The Art of Liquidity

My Personal Journey Into Global Crypto Science, Chapter One

by James Andrew

My Journey Begins...

It was 2003.

I had, for two years, already been operating an always-profitable, high-speed, automated trading system, which actively made markets, in every NASDAQ 100 stock.[1]

I called it, “Liquidity Routing.”

It routinely made more than 100% a month, with a Sharpe Ratio that was beyond any other trading system.

While program trading systems had been around for many years, mine was a new breed. Until now, the vast majority of systems operated by analyzing the time and price data represented, in a typical candlestick chart. Mine worked by scanning the full depth order-book, for patterns which presented an opportunity to trigger an arbitrage.

At the time, I used 3D Visualization to monitor the real-time operation of my programs. I realized that, rather than looking at a screen of flickering numbers, I wanted to see a space filled with shapes, moving in meaningful ways.

[1] First as a Member of Mercury Securities, LLC, then as a partner in Dimension Capital, LLC, and finally as founder of Dwellpoint, LLC.

Here we see an early iteration of this interface. The panel on the left allowed me to adjust key algorithm parameters, like “Shave Amount,” “Spread,” and “Latency Threshold.” [2].

In the 3D View, we see a row of cylinders. Each cylinder has several cubes, which can move up and down the central shaft. Each cylinder represents one side, of an order book (i.e., Bid or Ask), of a NASDAQ security. This entire UI represents the algorithm which trades the Bid of each stock. A separate computer traded the Offer.

The cube on top of each cylinder represents the ECN currently posting, on the inside of the market. I realized that I could normalize all the price data, by reckoning all calculations in terms of *depth*, rather than in terms of absolute price.

Other cubes, on each cylinder, are continuously repositioned on the central shaft, to represent each market’s current depth, into the order book. The bottom row, of “black box” cubes, represents my own orders, which I place and continually adjust, so as always to be a little bit *deeper* than the other markets.

Behind the scenes, I had coded a sophisticated order-management system,[3] which continually placed, canceled, and replaced orders, of up to 5000 shares, of the underlying security.

When someone hit my order, causing me to have a position in the security, I had an immediate profitable exit, on one of the other exchanges. Each time this happened, the cylinder would flash, bright green. In the few cases where I missed my exit, and had to break even or take a small loss, the cylinder would flash red.

In a single view, I could comprehend all the relevant information - information that was being driven, by thousands of rapidly changing, numerical inputs. See it in action here.

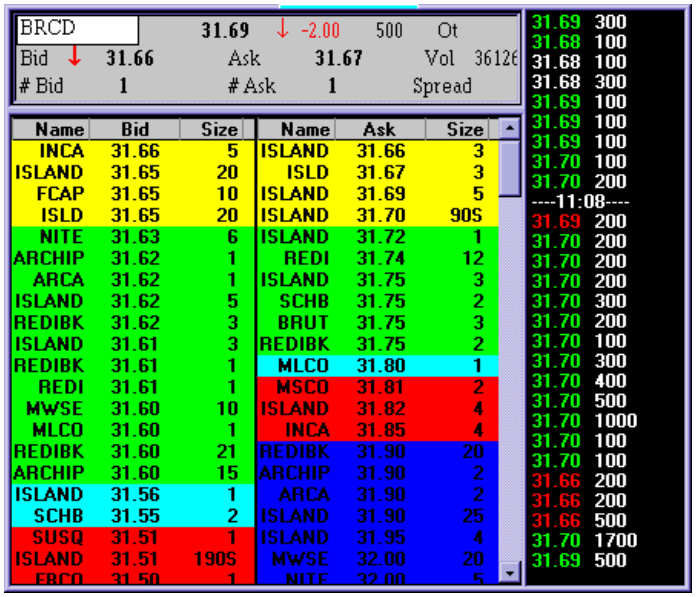
3D-2: Three Dimensional View of Market Microstructure over Time

With my program consistently producing such a profitable return, I had the rare business opportunity to get creative, on my own terms.

I decided to reimagine the NASDAQ Level 2 display. I would apply these same principles to creating a 3D geometric rendering of the price and quantity data present, in the “Level 2” order book and also the trade data, from the “Time & Sales” display.

