Chapter 4
Secondary Market: Trading, Price Discovery, and Order Matching

Reto Francioni, Martin Reck, and Robert A. Schwartz

4.1 Price Discovery

4.1.1 Importance of Price Discovery

Price discovery is achieved as orders are submitted to a market and turned into trades. A transaction price is, of course, determined each time a trade is consummated, but *price discovery* refers to something more fundamental. Price discovery refers to the search for a value that best reflects the broad market's desire to hold shares of a stock. In economic parlance, price discovery involves the search for an *equilibrium value*. While price determination occurs on a trade-by-trade basis, price discovery is achieved only as a substantial set of orders is brought together, generally over a succession of trades.

An exchange's ability to deliver good price discovery depends on its market structure, namely the rules, procedures, and technology that define the exchange's trading platform (we delve into market structure later in this chapter). More efficient market structure enables the delivery of more robust price discovery for the broad market. The challenge, however, is formidable; equilibrium values are unobservable, they are continually subject to change, and they are not easily attained.

R. Francioni (✉)
University of Basel/Deutsche Börse Group, Basel, Switzerland
e-mail: reto.francioni@unibas.ch

M. Reck
Deutsche Börse AG, Mergenthalerallee 61, Eschborn 65760, Germany
e-mail: martin.reck@deutsche-boerse.com

R.A. Schwartz
Zicklin School of Business, Baruch College, CUNY,
55 Lexington at 24th Street, 268, New York, NY 10010, USA
e-mail: Robert.Schwartz@baruch.cuny.edu

© Springer International Publishing Switzerland 2017

R. Francioni, R.A. Schwartz (eds.), *Equity Markets in Transition*,
DOI 10.1007/978-3-319-45848-9_4

An exchange's economic function is not simply to make the transactions but (and this is the bigger challenge) an exchange also "*produces*" *the prices* at which the transactions are made. Delivering quality price discovery is a defining function of a stock exchange. From a definitional perspective, any trading facility that has as its primary function the delivery of good price discovery can, de facto at least, be considered an exchange. Unfortunately, however, the price discovery function of an exchange typically receives insufficient attention in market structure discussions. This is largely attributable to the non-observability of equilibrium prices and, therefore, to the difficulty of quantifying the deviations of transaction prices from their equilibrium values.

4.1.1.1 Expectations

Market participants commonly believe that shares have *fundamental values*. The concept of a fundamental value would apply in an environment where everyone who is in possession of the same *fundamental information* forms identical, *homogeneous expectations* of future share value. This commonly accepted share value would then be the stock's equilibrium price, and the price of shares need not be discovered in the marketplace.

Homogeneous expectations, for good reason, are commonly assumed in academic modeling (to wit, it is a key assumption in the capital asset pricing model). The reason for this assumption is completely understandable—as a simplifying device, assuming homogeneity can make a complex theoretical model tractable. In actual markets, however, expectations are not homogeneous. Rather, participants in possession of identical information concerning a company's fundamentals generally form *divergent expectations* based on that information. A divergence in beliefs is attributable to the sheer magnitude, complexity, incompleteness, and possible unreliability of the information set that pertains to a specific company, an industry, or the broad economy. Simply stated, in a divergent expectations environment, if some participants think, for instance, that a stock should be valued at $25 a share while others assign a value of $30, what is the stock's fundamental value, $25 or $30? The answer is "neither." Stocks cannot have fundamental values when the expectations of market participants are divergent. In a divergent expectations environment, share prices are not discovered in the research offices of the analysts—they can be found only in the marketplace where buy and sell orders meet and are turned into trades.

4.1.1.2 Public Goods

The ability of an exchange to deliver reasonably accurate price discovery is of overriding importance. It is not just the parties to a trade who care about price; a far broader public uses exchange-produced prices for a wide spectrum of purposes that include marking to market, derivative pricing, valuations of mutual fund cash

flows, estate valuations, and dark pool pricing. To turn to a nautical analogy, an exchange-produced price shares properties in common with a lighthouse. A lighthouse aluminates the presence of a harbor or the location of a rock; an exchange-produced price sheds light on the value of shares. The beam from the lighthouse benefits any ship that is passing in the night; the light cast by an exchange-produced price benefits the broad investment community.

In economic terms, both a lighthouse and an exchange produce a *public good*. It is well understood in economics that public goods are undersupplied in a private economy and, accordingly, that they must be provided by government. This is indeed the case for a lighthouse, and it is for an exchange as well. Regarding price discovery, an exchange performs a quasi-governmental function of major importance.

Another key consideration is *liquidity* provision. Price discovery and liquidity provision interact in a mutually supporting manner: one would expect price discovery to be sharper in a more liquid market and, reciprocally, that liquidity provision would be more forthcoming in a market that delivers better price discovery. Liquidity, however, is a slippery concept to define and hard to measure; the accuracy of price discovery is even more difficult to quantify (as we have said, *equilibrium values* are not observable).

The quality of price discovery is assessable, however. For one critical reason, this can be done with the use of an intraday volatility metric. The reason? Prices, in searching for equilibrium values, exhibit accentuated volatility. Here is how it works.

4.1.2 Mean Reversion, Returns Autocorrelation, and Accentuated Volatility

The price path from one equilibrium to another rarely follows a straight line. Rather, prices bounce around, describing a jagged path that, with momentum moves (and herding), can cause prices to overshoot new equilibrium values and then reverse course. Prices that systematically fall after having risen (or which rise after having fallen) are said to *mean revert*. A good way to visualize mean reversion is to picture prices first swinging up and then down (or down followed by up) within a trading range. When prices mean revert, a sequence of returns (price changes) is negatively autocorrelated. With negatively autocorrelated returns, prices are not following a random walk. Instead, price increases (or a run of increases) are more apt to be followed by decreases, and price decreases (or a run of decreases) are more apt to be followed by increases.

Mean reversion and its counterpart, negative returns autocorrelation, are present in short-period price movements (e.g., intraday returns), but they decay as one moves to returns measured over longer intervals of time (e.g., a day or more). The price volatility accentuation that is associated with negative returns autocorrelations also decays as one moves to longer measurement intervals. Consequently, the quality of price discovery can be inferred by matching very-short-period price volatility with longer period price volatility.

In a *frictionless* world of perfectly accurate price discovery (that is, in a random walk world), the variance of returns will increase proportionately with the length of the interval used to measure them. For instance, the variance of a distribution of five-day returns will be five times that of a distribution of one-day returns. Thus, a five-day returns variance that is *less* than five times a one-day returns variance indicates that the one-day returns are negatively autocorrelated (i.e., are mean reverting). Equivalently stated, the lower five-day variance is evidence that the one-day return variance is *accentuated* (not that the five-day return variance is depressed). We suggest, first and foremost, that the accentuation is attributable to price discovery being a complex, noisy process which is replete with jagged price moves, overshooting, and mean reversion.

For this reason, the quality of price discovery can be inferred from an intraday volatility analysis. To do so, alternative volatility measures can be employed, with the most popular being variance (or standard deviation) and a high-low range.

4.1.2.1 Volatility Analysis: Evidence

Alan and Schwartz (2013) assessed the level of intraday volatility for a sample of 30 Dow stocks, presenting examples of stock/day-specific opening half-hour volatility for the year 2011. In this subsection, we present a condensed version of the relevant part of that paper.[1]

The purpose of Alan and Schwartz's analysis was not to assess an average level of volatility across a large, all-inclusive set of stocks, but to hone in on the higher levels that volatility can reach in a brief, opening half-hour interval. To achieve this, for all US stocks for each trading day in 2011, they first calculated, for each stock on each day, an opening volatility measure and a spread-adjusted opening volatility measure that are based, not on a variance statistic, but on a stock's high-low price range:

$$\text{Opening volatility} = \frac{P^{\max} - P^{\min}}{P^{\text{mean}}} \tag{4.1}$$

$$\text{Adjusted volatility} = \frac{P^{\max} - P^{\min} - Spread}{P^{\text{mean}}} \tag{4.2}$$

where $P^{\max}$, $P^{\min}$, and P^{mean} for a given stock and day are the highest, lowest, and average trade prices, respectively, during the first half-hour of trading (9:30 AM to 10:00 AM), and *Spread* is the (time-weighted) average bid-ask spread over the same half-hour interval. The opening high-low volatility measure captures the range of price movements over the 30-min period; to get a sharper read on price discovery, this measure is adjusted by subtracting the bid-ask spread from the interval's high-low prices.

[1] This material is printed with permission of the *Journal of Portfolio Management*.

Table 4.1 Selected stock/day examples of opening volatility

Company Name (Ticker)	Date	Avg Price	Hi-Lo	Spread	Volatility	Adjusted Volatility	Group*
JOHNSON & JOHNSON (JNJ)	04/06/11	$59.82	$0.20	$0.01	0.33%	0.31%	5
BOEING (BA)	06/29/11	$72.34	$0.38	$0.03	0.53%	0.49%	6
HOME DEPOT (HD)	04/20/11	$38.34	$0.25	$0.01	0.65%	0.62%	7
MERCK (MRK)	09/27/11	$32.06	$0.25	$0.01	0.78%	0.74%	8
TRAVELERS COMPANIES (TRV)	08/04/11	$53.32	$0.48	$0.02	0.90%	0.86%	9
EXXON MOBIL (XOM)	05/26/11	$81.92	$0.82	$0.01	1.00%	0.99%	10
PROCTER & GAMBLE (PG)	04/12/11	$62.52	$0.71	$0.01	1.14%	1.12%	11
MCDONALDS (MCD)	10/05/11	$86.19	$1.12	$0.04	1.30%	1.26%	12
WALMART (WMT)	01/20/11	$55.66	$0.80	$0.01	1.44%	1.41%	13
AMERICAN EXPRESS (AXP)	09/26/11	$46.59	$0.77	$0.03	1.65%	1.60%	14
UNITED TECHNOLOGIES (UTX)	02/24/11	$82.74	$1.53	$0.03	1.85%	1.81%	15
UNITEDHEALTH GROUP (UNH)	12/08/11	$49.25	$1.05	$0.02	2.13%	2.08%	16
VERIZON (VZ)	08/01/11	$35.80	$0.89	$0.01	2.49%	2.46%	17
DU PONT (DD)	08/05/11	$47.88	$1.45	$0.02	3.03%	2.99%	18
JPMORGAN CHASE (JPM)	08/25/11	$37.69	$1.57	$0.01	4.17%	4.14%	19
DISNEY (DIS)	08/10/11	$30.34	$2.31	$0.02	7.61%	7.55%	20

(*There are no Dow stock observations in the first four groups, therefore our table starts from Group 5)

The stocks were next sorted by their adjusted volatility and divided into 20 groups of equal numbers. Group 1 comprised the stock/day observations with the lowest adjusted volatility, and Group 20 comprised the stock/day observations with the highest adjusted volatility.[2] From this all-inclusive set of stocks, the Dow stocks only were selected for the analysis. For each of the 20 groups, Alan and Schwartz selected the single Dow stock that had the highest single-day volatility in the group. The process resulted in 16 observations, which are shown in Table 4.1. Note that no Dow stock/day observation was located in any of the four lowest volatility groups.

