right to place their machines very close to the exchange's machines. This gives participants an (expensive) option to minimize the distance that data packets must travel between their machines and the exchange's machines, thereby shortening the latency.

FIX

FIX stands for the Financial Information Exchange protocol. It was invented in 1992 as a more reliable and unambiguous replacement for communication between investors and brokers that previously had taken place over the phone. FIX defines a format for each data packet that allows the recipient to interpret its contents, and is tailored to the kind of information that is relevant to financial transactions. At this point, FIX is a common organization of communication between investors and brokers as well as between stock exchanges and their members. Despite its ubiquity and sufficient expressivity, however, FIX

is not a universal choice among financial participants today because it is not particularly optimized for minimal latency.

Proprietary protocols

Most stock exchanges offer their members multiple choices of protocols that control packet formatting. FIX is one choice, but typically the choice that is closest to the native implementation of the exchange software will produce the shortest processing times at the exchange, and hence the lowest latencies. Each exchange family offers the choice of access in a proprietary format that is designed to be highly compatiable with its internal implementation, and it is not surprising that use of these proprietary protocols tends to minimize the packet processing times, and hence the overall latency. As a side effect, this means exchange members have to customize their communication software to the various exchange families, and typically build translation modules to go between the various proprietary formats and FIX or any other formats they use internally.

Orders and Market Data Content

Now that we've touched upon the infrastructure that physically transports packets, and the protocols that are used to format and process them, we'll zoom out a level and talk about what pieces of information go into data packets and how these pieces of information are commonly structured.

Order Types

Brokers communicate their desires to buy/sell stocks to a venue through a menu of order types and associated parameters that the venue defines. A would-be buyer or seller of stock is typically torn between two goals: getting a favorable price, and getting the trade done quickly. Different order types allow participants to express different trade-offs and constraints relating to these competing goals.

There are a lot of important differences in the menus provided by different venues, but most are variations on a few common themes:

Market Orders

A market order is used to communicate an immediate desire to buy or sell, irrespective of price. When a market order to buy enters a venue, it can be immediately matched with any open orders to sell. Conversely, when a market order to sell enters a venue, it can be immediately matched with any open orders to buy. This order type represents an extremal point in the trade-off between price and time: full priority is given to executing immediately, completely insenstive to price.

Limit Orders

Limit orders are used to express a constraint on price. The presumed goal is to execute a trade as soon as possible, within the band of acceptable prices. A limit order to buy, for example, will specify a ceiling for price, and will trade at the first opportunity to buy at or below its limit price. A limit order to sell will specify a floor for price, and will trade at the first opportunity to sell at or above its limit price.

The combination of limit orders and market orders is enough to build up some illustrative examples of trading dynamics. A market maker might have active limit orders to buy 200 shares of a stock for $10.00 a share and to sell 200 shares of the same stock for $10.02 a share. While waiting and available to be filled, these limit orders are referred to as "resting" on the order book kept by the trading venue. A buyer who wants to immediately buy 200 shares might enter a market order to buy, which will be matched against the limit order to sell at a price of $10.02.

Midpoint Orders

A midpoint order is a way of delegating the determination of price to the broader market. Instead of declaring a specific limit for price, a midpoint order will execute as soon as possible at a price that is equal to or more favorable than the midpoint of the current spread. In other words, it behaves like a limit order, but where the limit is a dynamically adjusting price calculated by taking the midpoint between the highest open buy limit orders and the lowest open sell limit orders. This calculation is typically over the price limits being advertised across all exchanges (which does involve some latency as the relevant information travels from the origin exchange to the venue processing the midpoint order).

If we return to our toy example of a single market maker offering to buy 200 shares for $10.00 a share or to sell 200 shares for $10.02 dollars a share, and we assume there are no better prices available across all exchanges, the midpoint price is therefore $10.01. If a midpoint order to buy enters a venue under these circumstances, it will be willing to buy at $10.01 or lower, and therefore will not be matched against the limit order to sell at $10.02. It will wait to interact with a seller willing to sell at $10.01 or lower (which might be a market order to sell, a midpoint order to sell, etc.)