[2] Network Latency to each ECN was an input into my algorithm

[3] Custom “AI Blackbox” framework : C++ based. Active Object Pattern implemented, in conjunction with a Concurrent Hierarchical State Machine, Agent-based system. This design was adapted from the world of Embedded Systems, which require true real-time operation.



In 1998 I started out as a NASDAQ DayTrader.

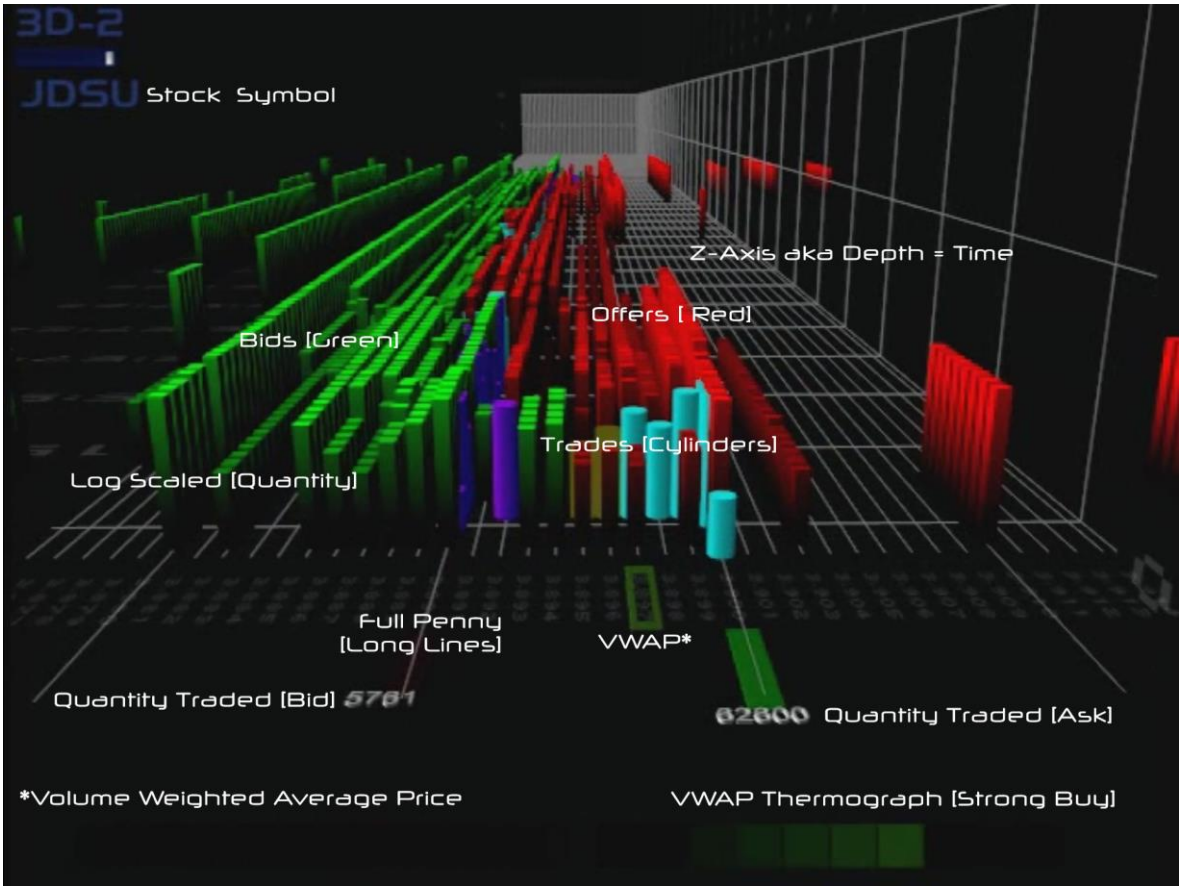
I learned to watch the “Level 2” for patterns which indicated an imminent burst, in the velocity of the price, of every stock. We would jump quickly in and out of stocks, largely based on these patterns. We called this “scalping” and “momentum trading”. The “Level 2” display was core to our trading. This was where we saw the full depth of the market.

We were the first generation of retail traders which had access to this pivotal information.

Using 3D, I realized I could visually represent this data over time. Now, at long last, I could see those patterns of behavior with my eyes - patterns which I had only been able to imagine, over the years I had spent watching the market.