Table 4.1 gives the company name and ticker, date of the observation, average price during the opening half-hour, dollar difference between the highest and the lowest price, average spread, opening volatility, spread-adjusted opening volatility, and group to which the observation belongs. On the low end of the spectrum, on April 6, 2011, Johnson & Johnson (at the time, a $60 stock) had a $0.20 price fluctuation in the first half-hour, a spread of $0.01 (2 basis points), and an adjusted volatility of 0.31 %. At the high end of the spectrum, Disney (at the time, a $30 stock), on August 10, experienced a $2.31 price fluctuation in the first half-hour of trading with an average spread of 2 cents (7 basis points). Concurrently, Disney's adjusted high-low was a very substantial 7.55 %. For all 16 observations, the spread-adjusted price volatility displayed in Table 4.1 is indicative of a component of volatility that we suggest represents appreciable price discovery noise.

[2] Alan and Schwartz further imposed a price filter that restricts the sample to stocks in the $30–$100 price range.

Fig. 4.1 Price path for SPY vs. DIS during the opening half-hour interval on August 10, 2011

Of particular interest is the high end of the spectrum. On this end, Disney, on August 10, 2011, clearly stands out; let us focus on it. On that date, the market for Disney (and the broad market as well) was under stress, as the markets were profoundly rattled by the European debt crisis. Might this explain the high first half-hour volatility? Macro uncertainty is certainly an underlying causal factor, precisely because price discovery is more difficult at times when uncertainty is high and people's expectations about what the future will bring are more divergent. Nevertheless, the question remains this: what could account for one person buying shares at a price that was 7.61 % percent higher than the price at which someone else sold shares within the same half-hour interval when the average spread was only $.02, as occurred on August 10 for Disney stock?

Alan and Schwartz questioned whether or not a fresh news release from Europe (or any other news event during that particular half-hour) could be the cause. A search of LexisNexis revealed no major news announcements at this time for either Disney or the broad market. Neither does Disney's price path suggest the advent of a major news announcement in the opening 30 min of trading on that day. Figure 4.1 shows, second by second, for that first half-hour, how Disney's (DIS) price evolved, side by side with the price of the SPDR S&P 500 ETF (SPY).[3] In Fig. 4.1, DIS's prices are on the left-hand axis and its chart is the solid line; SPY's prices are on the right-hand axis and its chart is the dashed line.

[3] To suppress the effect of price changes attributable to the bid-ask spread and to reduce the effect of out-of-sequence reporting, the prices shown in the exhibit are averages for all trades that occurred in each of the 1800 s that comprise the first half-hour (on that day, DIS averaged 37 trades per second, while SPY averaged 124 trades per second).

For DIS, there is initial volatility and a bump up in the first minute, a predominantly downward trend until 9:37 AM, a predominantly upward trend until 9:50 AM, falling prices for the next couple of minutes, and lastly an uptrend to 10:00. The picture for SPY is simpler: falling prices until 9:35, an upward trend until 9:48, and primarily falling prices to 10:00. Comprehensively viewed, both paths display mixtures of trending and reversals, and the two paths are weakly correlated with each other (the correlation is .19 for 30-s returns and .47 for 1 min returns).

From this evidence, one can infer that intra-half-hour news release is not the cause of the observed price movements for DIS. We suggest that the more plausible cause is the dynamic process of price discovery. Apparently, the August 10 opening price did not adequately reflect the broad market's desire to hold Disney shares. We suggest that the substantial price changes which ensued for at least the next 30 min largely reflected the market searching for a price that better balanced the opposing pressures exerted by a diverse population of buyers and sellers whose expectations, given the greater uncertainties that prevailed at that time, were on that day strikingly divergent.

After having focused on one stock (DIS) in particular, Alan and Schwartz proceeded to consider the full set of 30 Dow stocks over all 252 trading days in 2011. In this assessment, each stock/day observation was assigned to a volatility group.[4] Summary statistics of the adjusted opening volatility for each of these groups are given in Table 4.2. The mean, adjusted volatility ranges from 0.28 % for the lowest volatility group to 5.56 % for the highest volatility group.[5] The faster rise in average volatility among the higher volatility groups is striking: while group 18 has an average volatility of 2.71 %, the average reaches 3.49 % in group 19, and 5.56 % in group 20. Table 4.2 also shows the number (N) and the percent (%N) of Dow observations in each of the 20 groups. Out of the volatility observations for all Dow stocks in 2011, roughly 43 % fall into groups 11–20. In other words, almost half of the Dow stocks experienced an opening volatility that is higher than the median volatility across all stocks. Clearly, it is not just the small cap stocks that experience high volatility—the largest stocks in the economy clearly exhibit accentuated volatility in the first half-hour of trading as well.

4.1.2.2 Monitoring Volatility

Having a volatility auction at times of high volatility insures having (1) a price discovery process in place even when continuous trading has to be interrupted due to larger price movements, (2) and not only allows for the pricing of the underlying stock, but also provides a price point for the respective derivatives instruments related to that stock. Related to indexes, the calculation of their values can (3) continue and is possible at any time during normal trading hours and, like for stocks, (4) any derivative product defined based on such an index can continue to be priced and traded (Fig. 4.2).

[4] The same stock was allowed to fall into different volatility groups on different days.

[5] Except for the three highest volatility groups, means and medians are virtually identical.

Table 4.2 Summary statistics of the adjusted opening volatility by group

Group*	Mean	Median	Min	Max	N**	% N
5	0.28%	0.28%	0.16%	0.31%	77	1.02%
6	0.42%	0.42%	0.32%	0.49%	607	8.03%
7	0.56%	0.56%	0.49%	0.62%	939	12.42%
8	0.68%	0.68%	0.62%	0.74%	1032	13.65%
9	0.80%	0.80%	0.75%	0.86%	854	11.30%
10	0.92%	0.92%	0.87%	0.99%	780	10.32%
11	1.05%	1.05%	0.99%	1.12%	754	9.97%
12	1.18%	1.18%	1.12%	1.26%	600	7.94%
13	1.33%	1.33%	1.26%	1.42%	503	6.65%
14	1.50%	1.49%	1.42%	1.60%	412	5.45%
15	1.70%	1.70%	1.60%	1.81%	294	3.89%
16	1.94%	1.93%	1.81%	2.09%	259	3.43%
17	2.25%	2.24%	2.09%	2.46%	197	2.61%
18	2.71%	2.68%	2.46%	3.02%	148	1.96%
19	3.49%	3.40%	3.03%	4.14%	78	1.03%
20	5.56%	4.86%	4.19%	8.89%	26	0.34%

*There are no Dow stock observations in the first four groups, therefore our table starts from Group 5.

**Total number of observations is 7,560 (30 stocks * 252 trading days).

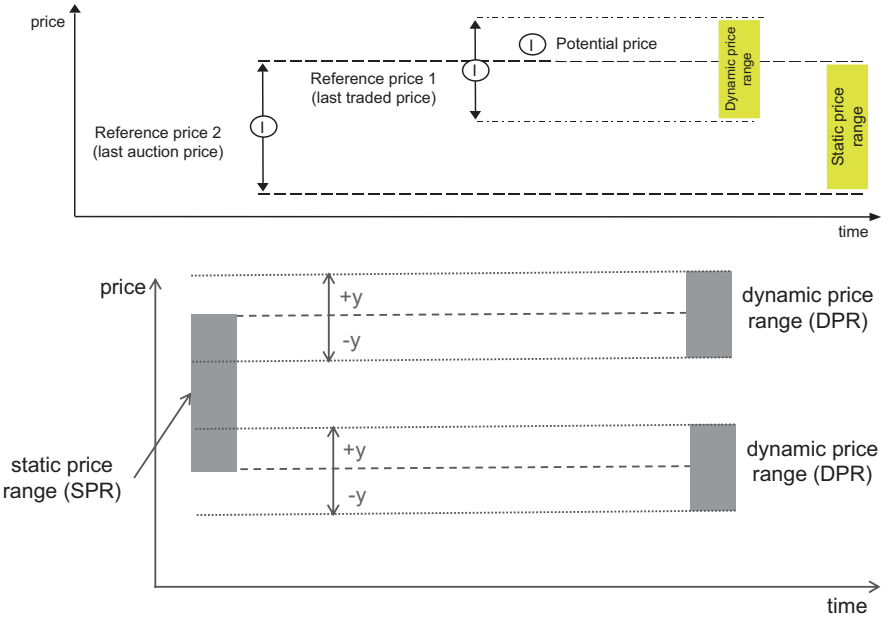

Fig. 4.2 Dynamic vs. static price range

An example illustrating volatility is shown in Fig. 4.3:

1. After rumors emerged in the market that CME Group was planning to make Deutsche Börse Group an acquisition offer, DBG's share price rose to a maximum of € 52.30 (+12 %).
2. After the communication of an ad hoc announcement, the price dropped to € 47.50. The share closed at € 49.30 (+5.6 %) with a turnover surpassing the daily average on 25 February three times.

Trading volume slowly decreased as DBG's communication department denied rumors until the release of the ad hoc (Fig. 4.4).

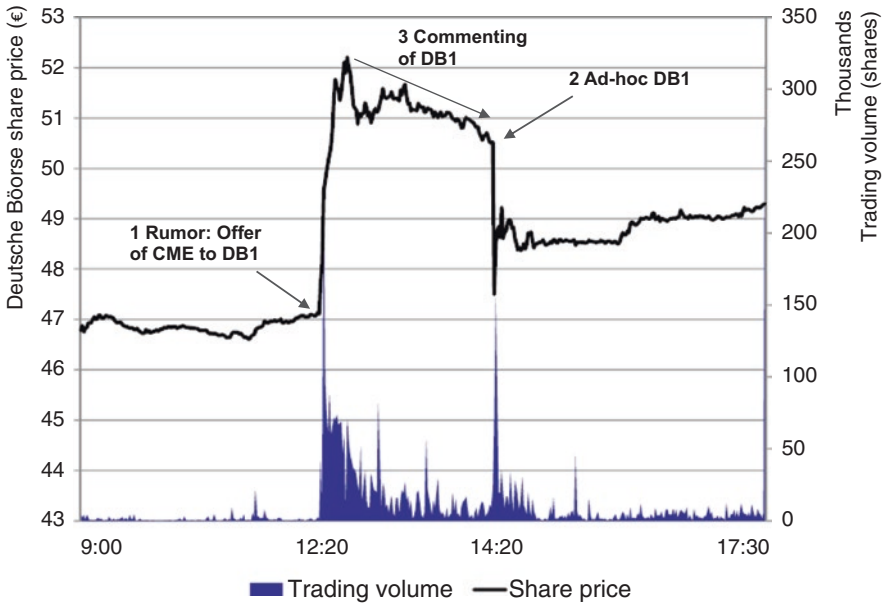

Fig. 4.3 Intraday volatility (Deutsche Börse Group example, 25 February 2013)

Volatility Interruptions in DB1 Shares on 25 February 2013						
No.	Price before VI	Start of VI	End of VI	Price after VI	Volume	Comments
1	48.120 €	12:18:02	12:20:25	49.000 €	29,025	
2	50.440 €	12:25:54	12:28:17	50.600 €	26,469	
3	52.110 €	12:36:33	12:38:56	51.900 €	5,208	
4	50.950 €	14:19:17	14:22:07	48.000 €	46,821	Extended Volatility Interruption
5	47.500 €	14:22:07	14:24:30	47.995 €	78,951	

Fig. 4.4 Volatility interruptions in DB1 Shares on 25 February 2013

4.2 Market Structure

Having recognized (a) the importance of an exchange delivering reasonably accurate price discovery, (b) that the quality of price discovery can be assessed by an intraday volatility metric, and (c) that intraday, first half-hour volatility can be strikingly high, we have one further matter to address at this time: the relationship between the quality of price discovery and a market's architecture.

