Report on the Results of Load Testing of the Moscow Exchange Trading and Clearing Systems on August 1, 2026

Table of contents

Testing objectives 2

Key results 2

- Moscow Exchange Securities Market Trading/clearing system 2
- Moscow Exchange FX Market Trading/clearing system 3
- Testing the FX Market TCS version with trading engine hot standby 4
- Moscow Exchange Derivatives Market Trading/clearing systems 5
- Testing the hot standby mechanisms for Spectra TCS core modules 6
- Transaction latency, Spectra TCS, TWIME, FIFO TWIME, SIMBA, and FAST UDP multicast market data servers 7
- Spectra TCS, CGate/Plaza2 data dissemination infrastructure 10
- Exchange Network and Colocation Zone Network 11

Conclusions 11

- The Securities and FX Markets 11
- Derivatives Market 11

Comparison of load testing parameters and peak load/data volume metrics in production environment 12

Testing objectives

To verify the operational capability of the Moscow Exchange trading and clearing systems (TCS) under increased load and high numbers of orders and trades. Trading was conducted in the systems of the following markets with the following target numbers of orders and trades per session:

- Securities Market – 400 million orders and over 20 million trades;
- FX Market – 110 million orders and over 3 million trades;
- Derivatives Market – 300 million orders and over 7 million trades.

Key results

Moscow Exchange Securities Market Trading/clearing system

Testing was performed on the production version and configuration of the Securities Market TCS software, with separate servers for trading and clearing operations.

The testing priority was to verify that the TCS information capacity meets the forecasted growth in the numbers of orders and trades for the coming years.

Achieved information capacity and performance parameters compared with the maximum numbers of orders and trades in production environment and in the 2025 load testing are presented in the table.

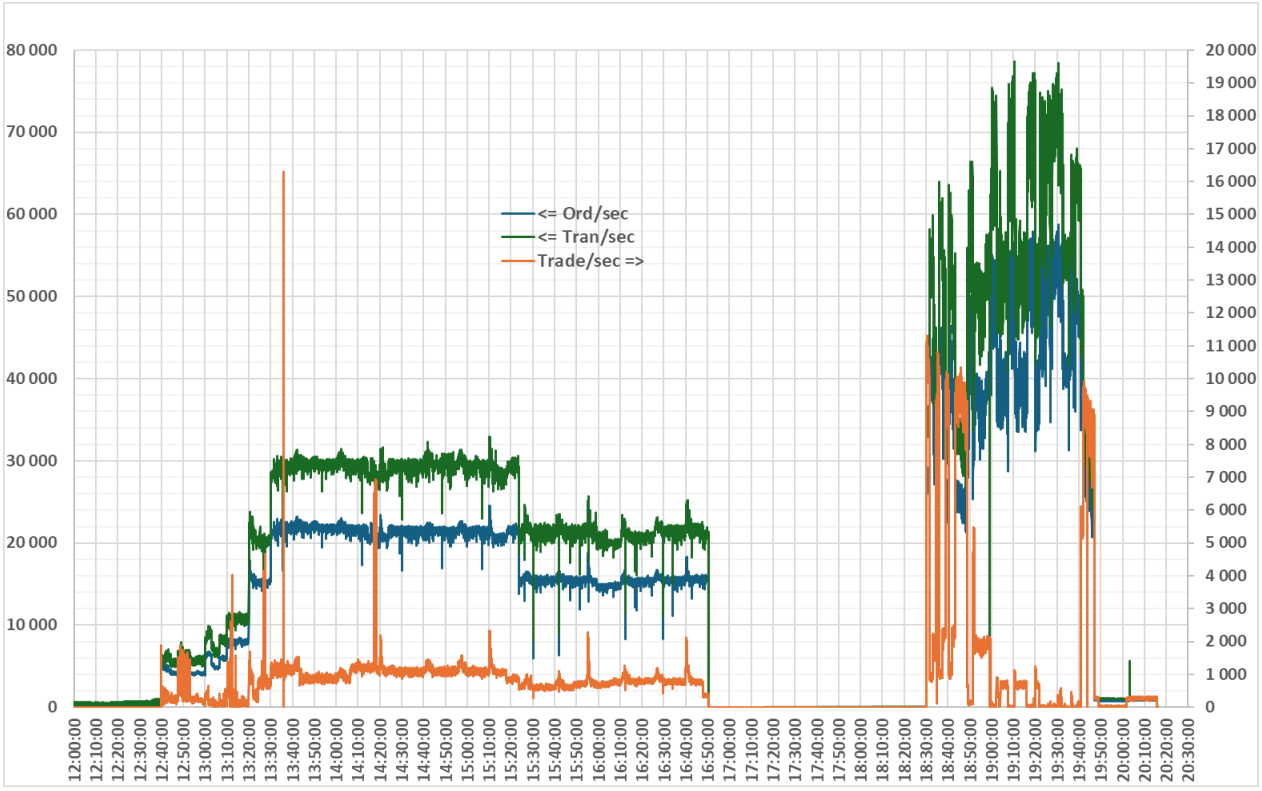
Accepted transactions are defined as incoming orders resulting in the registration of a new order or the successful cancellation of a previously registered order.