Over the next few months I used the latest 3D Technology[4] on top of a custom C++, Active Object based market data engine to realize the product I named, “3D-2”.

[4] “Quest 3D” by Act3D was a powerful node based authoring environment similar in many ways to Unity 3D. It encapsulated the full Direct3D 8 API and allowed me to create custom high performance C++ nodes to perform arbitrary, multithreaded computation, and then to render the results, in real-time 3D.



It worked like this…

- The horizontal, X-Axis represents ***price***.
- The vertical, Y-Axis represents ***quantity***.
- The depth aka Z-Axis represents ***time***.

The green and red columns represent orders, in the order-book. The foremost row represents the current order-book. Bids are green. Offers are red.

- The horizontal position of each column indicates the price
- The height of each column represents the total quantity, at that price.

The cylinders represent actual trades, as reported by the market.

- The horizontal position of each cylinder represents the price of the trade execution.
- The height of each cylinder represents the quantity of the trade execution.

Periodically, the program would take a snapshot of the front row, push it back towards the rear of the space, and make a new row representing the updated state of the order-book.

The horizontal green gradient, at the bottom of the display, is a visual indicator, of the Volume Weighted Average Price (VWAP) of this stock's recent trades. Tracking the average price of trades reveals the buildup of supply/demand imbalance. This is a leading indicator, of approaching price change in the stock.

To see this in action, use this video link. In this clip, we see pressure building as buys outpace sells. The VWAP indicator predicts the move, and finally, the price pops to the next level.

When I first saw this *pop* [video @ 1m42s], my eyes were opened. My brain comprehended now, in an instant, what I only could apprehend before, in bits and pieces. This was market microstructure, made simple as tinker toys - the superlative advantage over every poor, blind trader, stumbling about in the frothing sea and foaming chaos, of so many “impressive” Arabic numerals, while I bestrode the world, a colossus of Rhoads, high and dry, seeing everything – not only the present moment, but also the trajectory, of the present moment.

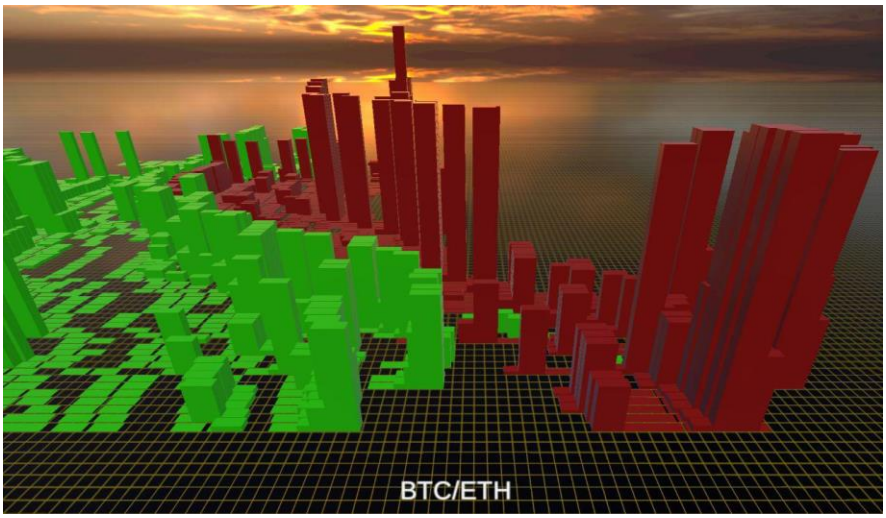
The competition had in their favor, as many single points of present-tense data, as they could decipher, one or two numbers at a time, if they were peculiarly gifted at the performance, of this impressive parlor trick.

I, now, on the other hand, needed no braille, no banister, no length of rope to guide me through the darkness. I was no longer a blind man. I no longer needed nor desired any such tedious description - numerical, verbal, or otherwise - of the geometric phenomenon, of the dangerous, winding road ahead of me, nor of the other cars in that race, for the zero-sum-checkered-flag, of the trading game.

Time passed, things happened, fast-forward:

In September 2017, I saw everything going on in the world of crypto trading and, looking at the order-books, on sites like GDAX and Poloniex, I was struck by how similar they were to the NASDAQ order-books I once traded. I was excited by this, and by everything I was finding out about Crypto and about Ethereum, in general. I decided to jump back in, and pick up where I left off many years ago.