By market architecture, we are referring to an exchange's rule book, trading systems, and technology. With each of these, clear alternatives exist. Here are some highlights. In a continuous market, trades that are generally bilateral are made whenever a buy and sell order meet or cross in price; in a periodic call market, orders are batched together for simultaneous execution at a single point in time at a single price. On an organized exchange, price is the primary rule of order execution (highest bids are matched with lowest offers), but when the most aggressive orders are tied in price, a secondary rule of order execution is called for; the rule could be time priority (first in, first out), size priority (the largest orders execute first), or pro-rata execution. Designated *market makers* may or may not be included to facilitate liquidity supply. Along with standard market and *limit orders*, an assortment of alternative order types and instructions are generally available (e.g., fill-or-kill orders, all-or-nothing orders, and hidden or iceberg orders). Small retail orders are typically handled in one way, and large block orders in another. A marketplace may be integrated or fragmented, and trading can be transparent or opaque. Systems can be predominantly electronic or driven by human intermediaries. A trading environment may be based on a single modality or it can be a hybrid.

This overview of market structures can be extended in both scope and detail, and we turn to major alternatives later in this chapter. The important point to make at this time is that market quality is not an exogenous variable it very much depends on how order flow is integrated in the process of delivering trades and producing prices. Choice exists, and unanswered questions as to what is best persist. Market architecture remains a work in process.

4.2.1 Continuous Trading in Order-Driven Markets

4.2.1.1 The Link to Continuous Trading: The Spread

No spread exists as a *call auction* book builds with buy orders meeting and crossing sell orders in price. But, after the call has been completed, unexecuted orders remain on the book (unless otherwise instructed) and, because all matching and crossing orders have been executed and are no longer on the book, a spread necessarily exists between the highest posted bid and the lowest posted offer in the continuous market that follows the call. This spread between posted orders continues to exist as the continuous market progresses, widening with the elimination of previously posted

orders and shrinking with the arrival of new orders that are not priced aggressively enough to execute upon arrival. The book at the completion of a call, and the spread between unexecuted orders that characterizes the start of the continuous market, is shown diagrammatically in Fig. 4.5.

After applying the matching rules, the market in the above traded stocks looks as follows:

- The execution price (EP) is 100: three shares are matched on each side and both orders execute fully.
- The spread is two, which means for the investor: the best bid to sell to is 99, the best offer to buy at is 101. If one were to inquire what the market is, the answer would be "99–101."
- The remaining orders in the *consolidated limit order book* (CLOB) stay there, ready for matching if incoming orders drive the price in their direction and a match is reached. These booked orders build both the market's breadth (at the price at which they have been placed) and its overall depth, thereby making the market more liquid.

Because there is no surplus of unexecuted orders at the execution price of 100, no order is left on either side of the book at 100. Therefore, the spread is now 99–101.

Referring again to Fig. 4.5, if two orders rather than one were placed at 100 (=EP), the cumulated number of buy orders at 100 would be four. Hence, one order would be left on the book on the buy side after the call has been completed. Accordingly, the quotes and their attending spread after the execution of three orders on each side would be 99–100 and 1, respectively. Reciprocally, if the overhang of one order was on the sell side, the quotes and the spread at the open of the continuous market would be 100–101 and 1, respectively.

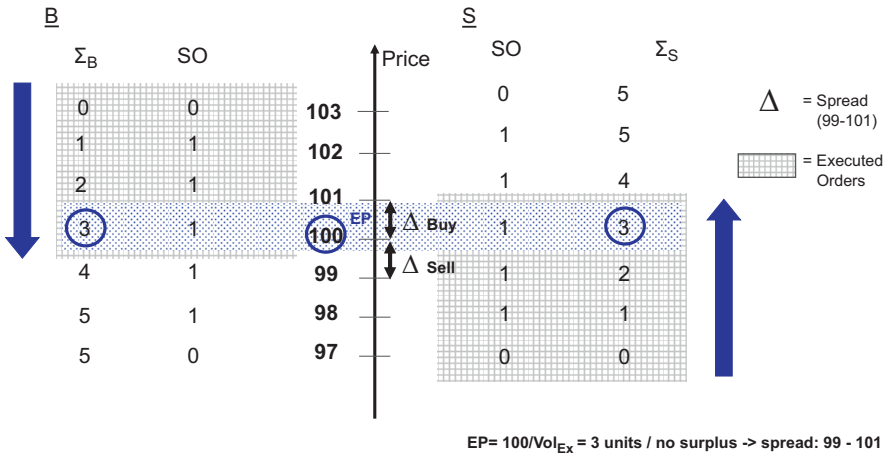

Fig. 4.5 The origin of the spread

Market Orders	Limit Orders
• Execute at the best counterpart • Execute immediately • Immediacy/liquidity demanding	• Execute at price of the order • Delayed or no execution • Immediacy/liquidity supplying

Fig. 4.6 Market orders vs. limit orders

Several broad points can be made about order placement in a continuous market environment[6]:

- Competition among dealers and limit order traders keeps spreads tighter.
- Dealer spreads depend on the costs they incur when providing immediacy.
- The threat of informed trading widens spreads.
- Uninformed traders in particular bear the cost of paying spreads that are wider because of the risk of adverse selection.
- Spreads are tighter for more liquid, better known, large cap stocks.
- Large anonymous traders are widely thought to be better informed.
- Limit-order traders give options to traders who can respond more quickly to changing market conditions.

4.2.1.2 Market Orders and Limit Orders

We alluded to some of the alternatives for market structure in the previous subsection. The two basic structures are continuous order-driven markets and continuous quote-driven markets, where "continuous" means that a trade can be made at any point in time that the market is open and a buy and sell order either meets or crosses in price. In a continuous market, trades are generally bilateral (i.e., one trader's buy order executes against another trader's sell order) as distinct from a call auction where trading is generally a multilateral (batched) matching (Fig. 4.6).

In a pure order-driven market, the orders of some public traders set the prices at which other public traders can buy or sell without the participation of an intermediary. In a pure quote-driven, dealer-intermediated market, a dealer's ask quote establishes the price at which a public trader can buy shares, and a dealer's bid quote establishes the price at which a public trader can sell shares. In this section, we focus on the former, the continuous order-driven market.

The viability of an order-driven market depends on the willingness of some public participants to place limit orders, and on the willingness of other public participants to place market orders. Without limit orders, market orders could not

[6] Schwartz, Robert A./Francioni, Reto (2004): Equity Markets in Action. The Fundamentals of Liquidity, Market Structure & Trading. Hoboken: Wiley.

execute; without market orders, limit orders could not execute. They need each other. Without market orders and limit orders coexisting, an order-driven market would fail. In essence, the limit order placers provide liquidity and immediacy to market order traders who are seeking liquidity and immediacy.

A limit order is so named because the trader who has placed it has stated a price limit at which shares are to be bought or sold. For a buy limit order, the price limit is the maximum share price the trader is willing to pay; for a sell limit order, the price limit is the minimum share price the trader is willing to accept. Alternatively stated, at any price greater than the buyer's limit, the buyer does not wish to acquire the shares; at any price lower than the seller's limit, the seller does not wish to dispose of the shares.

Market orders, on the other hand, are unpriced orders. The trader who has submitted a market buy order is willing to buy at the lowest posted offer, which would be the price established by the most aggressive (lowest priced) limit sell order. The trader who has submitted a market sell order is willing to sell at the highest posted bid, which would be the price established by the most aggressive (highest priced) limit buy order.

Limit orders are generally posted on the limit order book, with buy limits placed below the lowest market ask and sell limits placed above the highest market bid. A limit buy order priced above the best market ask (or a limit sell order priced below the best market bid) is referred to as a *marketable limit order*. Marketable limit orders that are larger than the best market ask (or bid) will *walk the book* (that is, buy orders will execute at successfully higher prices and sell orders will execute at successfully lower prices) until they are fully executed or reach their limit price, at which point any unexecuted portion of the order will be entered in the book as a regular limit order.

4.2.1.3 Costs and Benefits of Market Orders and Limit Orders

Trading by market order conveys one benefit: it enables the participant to trade immediately and, in so doing, to achieve certainty of execution. But a market order strategy entails a cost. Assuming that the order is not large enough to walk the book, the market order trader buys at the ask or sells at the bid and, in so doing, pays the spread. The spread, however, is the cost of a round trip, and thus half of the spread is taken to be the cost of each leg of a round trip.

Trading by limit order saves the spread, but incurs a cost of its own: a limit order on the book may not execute and, if it does execute, it might do so for an undesirable reason. Limit order traders, like market makers, are posting quotes that enable others to trade. A market maker is in business to provide liquidity and immediacy to others. He will sell, not because he wishes to hold fewer shares, but because a public participant wishes to buy; or he will buy, not because he wishes to hold more shares, but because a public participant wants to sell. A limit-order trader, on the other

hand, seeks to buy or to sell for the precise purpose of adjusting his or her portfolio holdings. If he or she posts a limit order and it does not execute, he or she has failed to make that portfolio adjustment and that, to him or her, is a cost.

A limit order executes when it (a) gets to the top of the book, (b) gains time priority by being first in the queue at the best bid or offer, and (c) is hit by a market order. The trade-initiating market order could have been part of a temporary buy/sell imbalance, or it could have been motivated by news that, from the perspective of the limit-order placer, was adverse information. When the limit order executes because of adverse information, the limit-order trader bears the cost of "adverse selection" and suffers what is known as "ex post regret."

But trading by limit order is beneficial when the order is executed because of a temporary buy/sell imbalance. In the microstructure literature, the imbalance is attributed to participants buying and selling shares for their own "liquidity" purposes when a fresh receipt of cash is realized or a new need for cash is incurred. A buy/sell imbalance that is informationless can push price to (and also past) the price of a posted limit order, trigger an execution, and then revert back to its former level. After his or her order has executed, the limit-order trader benefits from this reversion. The price reversion is referred to as *mean reversion*, the tendency of price to move back to its "mean" (i.e., average) value after it has been pushed away. Mean reversion in prices translates into accentuated short-period price volatility, as we have discussed previously.