	Transactions	**Orders**	**Trades**
Achieved values (units), 2026	591,826,215	439,290,571	23,867,186
Achieved values (units), 2025	548,599,105	416,441,099	21,534,788
Maximum values in production environment	241,883,032	179,411,854	10,664,761
Maximum processing rate of accepted transactions (units/sec), 2026	78,657	58,781	16,293
Maximum processing rate of accepted transactions (units/sec), 2025	38,022	35,795	3,929

Peak frequencies of the continuous transaction, order, and trade flows during testing exceeded the peak values in one-second intervals observed in production environment by more than 2 times.

Before the emergency shutdown, the Securities Market access servers and FIX, FAST, SIMBA, and FIFO TWIME ASTS services operated normally, with no issues identified.

Graphs of transaction and trade frequencies are provided in the figure:



Clients generated 4% of the transactions, three times the share in 2025.

At 15:50, the Securities Market TCS access servers stopped because their states diverged from the state of the trading engine.

The divergence was caused by a code defect triggered by unusual activity by one of the testing participants. The defect was fixed and TCS operation was restored by 18:15. Trading resumed at 18:30.

Due to errors in the access server restart configuration, the FIX/FAST/SIMBA/TWIME services did not resume operation.

Moscow Exchange FX Market Trading/clearing system

Testing was performed on the production version and configuration of the FX Market TCS software, with separate servers for trading and clearing operations.

Achieved information capacity and performance parameters compared with the maximum numbers of orders and trades in production environment and in the 2025 load testing are presented in the table.

Accepted transactions are defined as incoming orders resulting in the registration of a new order or the successful cancellation of a previously registered order.