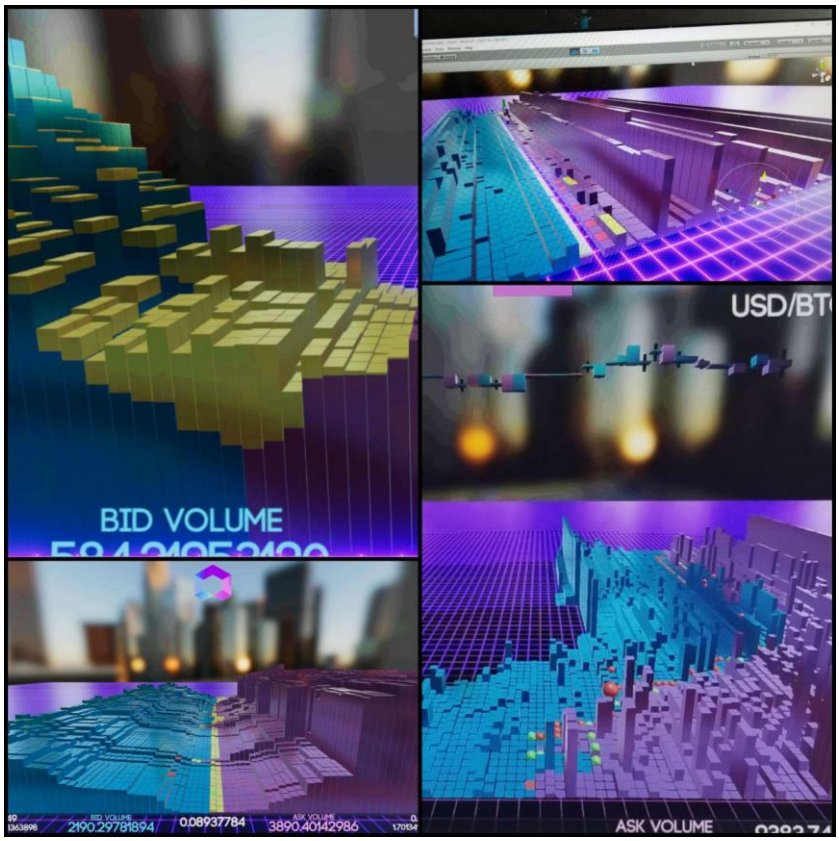
Here is a video I made on September 30th, the day I first got the basic 3D view working for BTC/ETH.



Then in October, I discovered the Omega One whitepaper, and I was impressed. It made me smile when I saw how similar certain aspects, of the proposed architecture were, to my own, automated trading system architecture, of the early years of my journey. I ended up meeting Omega’s charismatic founder, Alex, out at an event where he was speaking. I showed him a live demo of my 3D market visualization. He loved my invention, we got along famously, and the two of us were off to the races.

I began working with Omega One, just a few days later.

While at Omega One, I have been refining and advancing the 3D Visualization to support the aggregation of multiple markets, while building out the engine to support the unique requirements, of the Crypto marketplace.



- Market Data Links to four crypto exchanges (GDAX, Poloniex, Bittrex, Bitfinex)
- Aggregates liquidity from all exchanges into a single display
- Represents arbitrage opportunities, visually. [The gold bars in the image]
- Represents trades from each exchange as spheres, sized according to trade size.
- Integrated simple Candlestick charting
- Pricing data is normalized, by reckoning prices as a percentage of the current mid-price. [Each column represents a “bucket” of quotes within a range of prices. I.e., 0.01%, aka 1 basis-point]
- Cutting edge “compute shader” GPU rendering enables complex 3D visualization, at 60 FPS on a standard iPad, with low power consumption.

For the past six months, I have been travelling all over the world, as Technical Director for Omega One, and have enjoyed the unique opportunity to share this experience, with hundreds of people, in person. To say the response has been overwhelmingly positive would be the grossest of understatements. People love this, like starving men love the smell of bread, (or like sharks, the smell of blood) and they are begging me, for more of it. Simply put, this is the future of trading, and the future of Crypto, and everyone with a stack of chips is bellying up to the wheel, jostling for position, before I drop the little metal ball, on my forthcoming beta-test.