Recognizing that the compensation for limit-order trading is realized through mean reversion and accentuated price volatility, we note that a certain amount of mean reversion (and the associated volatility accentuation) is a natural property of a continuous, order-driven trading environment. If the limit order book is very thick and the bid-ask spread is tight, price dislocation and mean reversion will be minimal, the compensation for placing a limit order will be low, and fewer of these orders will be placed. At the other end of the spectrum, if the book is thin and spreads wide, mean reversion will be strong and a greater number of limit orders will be placed. When the book is in balance, the spread is just wide enough and mean reversion is just strong enough to appropriately compensate the limit-order traders for accepting the risk of not executing, along with the risk of executing because of adverse information change.

This balance between limit orders and market orders also underlies the natural (an economist would say "equilibrium") size of a stock's bid-ask spread. A spread that is "too tight" reduces the benefit of trading by limit order but does little to reduce the risks of non-execution and adverse selection. Thus, a spread that is "too tight" leads to more market orders being placed relative to limit orders, and hence the spread is widened. At the other end of the spectrum, a spread that is "too wide" leads to more limit orders being placed relative to market orders, and hence the spread is tightened. When the spread is of appropriate magnitude, the likelihood of it widening when it next changes equals the probability of it tightening.

We conclude this discussion with the thought that, because the risks of trading by limit order cannot be eliminated by placing a limit order sufficiently close to a

counterpart quote that has already been placed on the book, a bid-ask spread is a natural property of a continuous order-driven market, just as is mean reversion and accentuated short-period price volatility.[7]

4.2.1.4 Transparency and a Consolidated Limit Order Book

It follows from our prior discussion that a participant in a continuous order-driven market is led to make strategic decisions: most importantly, whether to submit a market order or a limit order and, if a limit order, the price at which that order is placed. Knowledge of the configuration of the limit order book is critical to making these strategic decisions, as is information concerning recent prices, quotes, and trades. Simply put, reasonable pre-trade transparency and post-trade transparency are both essential for a continuous order-driven market to operate efficiently.

Order flow consolidation is also important. Consolidating the order flow facilitates enforcing price priority across all orders that have been sent to the market. It also enables a secondary priority rule to be imposed across all orders (e.g., time priority). Consolidated order flow, with price and time priority enforced, bolsters competition between all orders that have been sent to the market. Additionally, the consolidation of market information facilitates the formulation of order placement strategies.

4.2.1.5 Limitations of the Continuous, Order-Driven Market

An order-driven market is an ecology that comprises a variety of participants who interact in a variety of ways: some are buyers, and others are sellers. Some are seeking to trade because of new information, others because of their individual reassessments of share value, and others in response to their personal liquidity needs and cash flows. Some seek to trade by limit orders and others by market order. Some are longer term investors and others are shorter term traders. Some are proprietary traders and others are intermediaries. Some are large, institutional players and others are relatively small retail customers. And so forth and so on. The important point is, for the order-driven market to work efficiently, it must be in ecological balance. If it isn't, the order-driven market can collapse.

First and foremost, for a continuous order-driven market to be viable, it must receive sufficient order flow. If the order flow is inadequate, the possible gains from mean reversion will be insufficient to compensate a sufficient number of traders for placing limit orders (and, by so doing, accepting non-execution risk and adverse selection risk). Hence, the limit order book will be unduly thin. A sparse book and

[7] For further discussion, see Kalman Cohen, Steven Maier, Robert Scwartz, and David Whitcomb, "Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread," *The Journal of Political Economy*, April 1981, pp. 287–305.

a correspondingly wide bid-ask spread impose a cost that discourages the placement of market orders and this, in turn, lowers the probability of a limit order executing. A vicious cycle can develop that results in market failure. For this reason, an alternative to the continuous, order-driven market is generally turned to for smaller cap, less frequently traded issues. A more appropriate market structure for the thinner traded securities is the quote-driven, dealer market. Call auctions can also be profitably employed.

While transparency is an important feature of a continuous, limit order book market, large traders, to contain their market impact costs, seek opacity for their orders. It is clearly inappropriate for a large participant to submit a large block as a market order it would walk the book and, if large enough, could clear out the entire contra-side of the book. Neither would a large block be posted as a limit order transparency would be totally lost and, given its size, the probability of the large order executing completely would be relatively low. Consequently, blocks are not submitted as such to the continuous, limit order book market; rather, they are delivered in a succession of small tranches. The "slicing and dicing" takes time, however, and thus immediacy is not supplied in this market environment. Alternatively, the large orders are commonly submitted to an alternative trading system (ATS), many of which are referred to as *dark pools*.

Recently, technology development has brought to light one further complexity for the continuous market: the incredible speed with which orders can be submitted and turned into trades. Accompanying this speed is the ability to measure time with high-frequency precision. In today's markets, time is measured in sub-second intervals, down to nanoseconds and even microseconds.

Fast order submission, trade execution, and information dissemination are clearly desirable; hyper-fast, however, may not be. Because a bilateral trade is made any time that a buy order and a sell order meet or cross in price, speed is not simply desired in and of itself; in *continuous trading* with supersonic speed, getting to the market quickly is not per se important; getting to the market first is what matters. And it is the race to be first that magnifies the importance of speed in continuous market trading. When sub-second readings matter, the continuous market can become hyper-continuous.

No human can follow the quotes, trades, and prices as they evolve with subsection frequencies. Consequently, high-speed trading decisions are made by computers, and computer-to-computer trading can, at times, lead to some undesirable results (e.g., flash crashes that have been experienced in recent years). The cost of acquiring the technology required to achieve such *high-frequency trading* is enormous. In the HFT world, some participants gain advantages through, for instance, co-location and development of sophisticated trading algorithms.

In a horse race, a winner must if at all possible be declared and, with a nanosecond time clock, winning by a nose can do it. But trading is not simply a horse race. In trading, the sequence of order arrival within tiny, sub-second intervals is not attributable to meaningful, underlying information change, and it conveys little information of fundamental economic importance to other participants.

For the most part, the sequence of order arrival in very brief intervals is a matter of chance, of who has the better technology, and the vagaries of the order flow. Recognizing this, a lot can be said for not adhering to a microsecond time stamp. Alternatively, all orders within, for instance, a 1-s interval could be given the same time stamp, a stamp that identifies the second within which each order has arrived. Thus, multiple orders that arrive in the same second should be given the same time stamp, and may be executed in a single multilateral match with the use of a call auction algorithm to determine the trades and prices.[8] Further understanding this possibility requires knowledge of the call auction approach to trading, a market structure that we turn to in the next section of this chapter.

4.2.2 Call Auctions in Order-Driven Markets

An auction is a standardized procedure for handling and matching orders in a consolidated limit order book for the purpose of establishing a clearing price, the number of shares that will trade at that price, and the specific participants who will participate (and to what extent) in the multilateral transaction. The execution price is the value that maximizes the total number of shares that will trade.

The order book for the call can offer different degrees of transparency:

- Regarding quantity and quality, the book can display information ranging from displaying all orders including price, size, and trader name to simply showing indicated clearing prices but not volume or any other information.
- Regarding time, from seamless to minutes or a couple of minutes.
- Regarding addressee, the professional traders get the market information in real time, interested public parties postponed.
- Regarding data dissemination (which is an additional business for stock exchanges), a variety of combinations of what/when/to whom are in place; here segmentation is key, as well as, e.g., data streams for professional traders or *algo traders*.

The order book is built in the following steps:

1. Buy and sell orders are entered with price and time of entry recorded.
2. Limit orders on each side of the book are cumulated, from the highest price to the lowest for buy orders, and from the lowest price to the highest for sell orders.
3. Market orders are cumulated and included in their respective totals.
4. The cumulated buy orders are matched against the cumulated sell orders.

[8] For further discussion, see Robert Schwartz, "Slow Down, Wall Street," Commentary in *Traders Magazine*, July 2014, and Robert Schwartz and Liuren Wu, "Equity Trading in The Fast Lane: The Staccato Alternative," Invited Editorial, *Journal of Portfolio Management,* Volume 39, Issue 3 Spring 2013, pp. 3–6.

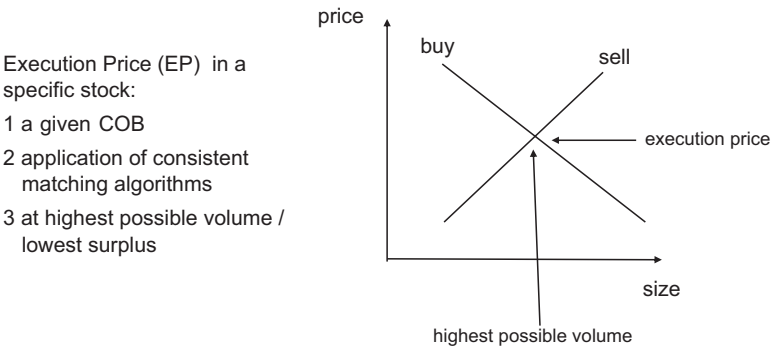

Fig. 4.7 Calculation of the execution price

With this preparation of the central order book, the matching algorithms can be applied to get the execution price (Fig. 4.7). The matching algorithms of an auction are an axiomatic system, whereby the overriding principles are the following:

- For one price without surplus (i.e., a clean cross): most possible executions, maximum turnover.
- For one price with surplus: maximum turnover *and* additional criterion.
- An additional criterion, like market pressure or most recent price, has to be introduced if two prices with the same surplus would be executable.

The algorithmic matching system must fulfil the following criteria:

- Consistency—Equal treatment
- Completeness—No loopholes or gaps in the procedure
- Simplicity—Least possible complexity

The first two bullets are necessary conditions, whereas the third is more of an criterion to improve effectiveness and efficiency.

4.2.2.1 Essentials

There are several necessary preconditions for a functioning call auction.

Fungibility means interchangeability: The shares of a listed company have all the same features and characteristics in content, form, and time. This is why buyers and sellers can trade at exchanges and clear through CCPs, which means netting and offsetting. In order to trade in an auction, the presence of buyers and sellers is mandatory besides a COB. Orders can be stored and deleted in a regulatory, compulsory way and, in applying the matching algorithms, orders get translated into trades. The whole call auction is embedded on the rules and regulations of an exchange (Fig. 4.8).[9]

[9] Cf. paragraph 4.2.2.4.