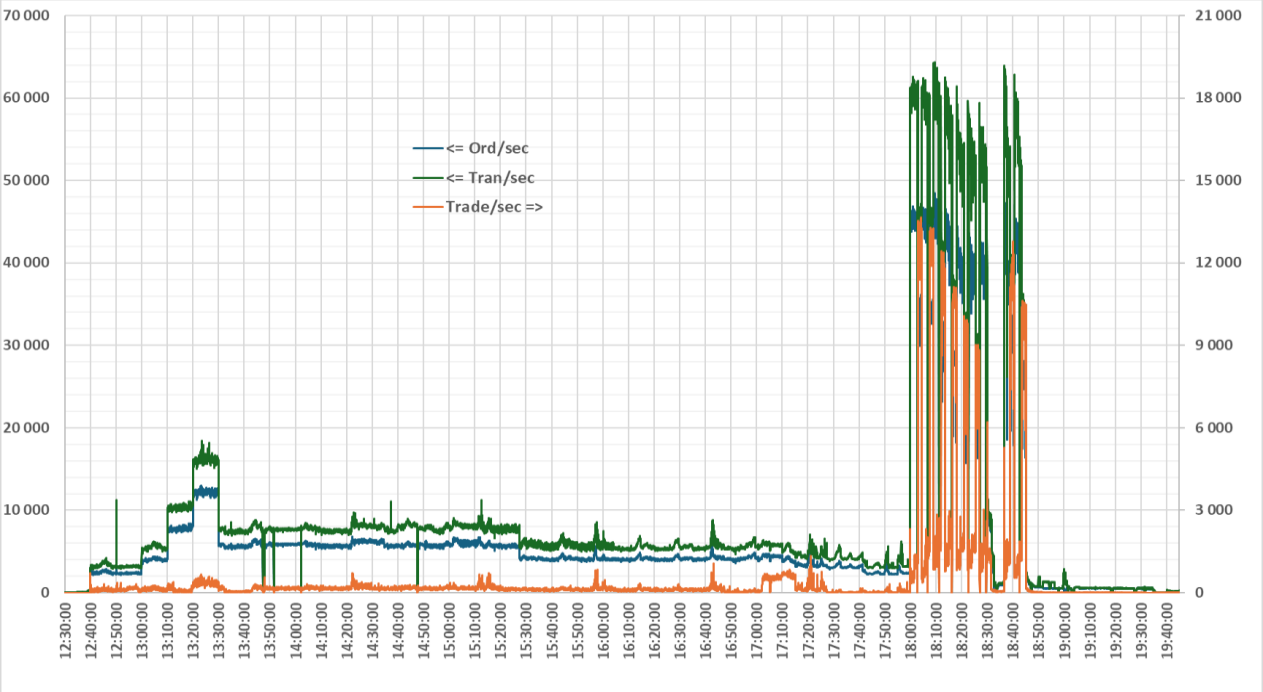
	Transactions	**Orders**	**Trades**
Achieved values (units), 2026	245,184,246	183,288,745	12,976,163
Achieved values (units), 2025	183,237,258	118,396,134	3,601,911
Maximum values in production environment	120,000,000	85,000,000	1,500,000
Maximum processing rate of accepted transactions (units/sec), 2026	64,389	48,427	13,654
Maximum processing rate of accepted transactions (units/sec), 2025	37,334	25,141	843

Peak frequencies and numbers of orders and trades were not reached during testing.

Peak frequencies of the continuous transaction, order, and trade flows during testing exceeded the peak values in one-second intervals observed in production environment by 2–3 times.

The FX Market access servers and FIX, FAST, SIMBA, and FIFO TWIME ASTS services operated normally, with no issues identified.

Graphs of transaction and trade frequencies are provided in the figure:



Clients generated 8.7% of the transactions, three times the share in 2025.

Testing the FX Market TCS version with trading engine hot standby

On August 2, 2026, additional testing was performed to verify failover to the standby trading engine by simulating a hardware failure of the primary trading engine. Representatives of trading members took part in the test.