Here you can see a video of the “Omega 3D” prototype in action.

However:

As I set out to design the technical architecture, which would deliver all this market data, to clients all over the world, I realized that if I designed it using a typical, centralized-server-based architecture, the system would surely fail.

There is a fatal problem with any such attempt, to provide access to all the globally distributed liquidity sources, my young padawan.

The global, geographically distributed nature, of both the liquidity and the universe of Crypto traders, demands that the technical architecture, of any market data delivery and trade execution platform, fit for the global Crypto marketplace, must also be decentralized and geographically distributed.

Crypto Exchange Architecture

With the advent of the crypto marketplace, trading is forever changed. The crypto markets are global. Dozens of significant exchanges allow for trading of the same instruments. These pools of liquidity are spread across the surface of the planet. Many run in Amazon AWS or Google Cloud Platform data centers. For the owners of these exchanges, the technical architecture has been relatively straightforward. From the perspective of a centralized exchange like Binance or GDAX, the right approach is to set up an exchange, which functions very much like a traditional centralized equities exchange (Nasdaq) or ECN (Island, Archipelago, etc).

- Set up one or more servers, either self-hosting in racks, or cloud hosted, via Amazon EC2 or similar.
- Publish a real-time order-book, trade data, and other relevant data using the public internet.
- Clients of the exchange either pull this data periodically (REST/JSON) or the exchange pushes a real-time, event based data feed. (WebSockets,FIX).
- Clients of the exchange place, cancel, and modify orders to buy and sell, in response to the data they get from the exchange. These orders are sent over the public internet, using the same protocols (REST/WebSockets/FIX).

The Problem with Liquidity Aggregators

Right now, the fragmentation of the Crypto markets is an insurmountable problem. The fact that there are so many exchanges is a barrier, preventing these markets from functioning nearly as well as they should. If a trader wants to fill a large order, the ideal situation would be to have access to all the liquidity currently represented, across all the centralized exchanges, and also across all the Decentralized Exchange Protocols[5] . There is an obvious need, for a service which allows traders to access all the available liquidity. There are a number of companies contending for this lofty prize, but to no avail, since they are all neglecting, quite thoroughly, to account for the globally distributed nature, of these exchanges.

[5] For now, I will leave out the solution for connecting to the DEX's and focus on the centralized exchanges, where the vast majority of trading happens.

Instead, they are architecting their services after the manner of one of the centralized exchanges, described above. They are setting up a server in some particular place, and they are connecting, over the public internet, to the various exchanges. Clients connect to this server which redistributes the aggregated market data, and relays orders and confirmations to the various connected exchanges, over the public internet.

This approach is as intellectually impressive, and as practically useless, as any perpetual motion device, of the 19th century.

Now in order to understand the nature of these drawbacks, we must first understand some core trading concepts.

Marketplaces provide two major functions.

1. **Liquidity** : This is the set of orders to buy or sell which may be executed against, by market participants.
2. **Price Discovery** : The “proper” price of the instrument is continually negotiated, by the set of buyers and sellers participating in the market.

While every market facilitates these two things, *the way in which it is done* can make or break the market, because all liquidity is not of the same quality, not of the same value, not of the same relevance, at any given time, in any global game, limited, not by any abstract concept of time, but by the simple and pesky reality, of the speed of electricity.

All liquidity is not “created equal,” but is determined by that critical factor, which we will call, “availability.” Accurate price discovery requires what we will call, “highly available” liquidity.

And so, to be perfectly vivid with my dear reader, those of us without a Tardis parked in our garage, had best pay attention, or go the way of the dodo bird.

Alas, poor dodo, I knew him well, a fellow of infinite jest.

With Liquidity, Availability is Everything, (so understand this, or perish in the battle to come):

In the context of electronic trading, let us define *availability* as, “the probability that the liquidity represented by a quote - an order to buy or sell - is still *valid*, by the time the order to execute a trade (which matches the quote) reaches the exchange, where the quote originated.”

Let’s define the *range* for availability, as a decimal, from zero to one. Zero means it is certainly not available. One means it is certainly available.