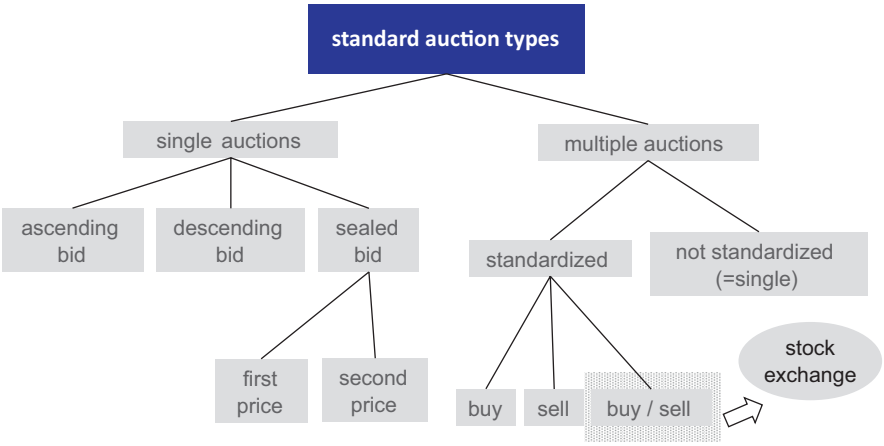

Fig. 4.8 Standard auction types

Auctions can be structured in different ways, as is illustrated in the diagram above. In a single (one-sided) auction, the highest bidder gets the object (e.g., a piece of art). In a descending auction, the first (lowest) bidder gets the object. If sealed bids are allowed, one alternative for the execution price is the first sealed bid; another alternative could be that the second sealed bid is the execution price.[10]

An execution price is determined by every auction, while the sequence of the repeated auctions determines the perfection of continuation (seamless). An auction can also cover the sell *and* the buy side simultaneously, as does an auction at the stock exchange. In contrast to a single auction, this is called a double auction.

4.2.2.2 Double Auction

With a double auction, bids and offers for a specific stock are matched to find the execution price. This price determination is achieved by applying a specific set of rules (matching algorithms).

A double auction can fulfil several functions:

- Open the market (opening auction)
- Reopen the market following a trading halt
- Close the market (closing auction)
- Mimic continuous price discovery by adding several multiple auctions in very short periods of time (even seconds are possible)

Figure 4.9 shows the situations that can occur before an *execution price* (EP) is determined:

[10] This is in accordance with "buy low/sell high," meaning for a single auction: in the first sealed bid it would be the one with the highest price, which is also true for the second sealed bid.

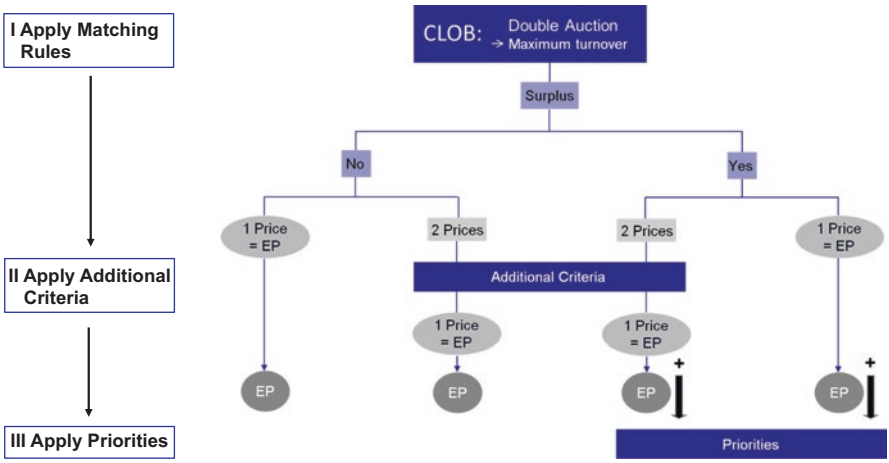

Fig. 4.9 Determination of the execution price in double auctions

If in a double-auction-price-discovery, the right side (*Yes*) applies, two possibilities can occur: either there is one price or there are two prices.

1. In the case of one price, this is the EP. And with an additional priority the handling of the surplus is defined. Some possible priorities are outlined in Sect. 4.5.
2. If two prices are possible, first the criteria to find the EP are applied, and afterwards the priority to handle the surplus.

A graphical outline of a double auction to open the market is shown in Fig. 4.10a, b:

1. Pre-trading: During the pre-trading phase, all incoming orders are collected in the COB. At the same time, orders may also be deleted. The COB is open, and the cumulated breadth and depth of the book can be seen in its entirety.
2. Auction (Deutsche Börse AG example): Once the call auction has started, the book is no longer transparent; there will be only a display of the indicative EP.[11] At a random end, the COB is frozen, so that no order may be entered or taken out of the book. Then the price determination is activated by applying the matching *algorithms*. At the end of the call the EP is defined and the market is cleared. The best (most aggressive) unexecuted orders that remain on the book set the *spread* which applies as continuous trading starts.

Through placing calls within the trading hours, the trading day gets a clear structure (Fig. 4.11).

[11] This is an optional feature.

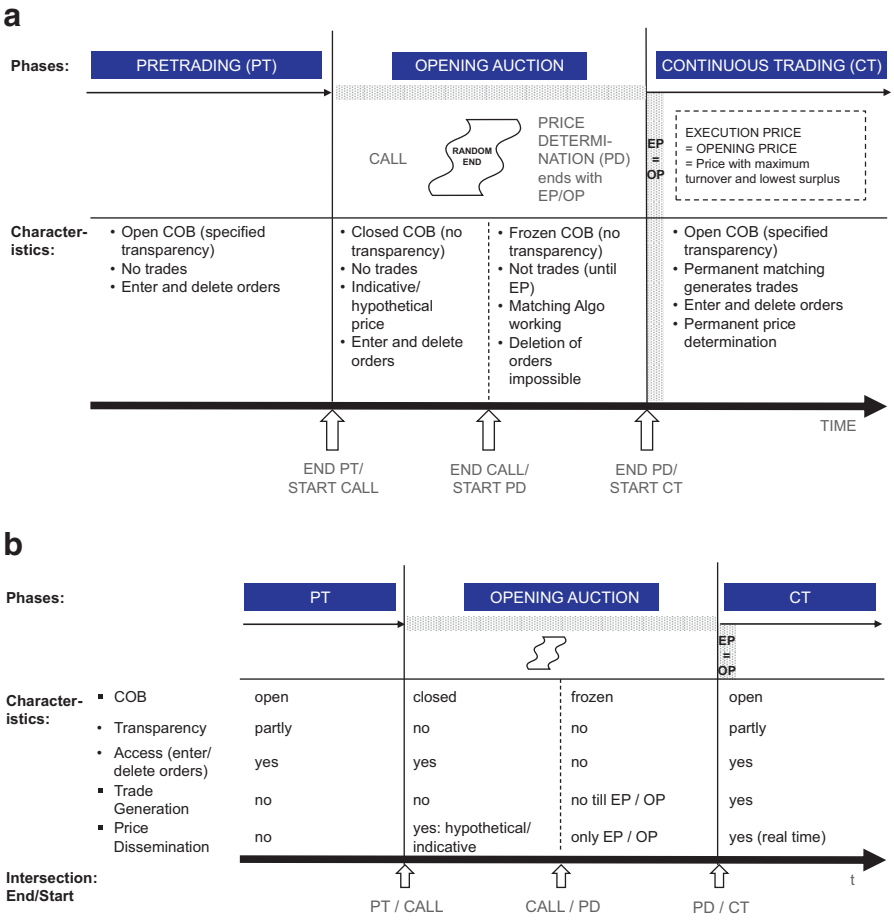

Fig. 4.10 (**a**) Opening the market through double auction. (**b**) Double auction characteristics

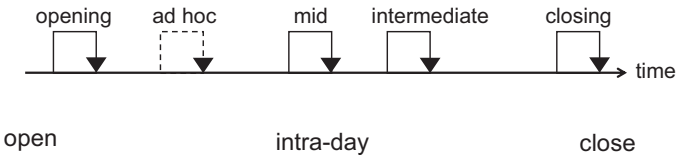

Fig. 4.11 Example structure of a trading day

4.2.2.3 Call Auction

A call auction is an order-driven market. Unlike the continuous order-driven market, in call auction trading orders that could otherwise be matched and executed (in bilateral trading) are held and cleared (in multilateral trading) at a single point in time, and at a single price. The matched and crossing orders which are executed at that single price include buy orders at the call price and higher, and sell orders at the call price and lower.

As noted, the clearing price at a call is determined by selecting the value which maximizes the number of shares that trade. This value is found by matching the cumulated buy orders (cumulating from the highest priced buys to the lowest) with the cumulated sell orders (cumulating from the lowest priced sells to the highest). Because order size and price are not continuous variables, a buy-sell imbalance (surplus) commonly exists at the market clearing price. Surpluses are typically handled by executing orders on the deeper side of the book according to the sequence in which they were transmitted to the market (i.e., by applying a first-in, first-out time priority rule).

During the first price determination in the process of getting an EP during a call auction, three phases have to be distinguished:

In the first phase, the COB has to be built and prepared to start the price determination. Then, in phase two, the actual determination of the EP takes place by applying matching algorithms. And eventually, in phase three, a possible surplus has to be handled.

As noted, if based on the application of the matching rules, two prices are possible, and an additional criterion is necessary to determine the execution price (Fig. 4.12). One of the following three criteria can be applied:

1. *Criterion of smallest surplus*: The objective of the smallest surplus criterion is to minimize the number of unexecuted orders. Therefore, in Fig. 4.13 the execution price (EP) is 99 because, at this price, the surplus is 500 shares while at 98.75 the surplus is 1000.
2. *Criterion of market pressure*: If, after the application of the two above mentioned criteria, two prices are still possible, the third criterion has to be resorted to. For instance, for the two possible prices 99 and 98.75:

   - The trading volume is 3000 shares *and at the same time.*
   - The minimum surplus is 1000 shares.

Because the surplus of the two possible prices is on the buy side, prices are driven up (if both surpluses were on the sell side, prices would be driven down): Therefore, 99 is the execution price (EP).

If the two equal surpluses at 500 shares were on the buy side (at 98.75) and the other one on the sell side (at 99), as shown in Fig. 4.14, the execution price (EP) is 99, because the reference price, usually the last paid price for the preceding day, is 99.50.