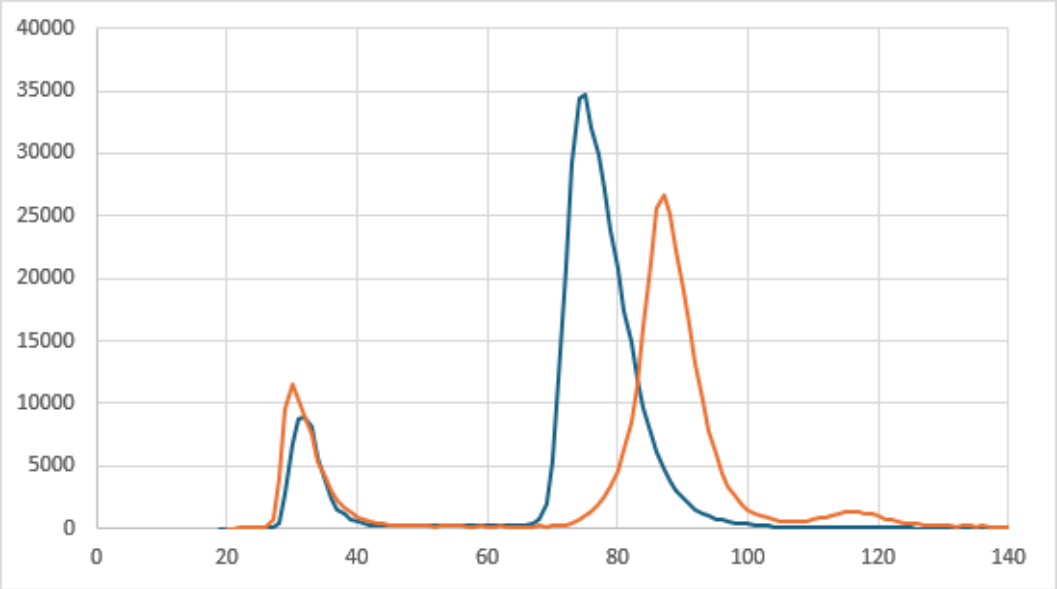
A forthcoming version of the FX Market TCS was installed in the production environment. All TCS modules participated. The algorithm used to generate trading activity was the same as in the August 1 load test.

The transaction flow was gradually increased to 20,000 per second. The primary trading engine was stopped at 13:00.

All client and load generator sessions successfully connected to the new trading engine within six seconds. For sessions sending a stream of orders to the TCS, the failover appeared as a wait of 4.5 to 6 seconds for the response to one transaction.

The TCS build used in this test had more aggressive optimization of the clearing subsystem code than the production build. Consequently, the response times and maximum throughput of the hot standby version were better than those of the production version.

The distributions below show the numbers of transactions by trading engine processing time, in microseconds. The times exclude message transit through the access servers and the network from the client and back. This component of response time does not change when hot standby is introduced. The blue plot shows the hot standby test; the orange plot shows the August 1 load test at the same transaction frequency of 10,000 per second.



After the failover demonstration, the TCS was restarted in hot standby mode and its throughput was measured. The maximum transaction frequency in the August 2 test was 89,000 per second, 38% higher than in the load test (64,300).

The same core code optimizations are planned for the next update of the FX and Securities Market TCS. The August 2 test values may therefore serve as a guide to the performance of the TCS after hot standby and these optimizations are introduced.

Moscow Exchange Derivatives Market Trading/clearing systems

Testing was conducted on SPECTRA version 9.6, put into production on June 22, 2026, on servers in the Data Space and Nord data centers. The TCS core modules (servers for pre-trade order validation, pre-trade collateral checks, and order matching) were switched to hot standby mode.

Testing focused on the following aspects:

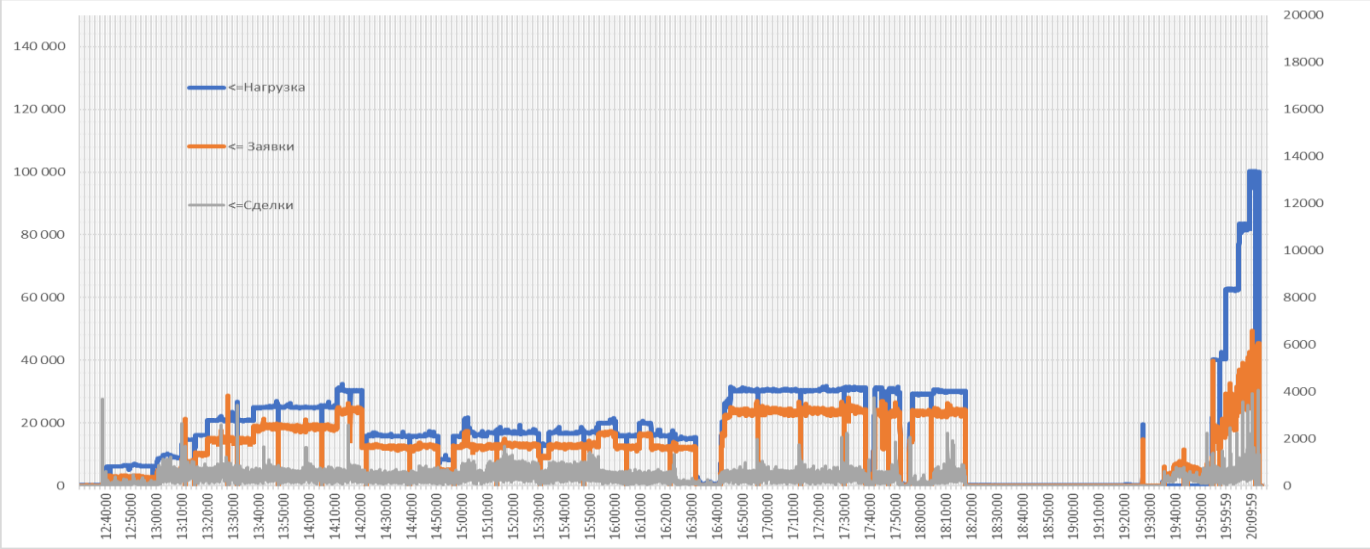
- Sustained operation of the TCS in the hot standby configuration:
    - Measuring latency and peak system throughput;
    - Simulating failures and testing failover of the hot standby TCS core modules, including measurement of failover time;
- Confirming the TCS capability for sustained operation under loads comparable to short-term peak loads in production environment;
- Verifying the operability of the entire TCS under forecasted 2028–2029 data volumes, taking into account changes in the order-to-trade ratio in production.

During testing, 334.4 million accepted orders were submitted and 9.4 million trades were executed. **The maximum transaction processing rate was 100,310 transactions per second.**

	Transactions	Orders	Trades
Achieved values (units), 2026	510,843,812	334,388,354	9,385,084
Achieved values (units), 2025	525,487,657	362,752,328	13,926,464
Peak performance (transactions/sec), 2026	100,310	49,375	4,073
Peak performance (transactions/sec), 2025	114,000	54,027	3,976
Performance in production environment (transactions/sec), 2025–2026	25 000		

Clients participating in the testing generated 1.1% of all transactions.

The transaction load graph for the Derivatives Market TCS is provided in the figure:



An evening clearing was performed during testing. At the start of the evening clearing, the data volumes were 8.3 million trades and 300.2 million submitted orders.

The evening clearing operations were completed in 50 minutes.

This time was in line with expectations, considering the increased volumes of reference data and transactions.

During testing, variable lag was observed between the CGate/Plaza2 data dissemination infrastructure and the system core. The issue was recorded from 13:36 to 16:50 Moscow time.

Testing the hot standby mechanisms for Spectra TCS core modules

During the period of increased load after 19:40, the leader processes in the hot standby cluster were terminated in succession using SIGKILL, and the transfer of the leader role to another cluster member was analyzed.

For all TCS core modules, election of a new leader and resumption of message processing in the TCS core network took 1.2–1.5 seconds, with no message loss.

When the module publishing the core transaction log to the Plaza2 network failed over, the Plaza2 infrastructure completed the switch in two minutes.

Transaction latency, Spectra TCS, TWIME, FIFO TWIME, SIMBA, and FAST UDP multicast market data servers

The Corvil system was used to measure Derivatives Market TCS latency by logging transaction flows from clients and exchange-side robots, together with a traffic analysis system and Grafana for graph visualization.

The TWIME reference robot was located on a server in the colocation zone and generated a load of 14 orders per second, simulating a typical client.



At transaction frequencies of 5,000–30,000 per second, median RTT for the TWIME gateway ranged from 165 to 223 microseconds.

In the previous load test on August 2, 2025, the RTT measured for a robot in the colocation zone ranged from 160 to 200 microseconds at comparable transaction frequencies.