In an ideal scenario, each quoted order will have an availability of 1.0. That is, there is a 100% probability that the order is still valid, at the time the trader sees it, at the time the trader places an order, and at the time the exchange receives the order to match against the quote.

In the real world this is never the case, largely because of network latency. Consider that with a network latency of *n* milliseconds, there will always be some probability that the counterparty has cancelled their order in the past *n* milliseconds, and that you, the trader, do not know that yet. As the

value for *n* rises, so does the probability that any given order in the exchange's order-book is no longer valid.

Traders call this a "stale" quote.

As the number of stale quotes increases, the value of the information presented by the exchange decreases.

And so…

(cue Jeopardy Think-Music)

…the value of the exchange itself decreases.

Do not pass go, do not collect two hundred dollars.

High Frequency Trading : A Mechanism of Superior Price Discovery

Over the past fifteen years, the most advanced trading firms have realized something very important.

Even in a market with perfect liquidity availability, the orders represented in the order-book do not present a complete and accurate picture of the supply, demand, and optimal price for a given instrument. The reason is that many of the orders are like icebergs. The counterparty reveals only a small piece of their complete order so as not to "tip off" the other market participants to their true intentions.

HFT firms have recognized that by placing small orders, they can force the counterparty to respond in a way which gives a more accurate picture of the real supply or demand for the instrument. This is because, by executing a small trade, they force the counterparty to respond in one of several ways. Let's say the counterparty is selling 1000 units. The HFT trader executes 100 units. This counterparty may then either refresh the full order, or cancel the remainder, or refresh only the remaining amount. This information is extremely valuable, as a means of detecting the "real" amount, waiting to be bought or sold at the moment. It's something like an "active sonar ping," where advantage is gained, by sending out a signal which interacts with the target, yielding partial information about its location.

Traders who employ this technique will, all other things being equal, outperform traders who do not. In the equities markets, HFT firms have drastically outperformed traditional firms.[6]

In the context of HFT, latency is the enemy. An HFT trade generates a signal in the form of the response. This signal is only fully valid at the moment it is generated by the counterparty. From an information theory perspective, the signal becomes "noisy," i.e., progressively more unreliable as the latency of the network increases. The longer it takes to receive the signal, i.e., to receive the updated quote, the less useful this information is, as a means of price discovery.

[6] Adam D. Clark-Joseph. *Exploratory Trading*

Sophisticated market participants will demand HFT capability. The winning solution for a comprehensive interface, to all global crypto-market liquidity, must therefore support HFT.

"...questions about HFT implications are inherently computational, as the very speed of operation renders details of internal market operations—especially the structure of communication channels—systematically relevant to market performance. In particular, the latencies between market events (transactions, price updates, order submissions) and when market participants observe these activities become pivotal, as even the smallest latency differential can significantly affect trading outcomes."[7]

The Naive Solution (Which Will not Stand in the Battle)

Companies which use legacy technologies from the pre-crypto era, for architecting their crypto-exchange aggregation solution, might attempt the following, in vain:

- Set up a server or set of servers in a server room or single datacenter near their place of business. (i.e. in Amazon AWS US-East Ohio)
- Connect over the public internet to all the crypto exchanges they desire to offer
- Expose a FIX or WebSocket API which will allow clients to connect over the public internet for market data and trading.

The primary problem here should now be obvious to the diligent reader: t**he geographical distance of each exchange and the geographical distance of each client, from this single datacenter, creates an unfair marketplace.** This problem is so severe as to make such a solution entirely unsuitable for servicing the globally distributed marketplace. Even the exchanges which are relatively close to the single datacenter, but not in the same datacenter, are still relatively distant in terms of latency over the public internet. Clients that want to employ HFT techniques, will be sorely disappointed with the performance of their algorithms. The will not be able to compete with those who have co-located their HFT capable algorithms with this exchange.

Consider what happens if a client in Asia wants to use this system to trade against Asian exchanges. The client is used to connecting directly to these Asian exchanges, either via their websites, or via their APIs, over the public internet. However, when connecting to this hypothetical exchange aggregator, he will see that same Asian liquidity, with additional delay that may be over 250 milliseconds. This is because it takes about that long for the Asian market data to make it to Ohio, and then all the way back over the Pacific, and this 250 milliseconds is only the median delay. Trans-Pacific public internet links are notoriously variable, in their latency and see many latency spikes imposing even greater delay.