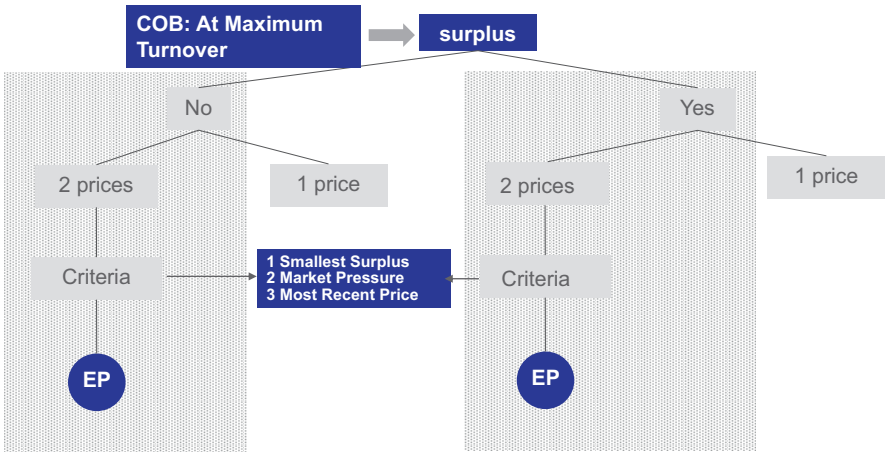

Fig. 4.12 Determination of the execution price in case of two prices

EXHIBIT 4.A2 Setting the Price: The Smallest Surplus Criterion

	Buy Orders, Number of Shares				Sell Orders, Number of Shares			
		Accumulated			Accumulated			
	Separate	Per Price	From Highest Price	Price (Limit)	From Lowest Price	Per Price	Separate	
	400 + 300	700	700	Market				
	....	....	....	> 100				
	200	200	900	100				
	300	300	1,200	99.75				Surplus
Surplus	400	400	1,600	99.50	4,700	500	200 + 300	
	200 + 300	500	2,100	99.25	4,200	700	700	
	800 + 100	900	3,000	99.00	3,500	500	500	500
1000	1,000	1,000	4,000	98.75	3,000	800	100 + 700	
	700 + 200	900	4,900	98.50	2,200	300	300	
				98.25	1,900	300	100 + 200	
				98.00	1,600	100	100	
				< 98	....	....	....	
				Market	1,500	1,500	700 + 800	

Fig. 4.13 The smallest surplus criterion

3. *Most recent price criterion*: In the unlikely case that two EPs with the same surplus but on different side of the market occur, the price that is closest to the last paid price—the reference price[12]—is selected. In the following example of Fig. 4.15, the reference price of 99.60 is closer to 99.0 than to 98.75.

[12] In the rules and regulations of exchange organizations, the reference price is usually the closing price of the previous trading day. The closing price is also used as a reference price for the derivatives market.

EXHIBIT 4.A3 Setting the Price: The Market Surplus Criterion

	Buy Orders, Number of Shares				Sell Orders, Number of Shares		
		Accumulated			Accumulated		
	Separate	Per Price	From Highest Price	Price (Limit)	From Lowest Price	Per Price	Separate
	400 + 300	700	700	Market			
	....	....	....	> 100			
	200	200	900	100			
Surplus	300	300	1,200	99.75			
↓	400	400	1,600	99.50	4,200	500	200 + 300
	200 + 300	500	2,100	99.25	3,700	700	700
1000	800 + 1,100	1,900	4,000	99.00	↑3,000	0	0
1000	0	0	↓4,000	98.75	3,000	300	100 + 200
	700 + 200	900	4,900	98.50	2,700	800	800
				98.25	1,900	300	100 + 200
				98.00	1,600	100	100
				< 98	....	....	....
				Market	1,500	1,500	700 + 800

Fig. 4.14 The smallest surplus criterion

EXHIBIT 4.A4 Setting the Price: The Most Recent Price Criterion

	Buy Orders, Number of Shares				Sell Orders, Number of Shares			
		Accumulated			Accumulated			
	Separate	Per Price	From Highest Price	Price (limit)	From Lowest Price	Per Price	Separate	
	400 + 300	700	700	Market				
	....	....	....	> 100				
	200	200	900	100				
	300	300	1,200	99.75				Surplus
Surplus	400	400	1,600	99.50*	4,700	500	200 + 300	↓
↓	200 + 300	500	2,100	99.25	4,200	700	700	
	800 + 100	900	3,000	99.00	↑3,500	500	500	500
500	500	500	↓3,500	98.75	3,000	800	100 + 700	
	700 + 200	900	4,400	98.50	2,200	300	300	
				98.25	1,900	300	100 + 200	
				98.00	1,600	100	100	
				< 98	....	....	....	
				Market	1,500	1,500	700 + 800	

*preveious price

Fig. 4.15 The most recent price criterion

If through application of the outlined criteria an execution price is determined and there is no surplus, all orders on both sides of the market are executed at the EP. If there is a surplus, it is calculated as the number of shares at the larger side minus the number of shares at the smaller side. The smaller side, which always executes completely, establishes the number of shares that trade, while shares on the larger size have to be rationed (Fig. 4.16).

Several rationing criteria are equally possible:

Time priority: Time is the most common secondary priority rule (with price being the first priority rule) (Fig. 4.17). The application is as follows:

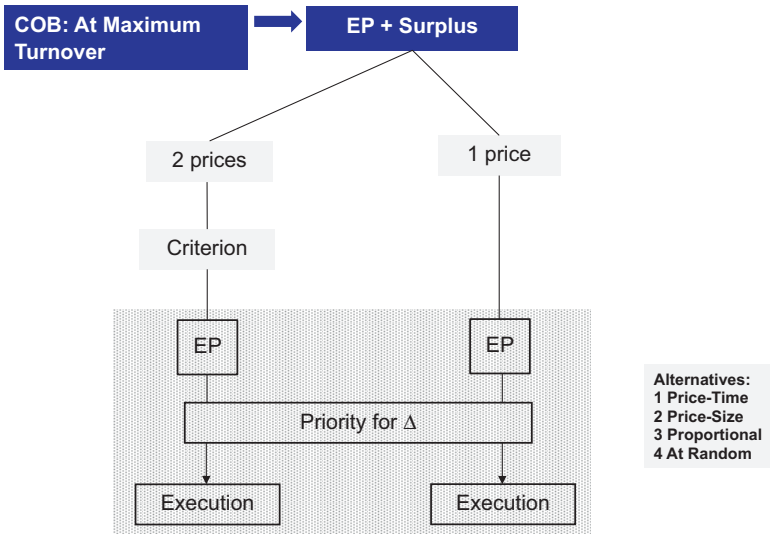

Fig. 4.16 Priorities on how to handle the surplus

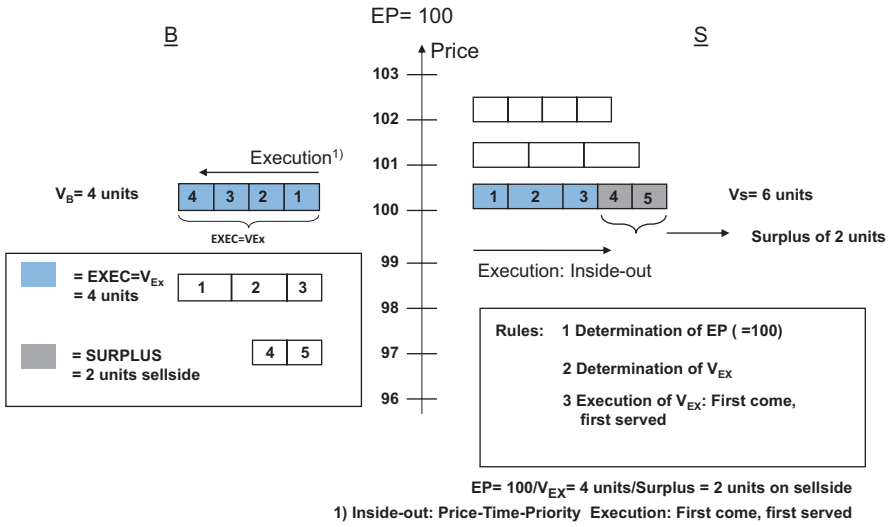

Fig. 4.17 Time priority

1. At the EP (= 100) arrange the orders based on their time stamp.
2. Then apply priority "first come first served/executed":

   - On the smaller buy side, all orders are executed: The executable volume at EP is 4 units/shares.

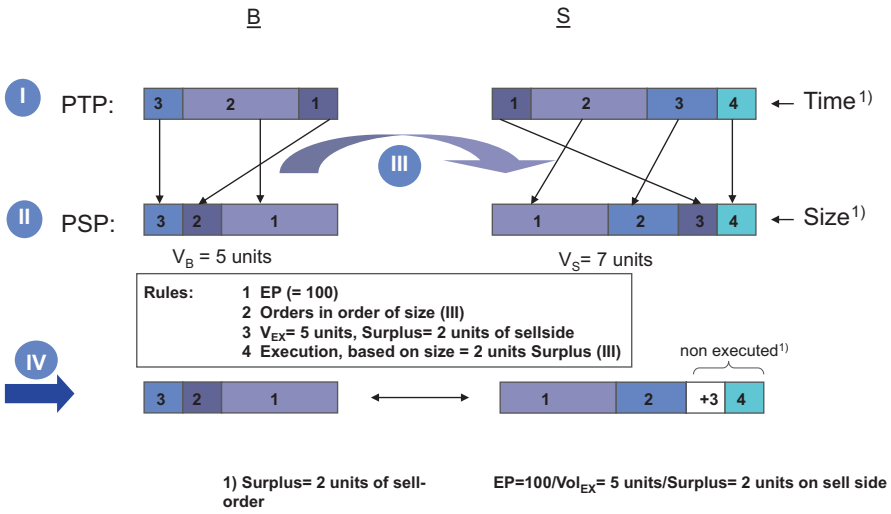

Fig. 4.18 Size priority

- On the bigger sell side, orders 1–3 are executed, and orders 4 and 5 (=2 units/shares) remain in the order book.

Size priority: If at the EP bigger orders receive priority, size priority is applied.

1. The COB is built as follows: two sided, applying price priority.
2. Within one price: size priority (bigger order size before smaller order size).
3. At the EP (=100) the executable volume is the smaller (buy) side with three orders adding up to five shares.
4. Execution:

- The whole buy side is executable: three orders adding up to five shares.
- On the sell side, 5 units—orders 1 and 2—are executable. Orders 3 and 4 are not executable and therefore build the surplus of 2 units or shares (Fig. 4.18).

Proportional-execution-priority: Proportional-execution-priority means that each order of the larger (surplus) side at the EP is executed proportionally to the smaller side (Fig. 4.19). Regarding the example illustrated above this means:

1. The proportion between the smaller (five units) and the bigger side (ten units) is 5:10 which equals 0.5 or 50 %. Therefore,
2. Every order on the bigger side is executed up to 50 %, which means half. The not executed part (five units) remains in the COB.

At random priority: Execution at random means that orders on the larger side are executed randomly until the sum of the executed order reaches the sum of the smaller side (=5 units) (Fig. 4.20). The unexecuted part (=3 units) remains in the COB.