In the August 1, 2026 tests, a FIFO TWIME reference robot was run to assess latency.

The FIFO TWIME reference robot was located on a server in the colocation fast links zone and generated a load of 14 orders per second.



At transaction frequencies of 5,000–30,000 per second, median RTT for the FIFO TWIME gateway ranged from 125 to 173 microseconds.

This was the first load test with a FIFO reference robot, so a comparison is possible only with production latency on the FIFO server, currently 70–100 microseconds.

For transactional connections in the regular (non-FIFO) zone, the 2026 tests showed:

- A negative effect on latency from the TCS core hot standby mechanisms;
- A positive effect on latency from changes to internal traffic routing and replacement of network load balancing tools.

For FIFO transactional connections, the TCS core hot standby mechanisms increased latency by 50–70 microseconds in the 2026 tests.

Please note that the software will be further developed before the planned introduction of hot standby for the TCS core modules in Q1 2027. These changes may affect latency.

The Derivatives Market SIMBA and FAST server configuration matched the production configuration. Normal operation was maintained throughout the entire range of market transaction frequencies.

Traffic was captured using Corvil equipment. A traffic analysis system and Grafana recorded and collected statistics on the publication times of TWIME order accepted messages (NewOrderSingleResponse) relative to new order messages from the TCS (Order Log feed) in the SIMBA protocol. The graphs below show the median difference “TWIME timestamp minus SIMBA timestamp”; a positive value means that SIMBA preceded the TWIME response by 37–50 microseconds on median.

Statistics of this type are collected continuously during normal trading.

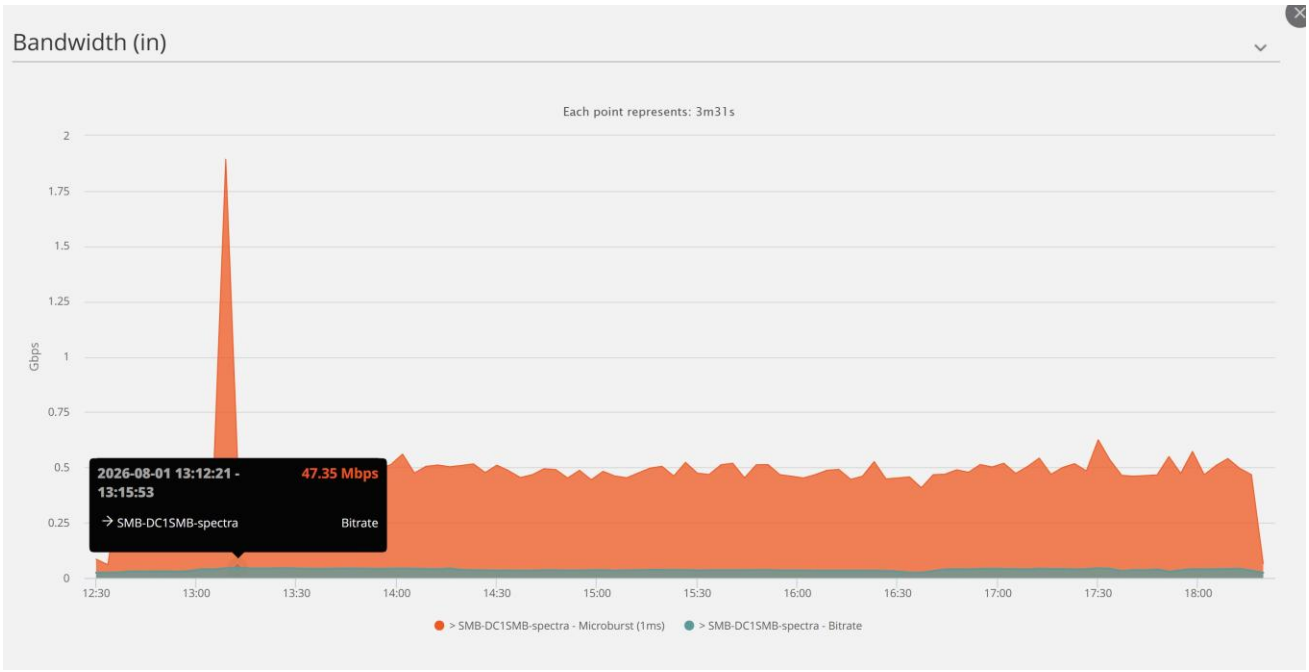
SIMBA vs. TWIME comparison:



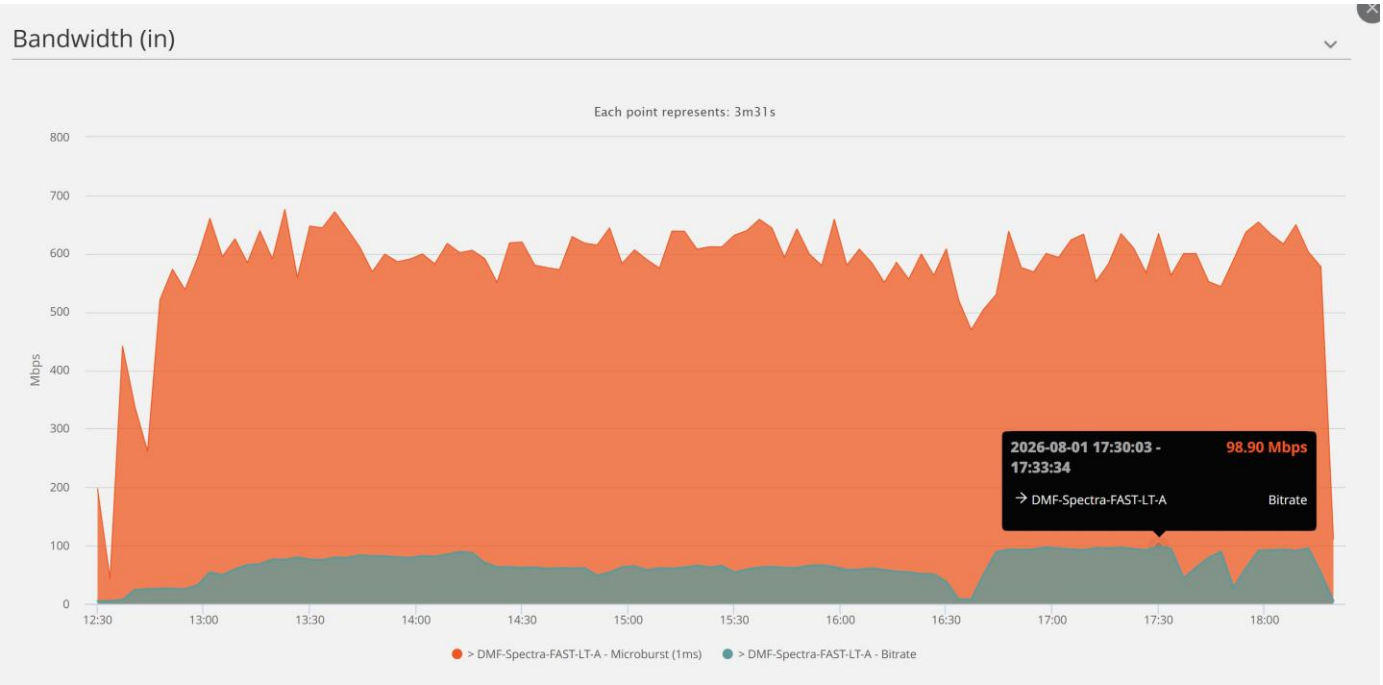
In 99% of cases, SIMBA market data for orders was sent at least 30 microseconds before the TWIME transaction response.

The median SIMBA publication time at 30,000 transactions per second was around 48 microseconds.