A sophisticated trader will reject this solution immediately.

There are other mortal flaws with this approach, as well.

- The single server in a single datacenter represents a single point of failure for the entire system. If this datacenter goes down, the entire system "goes dark".
- Scalability becomes a serious problem, as more and more traders want to trade more and more instruments. All the data must be pulled to, processed at, and redistributed from one location.

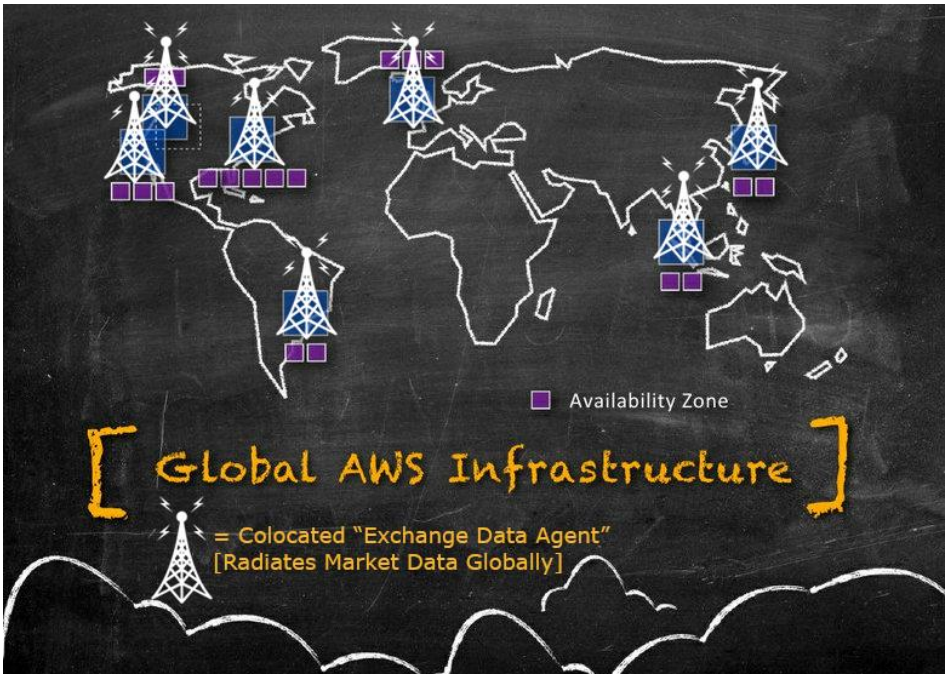
[7] Elaine Wah, Michael P. Wellman. *Latency Arbitrage in Fragmented Markets: A strategic agent-based analysis*

The Ideal Solution: The Global Liquidity Network (Which Shall Prevail and Reign Supreme, Without Rival in the Earth)

In the first wave of electronic trading, extreme centralization was the obvious approach to creating a highly competitive, level playing field for traders. The reason, however, that this approach was even possible, was that there was already a highly centralized monopoly on the business, of running an exchange. The Crypto Era is fundamentally different. Now, there exists a globally distributed collection of exchanges, and a globally distributed universe of traders. What must happen, before this will work properly, is the deployment of a globally distributed, decentralized solution, for connecting all the liquidity represented, across all the exchanges, with all the traders. The key features of this system shall be as follows:

- It will recognize that there is *no single globally correct order-book* for a given instrument. The trader's location on the globe and on the network will determine the ideal liquidity-picture, for that trader. "Closer" Liquidity is more valuable, because it is more available.
- It will recognize that moving data over the public internet is the least desirable case, and that the ideal solution will take advantage, of the more robust messaging solutions currently available.

The Crypto Exchanges are globally distributed, and so is the universe of crypto traders. The technical architecture for the ideal gateway into the crypto-trading *dataverse* will manifest the recognition of this geophysical reality.



What the world cannot do without, is a Global Liquidity Network.