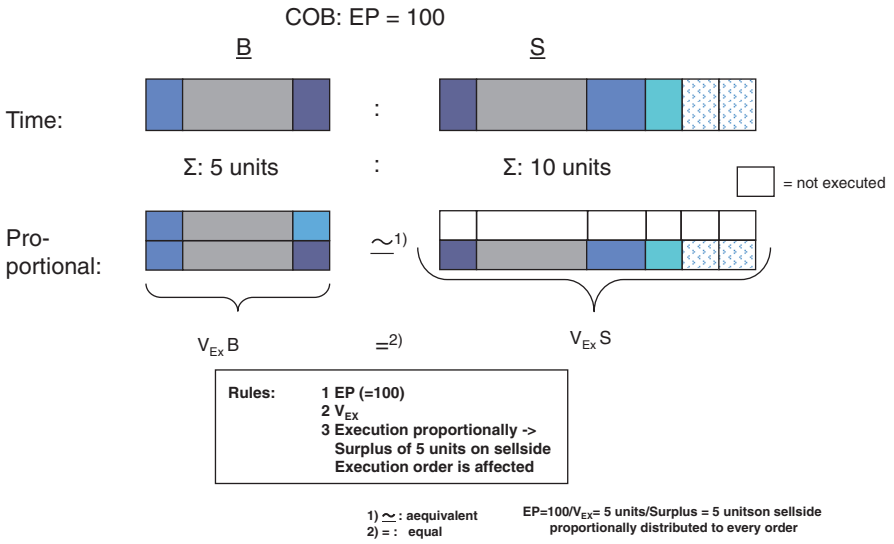

Fig. 4.19 Proportional execution priority

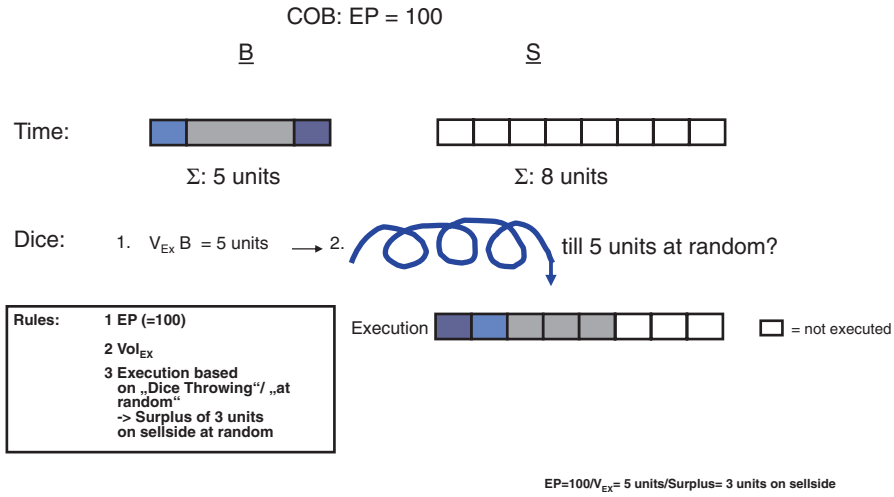

Fig. 4.20 At random priority

Figure 4.21 shows an illustrative limit order book for call auction trading.[13] The middle column displays the prices at which orders have been placed. The column on the left shows the cumulative number of round lot (100 shares) buy orders, cumulating down from the highest price at which each limit buy order has been

[13] Figure 4.21 is a screenshot from TraderEx, a computerized trading simulation software program. For further information, see www.etraderex.com.

TraderEx

DAY	TIME	SEED
1	9:37:00	10

TICKER	PRICE
	QTY
	TIME

MARKET	indicative	48.50
CALL	imbalance	-36

BIDS		OFFERS
93	49.40	1024
136	49.30	868
136	49.20	771
136	49.10	671
204	49.00	662
281	48.90	603
281	48.80	548
281	48.70	497
337	48.60	497
491	48.50	455
602	48.40	422
683	48.30	360
804	48.20	264
858	48.10	264
907	48.00	222
1033	47.90	203
1153	47.80	168
1237	47.70	168
1285	47.60	115
1307	47.50	78

Fig. 4.21 Illustrative limit order book for call auction trading

placed, and the column on the right shows the cumulative number of round lot sell orders, cumulating up from the lowest price at which each limit sell order has been placed. Given the displayed array of cumulated buy and sell quantities, 48.50 is the price that maximizes the number of round lots that would trade. At this price, a buy-sell imbalance (surplus) exists: cumulative bids (491) are greater than cumulative offers (455), and the number of round lots that can execute (being the lesser of these two values) is 455. This buy-side imbalance is handled by rationing the buy orders as we have just discussed (the criteria include time priority, size priority, proportional allocation, and random selection).

Note that no other price results in a number of executable round lots greater than 455. One price tick higher, at 48.60, the minimum of the cumulated bids and offers is 337 (resulting in a sell imbalance) and, one price tick lower, at 48.40, the minimum of the cumulated bids and offers is 422 (resulting in a buy imbalance). Thus a price of 48.50 maximizes the number of shares that trade and, accordingly, 48.50 is the clearing price.

Because limit orders submitted to a call execute at the clearing price established for the call, limit orders are price improved (with the exception of those placed at the clearing price exactly). This contrasts with continuous market trading where limit orders on the book execute at the price at which they have been entered. Because limit orders submitted to a call are commonly price improved, they should be priced more aggressively (i.e., higher priced buy limits and lower priced sell limits) than limit orders submitted to a continuous market. Another difference is the muted distinction between market orders and limit orders: market orders submitted to a call auction are nothing other than infinitely aggressively priced limit orders (infinitely high prices for buy orders, and zero prices for sell orders). Moreover, in contrast to continuous market trading, in call auction trading market orders do not execute with immediacy but only when the market is called.

Several advantages attend call auction trading. Batching orders together for point-in-time trading consolidates liquidity temporally. Systematically finding the clearing price with reference to the full set of cumulated buy and sell orders sharpens price discovery. Vis-à-vis continuous market trading, the order batching, single price auction procedure is fairer, and more difficult to manipulate. Recognizing these advantages, one might anticipate that call auctions would be widely used as a trading modality.

Call auctions were prevalent in the early days of trading, but nonelectronic calls had severe shortcomings and, as volumes increased in the precomputer age, the auctions were replaced by continuous trading. However, around the turn of the twenty-first century, calls started to reemerge in markets around the world. They have done so as modern, electronic facilities that are typically being used to open and to close trading in a hybrid combination with continuous trading. As we have noted, call auctions are also used to reopen markets after trading halts.

Uniting call and continuous trading eliminates one disadvantage of a call auction-only model: a participant need not wait for a market to be called in order to trade. The application of computer technology eliminates a second disadvantage of a nonelectronic call: investors can participate in an electronic auction in real time without being physically located on an exchange's trading floor.

When it comes to designing a call auction, a considerable number of alternatives exist. An auction can be totally opaque (closed book) or completely transparent (open book), or it can reveal only partial information about booked orders and an indicated opening price. A secondary trading priority rule (most prevalently time priority) can be applied to the order imbalance at a clearing price only, or to all executable orders on the deeper side of the market. The precise time when a market is called is generally determined by random draw within a prespecified, brief trading interval preceding a preannounced time (e.g., at the open or the close of a trading day). Calls can also be initiated at the request of a participant. A call can accept unpriced market orders, or it can be required that all orders be priced. Call auctions are generally price discovery facilities, but a variant exists: a *crossing network* matches customer buy and sell orders at an exogenously determined price (the midpoint of the bid-ask spread in a concurrently running continuous market, or at the closing price in the continuous market for after-hours trading.

This list of design alternatives can be extended. The important point is that not all calls are alike. As one might expect, some call designs will operate more efficiently than others. Much care must be taken to structure a call properly and, as is always the case with system design, one should recognize that the devil is in the details.

4.2.3 Market Making

4.2.3.1 Quote-Driven Market vs. Order-Driven Market

Order-driven markets consolidate liquidity in a single space—the order book. In the order book, limit orders and market orders representing bids and asks are placed and rules determine how trades occur. Basically, all orders are treated equally. Typically, only the type of order (limit or market), the limit order price, the time of order placement, and the order size (number of shares) matter. All traders can trade with each other in the same way, and there are no specific roles defined or incentives given to perform certain actions. As we have described, liquidity is gathered by limit orders submitted to the order book, before these orders grant other orders the option for an immediate execution.

However, liquidity provision and options to trade might be low in certain market conditions, especially for less frequently traded stocks. In these cases, it can be difficult to sustain continuous trading, and additional sources of liquidity will be necessary.

Market makers as a specific type of intermediary fill this role in many of today's markets. Their role is to provide two-sided markets, which means that they are mandated to continuously post bid and ask quotes to the market, and thus give other market participants the possibility to trade. Those quotes must be good for a minimum size and a maximum spread (the difference between the price of the ask and the price of the bid).

In quote-driven markets, the market is split into liquidity providers and liquidity takers. That split is a main difference between quote-driven markets and order-driven markets. Typically, multiple market makers operate simultaneously as competitors in providing their services to liquidity takers in a marketplace.

All bids (and offers) provided by market makers give other market participants the possibility to sell (and to buy). A market maker's quotes are options to buy or to sell. Liquidity takers cannot trade with each other; they are pure liquidity takers. They have to trade with a market maker.

A trade occurs if and when a market participant chooses one bid or ask of a market maker to trade with. By hitting the bid, or taking the offer, the constituent parts of the trade are determined (price, volume, and the two market participants).

In most dealer-driven markets, there are no secondary priority rules of order execution. Traders can choose the market maker they want to trade with. They can direct their orders to specific dealers, a practice known as "preferencing." Here we

see a further difference with order-driven markets which normally do not allow that type of practice.

Multiple market makers can be present in a marketplace. They compete with each other in the provision of liquidity.

4.2.3.2 A Market Maker's Role in an Order-Driven Market

The role of liquidity provision through a market maker can also be attached to an order book in a hybrid trading system. The order book works as described in the section on order-driven trading. In that case, in addition, market makers have the obligation to provide liquidity as in a quote-driven setup. Different from a pure quote-driven setup, they send their quotes into the order book and compete with limit orders in the order book. In this setup, all market participants interact via the limit order book in an equal way. The exclusivity of liquidity provision of market makers is broken up, and market makers' quotes become subject to the matching rules of the order book.

To compensate market makers for conducting their role, they in return receive certain benefits as an incentive. These can be discounts on trading fees or even a suspension from all charges of trading and post-trade clearing. Also, anonymity which is common in today's markets can be abandoned for market makers. Consequently they, and only they, can see with whom they trade. That privilege is supposed to help market makers identify so-called informed traders and thus reduce the market maker's risk of trading with these counterparts.

A further release from a strict quotation requirement is sometimes granted to market makers if they are obliged to provide a quote on request only, and not place it more permanently on the order book. Market participants can request a quote, and market makers must respond by sending the quote into the order book within a defined span of time.

An example of a market maker linked to an order book is the *designated sponsor* on Deutsche Börse's Xetra. The designated sponsor has the obligation to provide a quote, on a constant basis, into the order book of some stocks with low or medium liquidity. In addition, a quote request can be sent and the designated sponsor sees the name of the requestor. Furthermore, discounts on fees are granted to designated sponsors by the market operator.

4.2.3.3 A Market Maker's Role in Low- and Mid-Cap Stocks

Market makers like Deutsche Börse's designated sponsors contract with an issuer to provide their services to the market. The market maker is compensated by the issuer for providing liquidity in his or her stock. The market maker conducts research on this stock and provides analysis to the market. Deutsche Börse measures the performance of the designated sponsors in a stock and publishes performance figures on a regular basis. This information gathering provides

important guidance in the process of liquidity provision. Market maker revenues (spread and short-term trading in a mean reverting environment).