Order Attributes

Many order types allow those submitting orders to specify additional parameters that control more of the fine-grained mechanics of how the orders behave. Limit orders can be "lit" (aka "displayed"), meaning that their sizes and limit prices are visible to other market participants who might want to take the other side of the trade, or "dark," which means they are not visible. Market makers typically use lit orders to advertise and generate trades.

Dark orders can often specify a minimum quantity, meaning that they will not execute for less than that specified number of shares. For example, an order for 1000 shares that has a minimum quantity of 500 shares can not be matched in a trade for 100 shares. It will wait until at least 500 shares can be traded at once.

Some orders can also have pushy designations like "immediate or cancel" (IOC), meaning that they must be executed immediately or not at all, or "fill or kill" (FOK), meaning that they must be executed in full or not all. There is a virtual alphabet soup of overall options, so we will not try to give a comprehensive accounting here. We will discuss more order types later to the extent that they are relevant to higher level issues of market interactions.

Now that we've given a brief overview of how brokers communicate their desires to venues, we'll move to discussing how venues communicate information about trades and prices to their members or other financial participants:

Tapes

Publicly traded stocks in the US are organized into 3 groups called Tapes. Tape A has NYSE-listed securities, Tape C has Nasdaq-listed securities, and Tape B has everything else. As discussed above, a security is "NYSE-listed" if NYSE is the venue responsible for

running its official opening and closing auctions (among other duties), but it can trade on all venues at all times.

For a particular stock, these are the types of information that participants are most interested in:

Top of book and depth of book data

"Top of book" and "depth of book" both refer to data about the best prices and associated quantities that would-be buyers and would-be sellers are currently advertising on a particular venue. The prices and quantities that buyers are advertising are called "bids," and the prices and quantities that sellers are advertising are called "asks" or "offers." For example, a bid might be "would buy 100 shares of security X for $40 a share." This is of course a wordier version than the shorthand formats that participants use, but this is what they mean. Similarly, a sample offer means: "would sell 200 shares of security X for $40.02 a share."

At a particular trading venue, if a buyer was advertising a price that was equal to or higher than a price being advertised by a seller, the buyer and the seller could just trade with each other (at least for the minimum of the two share quantities). Since the venue will catch this and match the trade, the highest bid price at a given venue should always be lower than the lowest offer price. The difference between these is called the "spread" - we've mentioned this before as a driver for compensation to market makers who are constantly willing to buy or sell. The collection of bids at the current highest bid price and the collection of offers at the current lowest offer price are referred to as the "top of book." Data feeds that are categorized as "top of book" include only information about these advertised orders, and not any bids at lower prices or offers at higher prices. Naturally, the current highest bid price and current lowest offer price are moving targets as trades happen and advertised bids and offers are canceled or adjusted. Top of book data feeds are typically event-driven, meaning that as soon as the top of book information changes, the new information is immediately disseminated. A top of book data feed might aggregate the sizes of bids or offers at a particular price, or it might give granular information about each bid/offer.

"Depth of book" refers to information about bids and offers that goes beyond the highest priced bids and the lowest priced offers. A typical depth of book data feed will include information about all bids and offers, but there are some variations on this. Some depth of

book data feeds aggregate sizes of the bids or offers at a particular price, some given granular information on each bid/offer, and some place a limit on the number of distinct prices (sometimes called "levels") that are included. For instance, the CBOE exchange family offers some data feeds that include information only about the 5 highest price levels for bids and the 5 lowest price levels for offers.

As an interesting side note, the highest bid price and the lowest offer price *across venues* can be equal, a state which is referred to as a "locked market." The highest bid price can even be strictly higher than the lowest offer price when we look across venus, which is referred to as a "crossed market." These situations do arise fairly frequently, but are typically quickly resolved, as someone recognizes an opportunity. You might expect that if a seller on one venue and a buyer on another have compatible desires, then both would likely consume the relevant data, recognize the compatability, and one of them would send an order to the other venue to trade. But even if both participants are tracking the bids and offers on the other venue and see each other, billing practices at venues can complicate this story. Many of the exchanges are "maker-taker" venues, meaning that they charge a fee to the entity who submits an order that executes against an advertised bid/offer (aka the "taker") and they give a rebate to the entity whose advertisement led to the trade (aka the "maker"). If a buyer is advertising a price P on a maker-taker venue and a seller is advertising that same price P on a different maker-taker venue, both entities might be banking on receiving the rebate instead of the fee to view that price P as worthwhile. This sometimes creates an amusing standoff until someone changes their view of the price they are willing to pay, or another participant intercedes. This scenario couldn't explain a crossed market though, since the range of rebates/fees is capped to be smaller than the 1 penny price increment allowed for bids and offers.