Stage 1: Exchange Data Agent [EDA]

The solution begins with the design and deployment of a global network of "Exchange Data Agents" which "live" at the AWS and GCP data centers and which interact, at very low latency, with their corresponding exchange. For example, Binance is hosted at Amazon AWS datacenter called, "ap-

northeast-1". It is possible to operate an Amazon EC2 instance, running Windows or Linux, in the same datacenter as Binance. The result is that network communication with Binance takes place on a LAN, and not over the public Internet. The result is a much faster, much more stable, much less variable connection to Binance.

The EDAs will "broadcast" their data on a globally available, guaranteed delivery, message queue. The rise of the "realtime internet" has seen the rise of a number of protocols (AMQP) and middleware service providers (Ably, PubNub, Pusher, AWS IOT) that specialize in reliable global message delivery. We will call these services, "Global Message Queue Services." These services are commonly used, for real-time application involving IOT devices, and services like Uber, which need massively concurrent global messaging, among many thousands, or even millions, of devices. These services act like an "information superhighway," compared to the windy back-roads of the public internet. It's like a CDN for messaging, with some remarkable and relevant features such as:

- Guaranteed, ordered message delivery
- Latency Based Routing
- "Radiated" Global Message Delivery (local clients get messages sooner)
- Connection State Recovery
- Redundancy - No Single Point of Failure
- Message Persistence supports seek and replay of all messages.

Every exchange in the network will have a corresponding Exchange Data Agent. The EDA will:

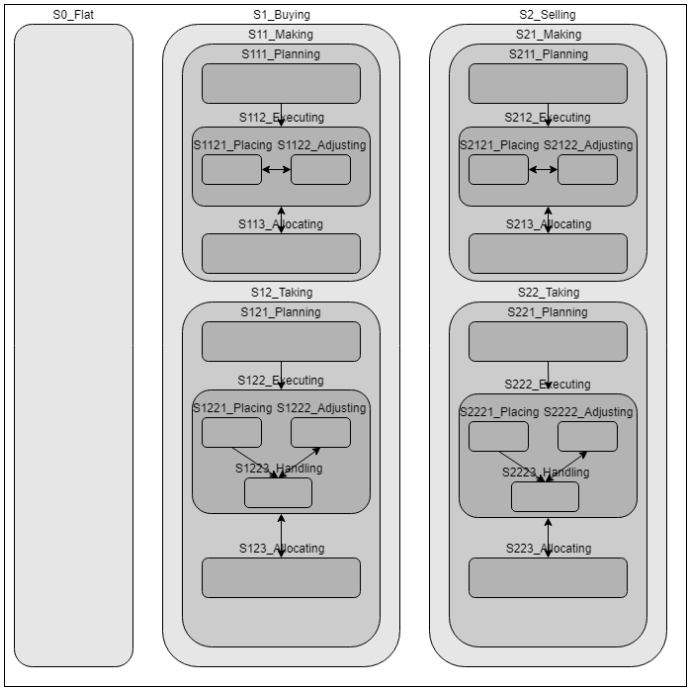
- Subscribe to full depth order-book data, via websocket
- Subscribe to trades (time & sales) data, via websocket
- Measure and Calculate Exchange Metadata (i.e., uptime, availability, latency)
- Publish full depth order data, via a GMQS
- Publish trades data, via a GMQS
- Publish Exchange Metadata, via a GMQS

Subscribers to the network will "tune in" to this data, from wherever the trader is on earth, and will receive an optimal view, of global liquidity.

Stage 2 : Exchange Trading Agent [ETA]

Using the same network architecture, this system will surely be extended to support trading, via the introduction of the Exchange Trading Agent.

Each ETA will handle trade execution on behalf of the global network of traders using the system. The idea is simple. Colocate these agents with each exchange. The ETA's will be assigned jobs on behalf of traders. Each ETA will use HFT techniques and their low-latency exchange connection in order to harvest liquidity, and in order to make markets with the highest possible efficiency.



The ETA will surely be implemented, as a hierarchical finite state machine, which operates according to this diagram. The ETA will trade on behalf of traders placing orders, using the traders' exchange API keys.

Stage 3: Decentralized Matching Engine, aka, Dark Pool

Once there is established a global network of ETAs, all continually working to fill client orders, there will be the basis, for a decentralized dark pool. By making the agents aware of what orders the other agents are "loaded" with, the network can begin to match buyers and sellers, without needing to harvest liquidity, or make markets in the external exchanges.

And then the world… will shake.

To Be Continued...