Market makers are compensated from two sources for providing their services (cf. "The Equity Trader Course," pp. 243–251): the bid-ask spread and trading the order flow. Market maker trades in a quote-driven market are typically "net trades"; namely a commission is not paid. A market maker realizes the bid-ask spread by buying low and selling high. Competition among market makers leads to a tightening of the spread. Wider spreads increase the market maker's profits, while competition resulting in tighter spreads reduces profits. Market makers with a "*long position*" profit when prices rise, and market makers with "*short positions*" profit when prices fall because they can cover their positions at lower prices.

Market makers need to manage their inventory. By adjusting a quote downward, a market maker attracts buyers who react to his or her aggressive ask. Consequently, inventory goes down. Vice versa, if the quote is raised, the market maker attracts sellers who react to the more attractive posted bid, and the market maker's inventory of shares goes up. To manage inventory, market makers can also trade with each other; this is called "*interdealer trading*."

Revenues may also arise for a market maker when successfully "trading the order flow" (cf. Equity Trader Course, pp. 243–251). If a dealer has a good sense of where the market is going short-term, he or she can profit from this insight. To do so requires the ability to detect trends and mean reverting behavior in the market. Timing is of the essence. A market maker profits when knowing when, on net, to buy or, on net, to sell.

4.2.3.4 Market Maker Costs (Costs of an Unbalanced Inventory and Asymmetric Information)[14]

In both types of markets, the order-driven as well as the quote-driven market, the natural buyers and sellers remain the ultimate source of liquidity. The "naturals" generally seek to hold positions in a portfolio for a longer time. Market makers seek to hold inventories (long or short) on a short time base only. They buy not for their own investment purposes, but to grant others the option to buy or to sell immediately. In so doing, they accept the risk of carrying an undiversified portfolio. Market activity (be it preferencing, volatility of prices, infrequent order flow, and stochastic nature of order flow) makes running an inventory more difficult and costly, and thus increases the spreads a market maker is willing to post.

Market makers, like any other traders, expect to incur losses from trading with better informed market participants. For example: a market maker buying stocks from an informed trader coming in before the stock's price is about to fall will lose from that trade. Market makers are compensated for that loss when

[14] The Equity Trader Course, pp. 248–251.

trading with "liquidity" traders (sometimes also referred to as "uninformed" traders). The volume of dealing with them must be large enough to compensate for trades that a dealer is making with better informed traders. That "ecology" of a quote-driven market is necessary for dealers to stay in business, i.e., for the market to exist.

4.2.3.5 Market Makers as Liquidity Providers[15]

Liquidity provision is the main role of market makers. As noted above, the role involves offsetting temporary imbalances between buyers and sellers (demand and supply) in the market.

In that sense, market making is the immediate provision of liquidity. The market maker is permanently present in the market, supporting liquidity provision on a continuous basis. That concept may be extended to a periodic service when it is linked to a periodic type of trading like the call auction. Market makers attached to a continuous order-driven trading market can also be required to provide liquidity (a quote) to call auction trading. The market maker's presence during the entire call phase in the auction may be required in such a setup.

Whether acting as the single source of liquidity in a "pure" quote-driven environment, or acting in combination with an order-driven format, "hybrid" market makers represent a flexible solution to providing liquidity.

4.2.3.6 Market Makers as Facilitators[16]

Trading only occurs when buy and sell orders meet in both space and time. We have described how market makers temporarily step in when there is an order flow imbalance. They do so by providing two-sided liquidity. Their role can go even further. Their activity may trigger orders which "are not yet" displayed to the market. An active trader may attract more liquidity to come to the market, even in a way that triggers a "burst of trading."

4.3 Functions of Market Models (a Designer's Perspective)

Viewed abstractly, trading is a process of information transformation that produces transactions. The carriers of information are the orders that meet in a market, along with the dialogue conducted by traders. The place for this information exchange is the trading system which comprises either a trading floor where participants meet

[15] The Equity Trader Course, p. 240.

[16] The Equity Trader Course, pp. 240–241, Animation.

face to face, or an electronic system where they meet virtually. The previous section has shown the importance and complexities of price determination, but the challenges of finding a good market structure extend beyond the construction of a robust price discovery mechanism.

In addition to price and quantity, trades comprise information about who trades, the type of asset traded, when and where the trade has taken place, and information about how the trade is settled (i.e., the modalities of the post-trading phase). A chosen market model prescribes how this information is generated from information that has been received. In this sense, a market model is the definition of a function. If we take today's electronic trading systems that have been applied in many market structures, a defined function is implemented through an algorithm that makes an outcome deterministic (i.e., the results are always the same when all of the inputs into the algorithm are identical).

In the text that follows, we briefly discuss the diverse functions that various market models define.

4.3.1 Determination of When a Trade Occurs

For a trade to be triggered, certain conditions must be fulfilled. The market model defines these conditions. Buyers and sellers must be in agreement on all conditions of a trade. Achieving this can be the result of a negotiation process among two market participants or, for instance, the result of placing orders in an order book at an exchange. The triggering of one or multiple trades may then occur ad hoc or at prespecified points in time.

In continuous trading, a trade occurs whenever two orders match. Consequently, a mechanism must be in place that constantly checks for a situation in the order book that will allow this to happen. Every new order that reaches the order book is tested to determine whether such an order exists on the other side of the market.

Periodic call auction trading demands less effort than order book trading in a continuous market. This is because call auctions are inherently less complex. Call auctions are typically triggered whenever a certain, predetermined point in time has been reached (e.g., the opening of trading in the morning, the closing of trading in the evening, or at midday). Accordingly, auctions do not require a constant check of the market. Even triggers for volatility auctions as described previously do not come from the auction market itself; they arise in continuous trading, but only when a predetermined volatility condition has occurred.

In electronic trading systems, two types of triggers can be calculated, one depending on the order situation, and the other depending on time. Alternatives exist. In bilateral, negotiation markets, for instance, a trade occurs whenever one party to the negotiation accepts an offer that has been placed by a counterparty.

4.3.2 Determination of the Location of a Trade Occurs

Market models can be hybrids. In such situations, a combination of market models exists, and they may interact. Market models may be combined sequentially. For instance, in many markets trading starts with an auction at the opening that is followed by continuous trading, which is followed by an auction at midday, which is itself followed by continuous trading, and that then closes with a call auction in the evening. Depending on the order specification and market conditions, the order book that is eligible for the trade is determined.

4.3.3 Determination of the Counterparts of a Trade

Counterparty determination depends on how many parties interact with each other at the same time to find a trade. On one end of the spectrum (bilateral negotiations), counterparty determination can be relatively easy. However, searching for and selecting a party to start a negotiation might be costly. Negotiation starts when two parties enter a process of finding the details of a trade, including price, quantity, and post-trade modalities. That process ends successfully after the passage of some time, or it terminates without any result.

The market model is more complex when many parties interact simultaneously in the same place, and priority rules are imposed on all of them. If there are multiple buyers and sellers at the same time with orders in the market and their orders are all eligible to trade, priority rules are required to specify who gets to trade first. As such, the rules determine who trades with whom. The orders that are submitted to the market must carry certain requisite information. In most cases, price is the primary criterion used (the most aggressive orders trade first). Price priority (the primary rule) is typically followed by time priority (a secondary priority rule), and this requires that each order be time stamped when it enters the market. Of course, this in turn requires the mechanism of a clock that imposes a sequence on all orders coming in. Ideally that clock is a central mechanism positioned at the "gate" that all entering orders have to pass through.

The continuous trading and periodic call auction market models illustrate the difference between sequential, multiple-price, and bilateral matching type of trading on the one hand and simultaneous, single-price, and multilateral matching type on the other.

4.3.4 Determination of the Price of a Trade

The complexities and challenges of price determination are addressed in the previous section of this chapter. Notably, some market models exist that do not comprise price determination but which still lead to trades. Venues that follow

such a model base trades on prices that are taken from other markets. The market model in these situations must define where to take a price from, and what specifically the reference price is; this is required for the production of either bilateral or multilateral trades.

Compared to a market model that is designed to produce prices based on the information that has been received and the rules that must be followed, building and operating a market model that comprises a reference pricing principle are clearly less complex.

4.3.5 Determination of the Quantity Traded

Price determination and quantity determination are two closely related functions that market models define. In those market models which include price determination, the quantity (i.e., volume of equities) that is offered (sellers) or sought for (buyers) depends on the price which buyers are willing to pay and sellers are demanding.

Both values—price and volume of transactions—in these cases are determined simultaneously. In automated order-driven models of trading those two functions are conducted algorithmically.

Market models which don't include price determination and use reference prices comprise a function for the matching of volumes to buy and to sell either periodically or continuously. Volumes in the first case are matched over a certain period of time and then executed with a quantity equal to the lower of the two quantities (buy quantity, sell quantity) available at that time. In case of a continuous matching a newly incoming order's volume is checked for execution against already "waiting" volume on the other side of the market or is going to wait until matching volume is available.

4.3.6 Determination of What Is Traded

Market models differ with respect to the standards they set for the assets to be traded. In highly standardized markets, all parameters that define an asset are preset and agreed upon. Traders accept and commit to these standards when they enter the marketplace. Security exchanges are of this type: their products are highly standardized and homogenous. A different situation exists when a market structure does not define, ex ante, the specifics of the assets to be traded. In these situations, market participants must agree on these qualities in the course of agreeing on a transaction, and the information exchanged must be specified accordingly.

4.3.7 Determination of Post-trade Modalities

Transactions that have been agreed to by market participants must be fulfilled, and the time and place for this to be done must be agreed upon. In exchange-type markets, that point is prespecified; typically, it is the national *Central Securities Depository* (cf. Chap. 6) where settlement takes place. Today, in most cases, that comprises the simultaneous, irrevocable, and final transfer of assets between the participants in the trade (typically money against securities). In cases where market structure does not specify that procedure, market participants must agree about how they want their obligations stated and fulfilled. That comprises finding a common place to which both parties can send their instructions for delivery and payment (i.e., the settlement of their trade).

The post-trade structures of exchanges standardly comprise the involvement of a central counterparty (cf. Chap. 5). For off-exchange transactions, regulators may demand the involvement of a CCP that offers clear standards to deal with consummated transactions.

4.3.8 Provision of Market Transparency Pre- and Post-trade

Market models specify the dissemination (or lack thereof) of information that reflects the trading intentions and orders of buyers and sellers. A market model transforms both the stream of incoming orders and the sequence of trades produced into information streams that reflect both orders and trades. The degree and timeliness of information available about pre-trading intentions are referred to as *pre-trade transparency*. The degree and timeliness of information available about trades that have occurred are referred to as *post-trade transparency*. Continuous trading in a so-called open limit order book fully provides both pre- and post-trade transparency. Information concerning volumes offered and requested at all price steps in the market is visible in real time to all market participants and to the public. The information about trades (i.e., price, the volume traded, and the exact time of the transaction) is published in real time as well. The term “lit-trading” is used for venues that, in this regard, are transparent. The term “dark-trading” is used by market participants who avoid (for various reasons) the publicity of lit-trading. Large traders, in particular, do not wish to have information about own trading intentions and completed transactions conveyed to other participants.