Graph of SIMBA UDP multicast traffic in Copy A and millisecond interval peaks:



Graph of FAST UDP multicast traffic in Copy A and millisecond interval peaks:



SIMBA UDP multicast traffic in the test reached 47.35 Mbps in each of Copies A and B, with peaks of up to 1.89 Gbps in millisecond intervals.

FAST UDP multicast traffic in the test reached 98.9 Mbps in each of Copies A and B, with peaks of up to 675 Mbps in millisecond intervals.

The bandwidth requirement for clients using the FAST service to receive ORDERS-LOG in real time is at least 100 Mbps per feed. A 1–10 Gbps channel is recommended when receiving both FEED A and FEED B or data from multiple markets.

Spectra TCS, CGate/Plaza2 data dissemination infrastructure

Two new components were used in the data dissemination network during testing:

- A module for publishing TCS core data to the Plaza2 network, modified to support TCS core hot standby;
- A function that smooths traffic peaks in periodically published auxiliary data feeds, including volatility data and indicative variation margin.

Two significant issues were identified:

1. During testing, variable lag was observed between the CGate/Plaza2 data dissemination infrastructure and the system core. The issue was recorded from 13:36 to 16:50 Moscow time and was caused by resource contention between TCS modules after hot standby was enabled.
2. Failover of Plaza2 data publishing took more than one minute when failures of certain core nodes were simulated.

Both issues will be resolved before the TCS hot standby release.

The function for smoothing peaks in auxiliary feed traffic proved useful in giving priority to trading traffic during surges in market activity and remained stable under increased load. It will be enabled in the production TCS configuration in the next release.

Exchange Network and Colocation Zone Network

Monitoring of the network within the exchange perimeter and in the colocation zone (up to and including the colocation zone core) showed no deviations from normal operation. No network retransmissions, packet loss, or deviations in packet delivery speed from the current load level in production environment were recorded.

Recommendations and requirements for the bandwidth of channels to the exchange are available at: https://www.moex.com/a1160.

In the dedicated network used to publish Securities and FX Market SIMBA data, cancellation of hundreds of thousands of active orders produced traffic peaks of up to 750 Mbps for each copy of the OLR feed. Software using these services should be designed with these traffic volumes in mind.

Conclusions

The Securities and FX Markets

1. Load testing confirmed that the FX and Securities Market TCS have a significant margin in terms of peak numbers of orders and trades relative to the maximum numbers in production environment.
2. The planned replacement of TCS servers in 2025 increased TCS throughput and information capacity.
3. On August 2, the hot standby mechanism was successfully tested on the production FX Market TCS infrastructure with representatives of trading members. The response times and maximum performance measured were better than those of the production TCS version.

Derivatives Market

1. The test confirmed the operability of the current TCS core version with hot standby mechanisms and the forecasted changes in latency and core module failover times.
2. The test confirmed that the TCS can handle forecasted daily loads for 2028–2029.
3. Issues were identified in the interaction between hot standby TCS core modules and the Plaza2 data dissemination infrastructure. They are scheduled to be resolved before 2027-1 release of the Derivatives Market.

Comparison of load testing parameters and peak load/data volume metrics in production environment

This section provides a table comparing load testing parameters with the maximum values of similar parameters in production environment.

Because information tables expand automatically, the initial table sizes do not define peak numbers of orders and trades. The RAM size limits on the TCS central servers are at least twice as high as the measured values.

Parameter	FX Market	Derivatives Market	Securities Market
Peak number of trades per day	1,500,000	4,376,167	12,000,000
Trades in load testing	12,976,163	9,385,084	23,867,186
Peak number of orders per day	85,000,000	175,476,181	179,000,000
Orders in load testing	183,288,745	334,388,354	439,290,571
Peak numbers of orders, trades, transactions, production configuration	Not applicable		
Peak number of transactions in one-second intervals, production environment	12,000	25,000	21,000
Peak transaction frequency, load testing	64,389	100,310	78,657