Trades

In addition to quotations of what would-be buyers and sellers are advertising, market participants also want data about trades as they occur. Once a trade is executed at a venue, the venue sends confirmations to the involved parties, as well as disseminating some information to the larger market ecosystem about the trade. The broader population can learn what security was traded, how many shares, at what price, and at what time. The identities of the buyer and seller are not revealed. In fact, even the trading parties themselves will

typically not know each other's identities. If a trade happens on an exchange, the identity of the exchange will be publicly visible as well. If it happens on a dark pool, the trade will be identifiable as having occurred off exchange, but the identity of the specific pool will not be publicly visible.

Auction Data and Alerts

Leading up to an auction, an exchange may choose to make some data available concerning the possible price or the imbalance between orders to buy and orders to sell. This kind of data may be used by participants to adjust their orders leading up to the auction, and may motivate new buyers and/or sellers to participate.

There are also alerts dissimentated through data feeds that participants might need or want to consume. For example, alerts may be sent out when trading is halted in a stock.

Market Data Dissemination and Trade Reporting

Now that we've talked about what kind of market data participants typically seek and receive, we'll discuss the data feeds that disseminate these types of data and the entities that control them.

SIPs

The SEC has mandated that some market data be made publicly available via Securities Information Processors (SIPs). Currently there are two SIPs, which are operated by NYSE and Nasdaq, respectively. The NYSE SIP (CTA) publishes trades and quotes for Tape A and Tape B securities, and the Nasdaq SIP (UTP) publishes trades and quotes for Tape C securities. In terms of quotes, the SIPs provide only top-of-book information from each exchange. A change in the top-of-book at a given exchange is first communicated by the exchange to the SIP and then from there disseminated to recipients of the SIP data feeds. This two hop approach is unsuprisingly slower than sending the new data directly from each exchange to the ultimate recipient.

Even though these are "public" data feeds that lack depth-of-book or auction imbalance information, they are still extremely expensive to connect to, consume, and redistribute,

especially for real-time data.

Proprietary data feeds

Exchanges also offer multiple kinds and tiers of proprietary data feeds. These products run the gamut from trades and top-of-book quotations to full depth-of-book at the particular exchange. Because they can be consumed directly from the exchange without the data first traveling to a central processor as is the case for the SIPs, these proprietary feeds can deliver data at the lowest latencies (especially to recipients who are co-locating their equipment in data center space provided by the exchange). Some of the data products are free (e.g. all of the data feeds from IEX or NYSE National), some of them are a bit cheaper or comparable in pricing to the SIPs (e.g. Nasdaq Basic), but the full depth-of-book feeds at the main exchanges with the highest market share are an order of magnitude more expensive.

For more details on market data pricing for the real-time data products offered by exchanges, see our market data series of blog posts (part 1, part 2, part 3) and our market data pricing visualization tool.

TRF

The FINRA Trade Reporting Facilities are where information about trades that happen off of exchanges is reported and then relayed to the appropriate SIP for public dissemination. There are currently two TRFs, also operated by NYSE and Nasdaq, but unlike the SIPs the TRFs are not partitioned by Tape. Instead, each dark pool may strike an agreement with the TRF of its choosing to report its trades. Such trades are ultimately visible to others via the SIPs, though with potentially greater latency than exchange trades and with the precise venue not specified. In other words, the TRFs allow other market participants to learn the price, size, and time of off-exchange trades, but only the binary attribute of "off-exchange" is visible/inferrable, not further granularity on which off-exchange venue produced the trade.