May 9, 2013

Via *E-mail*

Technical Committee of the NMS Plan
BATS Exchange, Inc.
BATS Y-Exchange, Inc.
BOX Options Exchange LLC
C2 Options Exchange, Incorporated
Chicago Board Options Exchange, Incorporated
Chicago Stock Exchange, Inc.
EDGA Exchange, Inc.
EDGX Exchange, Inc.
Financial Industry Regulatory Authority, Inc.
International Securities Exchange, LLC
Miami International Securities Exchange, LLC
NASDAQ OMX BX, Inc.
NASDAQ OMX PHLX, LLC
The NASDAQ Stock Market LLC
National Stock Exchange, Inc.
New York Stock Exchange LLC
NYSE Arca, Inc.
NYSE MKT LLC

Elizabeth M. Murphy
Securities and Exchange Commission
100 F Street, NE
Washington, DC 20549-1090

Re: Cat or White Elephant?
CAT RFP SEC File No. S7-11-10

Dear SRO Representatives and Ms. Murphy:

Pragma is provider of agency algorithmic trading software to both brokers and buyside institutions. An important part of our business involves analyzing historical market and

order data to understand market structure and how it affects our clients' execution quality. Thus we have an interest in the CAT both as a technology provider who will help brokers comply with their new regulatory obligations, and as an experienced practitioner in the area of analyzing large financial data sets.

In principle, the consolidated audit trail (CAT) is a great opportunity to enhance regulators' ability to surveil the U.S. markets. To that end, the goal of rule 613 is "creating a comprehensive consolidated audit trail that allows regulators to efficiently and accurately track all activity in NMS securities throughout the U.S. markets." It is expected to be a massive database, growing to over 21 petabytes (or 21 million gigabytes) after 5 years.

However, the SROs' request for proposals (RFP) underscores that the Rule fails to address critical practical issues around how the enormous data set will be used and what constraints that creates for the design of the CAT. Simply collecting data does not guarantee it can be used to achieve any particular objective. On the contrary, finding useful patterns in large volumes of data is far more technically challenging than just collecting and storing the data.

Without the proper consideration to how the data will be used, the technology system that the winning bidder delivers may be fundamentally and fatally incapable of achieving the regulatory objectives. Given the potential cost of hundreds of millions of dollars that will fund the project – ultimately at investor expense – we urge the SROs to revise the RFP to include requirements around how data will be used that ensure the CAT will achieve its objectives.

The RFP devotes only one out of the RFP's 54 pages to "data access requirements." It states that an "online query tool" will allow result sets restricted to certain symbols, reporter, customer, date ranges, and so-on, and that the system must also provide for a "bulk extraction and download of data" including "capabilities to define the logic, frequency, format and distribution method."

This lack of detail is jarring given the scale of the project, and raises the question of how the results of a query are expected to be delivered and used. If a regulator had a 100Mb telecom line dedicated just to receiving data from the CAT, a download of just one day's worth of data would take about 13 days.[1] Even a 1Gb line – the same bandwidth as the kind of high-capacity cable most typically used to physically connect a workstation or server to a local networking switch in the same office or datacenter cabinet - would only let a download keep up with the CAT, meaning taking 1 year to download 1 year's worth of data.

Storing any significant slice of the data would require a massive local storage system. For example, a dataset allowing a search for a certain pattern of activity over a universe of

[1] (100e6 bit / 8)(bytes/sec) * 60 * 60 * 24 = 1.08e12 (bytes/day) = ~1 terabyte/day.

securities over a 6 month period might require 1.6 petabytes of local storage at the regulator's site. Just the disk systems to hold this would require a small room full of densely packed computer equipment with significant power and cooling requirements.

Finally, and by a significant margin the most challenging, is the computational system that would effectively analyze this dataset. Tools like Excel that are familiar to most of us are laughably inadequate to the work, with a limit of 1 million rows as compared to the 7.3 trillion records expected to constitute the dataset in the example above. Even statistical packages like SAS, R, and Matlab can only efficiently analyze the amount of data that can fit into the computer's RAM – generally a paltry few GB, or a *millionth* of the dataset discussed in this example. For this reason, financial analysis programs that work on large datasets are often written in low-level programming languages like C or Java, with the data stored in files structured to be optimal for the analysis to be performed, for example in time series. While there are general distributed computational frameworks that have been designed to work with massive datasets found in domains like search and machine learning, for example Hadoop and related projects, they must be carefully engineered and meshed to the data store they are meant to operate on.

If the CAT does not include such an analytical engine, regulators must plan on building and maintaining powerful computational and storage clusters in datacenters that are connected to the CAT via high-capacity cross-connects, and designing these computational clusters to use powerful analytical tools capable of working effectively with massive datasets. But in addition to seeming redundant with the CAT infrastructure, such an undertaking seems technically ambitious and costly not to be incorporated into the scope and financing structure of the CAT.

An analytical system might be grafted onto the CAT after the fact as a separate project, but decisions made in implementing the CAT could have a huge effect on the cost and effort involved. In the worst case, the CAT provider might have selected an architecture like a relational database that meets the reporting requirements of the RFP – essentially selecting a subset of data – but is totally unsuitable for supporting the type of flexible analysis that will allow regulators to answer the questions that interest them. The best solution appears to be to include a powerful analytical engine within the scope of the CAT itself.

To this end, the SROs should amend the RFP to include a set of questions that they expect the CAT to allow them to answer, and require bidders to explain in detail how the analytical tools included in their proposal can be used to answer the questions that have motivated the creation of the CAT, for example:

- Detect and identify wash trades used to manipulate volumes and/or prices, including wash trades among a network of collaborators.
- Do high frequency traders withdraw liquidity from the market in times of increased volatility?

- Identify instances of gaming in which a trader impacts public quotations and then trades in the opposite direction in a darkpool.
- Scan for participants with an unusual ratio of order activity to executions.

The SEC highlighted a few such examples in the Rule, and no doubt regulators have a long list of such questions that they hope to be able to answer once the CAT is in place. If bidders are not asked to demonstrate how such questions can be answered, the SROs should explain to the public and to the SEC in technical, practical detail how they intend to use the CAT to do so. Without such an explanation, we must expect the CAT, despite the enormous financial burden it will place on the industry, and ultimately on investors, to be almost useless to regulators.

The goals of the CAT are appropriate, and the project has the potential to be successful and worthwhile. However, because working with large datasets is far more challenging than assembling them, these goals are only likely to be fulfilled if the CAT project incorporates requirements to ensure that data can be effectively used.

Respectfully,

David Mechner
Chief Executive Officer