DELL Technologies NVIDIA

Algorithmic Trading HPC & AI Reference Guide

Authors

Erik Vynckier
Erik Vynckier is board member of Foresters Friendly Society and chair of the Institute and Faculty of Actuaries (Research and Thought Leadership Board), following a career in investment banking, insurance, asset management and the petrochemical industry.

He co-founded EU initiatives on high performance computing and big data in finance and co-authored **"High-Performance Computing in Finance"** and **"Tercentenary Essays on the Philosophy and Science of Leibniz"**. Erik graduated as MBA at London Business School and as chemical engineer at Universiteit Gent.

Gabriel Pirastru – Dell Technologies, HPC & AI Team

John Ashely – NVIDIA FSI team

Keith Manthey and Darren Miller – Dell Technologies, UDS

For any enquires regarding Algorithmic Trading and Dell Technologies:
Gabriel_Pirastru@Dell.com

For any enquires regarding Algorithmic Trading and NVIDIA:
jashley@nvidia.com

For any enquiries regarding Algorithmic Trading and storage solutions:
Keith.Manthey@dell.com

Contact your local HPC contact:

- HPC_NA_Sales_Team@Dell.com
- EMEA_HPC_Team@Dell.com
- APJ_HPC_Team@Dell.com
- HPC_Latam@Dell.com

Dell Technologies Useful Links:

- Guides to Connected Finance
- High-Performance Computing
- HPC and AI Innovation Lab
- Reference Architectures

Contents

Introduction

Information technology has been a driver of innovation in the financial industry over at least the past thirty years. What is the status quo today, with more data and computing power at our disposal than ever before?

This paper will highlight algorithmic trading as one example of doing more with information technology in finance. Algorithmic trading is automatic electronic trading using computer programs to make buy and sell decisions without immediate human intervention. The human touch is in the intelligence the program embeds. However, the step to artificial intelligence is now being made.

Algorithmic trading describes the overall industry of both algorithm development and high-frequency trading. Algorithmic development refers to the design of the algorithm, mostly done by humans. High-frequency trading, on the other hand, involves putting the developed algorithm in practical use for trading. This latter is a very low-latency and high-volume strategy and is being done by a machine.

We'll look at how the industry operates, before moving to the technical challenges and how these are being addressed. Finally, we'll explore how to size the best solution combining high performance computation and data storage to serve your needs best.

Level Setting

i. Why do we do algorithmic trading?

The ability to remove humans from trading has several advantages such as automation, speed, accuracy and reduced costs, data retrieval and trading speed but most importantly not letting human emotions impact trade.

- **Accuracy;** algorithms are made of complex correlations of data and events/ pattern detection that enable them to react to information (news, market price changes, events…) and place trade accordingly to make profit.

- **Aggregate large volume of data;** by leveraging compute and storage technology, algorithms can now crunch lots of data quickly. Additionally, they can be tested on historical data back from 10, 20 or more years for better accuracy.

- **Data Retrieval Speed;** updating the algorithm in reaction to new data (market data, events, strategy change…) is critical. Therefore, having enough compute power, the right database and networking technology to make necessary changes is key.

- **Emotions;** removing emotions and placing trade at the right time ultimately results in profitability improvement.

- **Trading speed;** High-Frequency Trading (HFT) uses latency to leverage market information to make profits before competition. Speed is in milliseconds; in HFT latency is often the only determinant for profit/ loss. Not all algorithmic trading is high-frequency, traders will still optimise their trading platform to execute trade with low-latency. This is to avoid large price changes between the moment the computer places the trade and the exchange receives it.

ii. Pattern detection and Risk

One of the primary purposes of using algorithms is to detect patterns amongst data against a trading strategy. Basically, each data input, e.g. market movements (up/down), important news, impacts the market. Having this understanding enables the trader to decide which action to take as a result.

Secondly, the risk associated is important to take into consideration. Does the trader want to take a high-reward/ high-risk strategy or a safe-strategy, resulting in low but steady returns? To assess the risk level, Monte Carlo and other risk analyses are being performed.

The combination of data-correlation, pattern detection, risk analysis and portfolio optimisation will form the algorithm.

iii. Back-testing

Another advantage to algorithmic trading is the ability to do back-testing. Previously, traders had no idea whether their strategy would work before trying. Now, algorithms first run on past data to see whether they yield satisfying results. Once they reach the right level of profit-making, they are then applied to the real world.

Additionally, using emerging methods such as VAEs (Variational Autoencoders) and GANs (Generative Adversarial Networks), algorithms can be applied in simulated market environments. VAEs and GANs are generative models and are a great way to learn any kind of distribution using unsupervised learning.

iv. Short-shelf time

When algorithms are being deployed and start trading, the market and competition will inevitably react. Competitors play an endless game to stay ahead and make profits, in this play there are only winners or losers. Whether profits are made is often dictated by the ability to deploy large and powerful IT infrastructures that will enable the firm to correlate large amount of data and use it to trade at the lowest latency possible.

As more and more algorithmic trading strategies are being used, it can be more difficult to deploy them profitably. Indeed, competition is so stiff that entry barriers are high, especially regarding the cost of a performing IT infrastructure.

v. Who is doing algorithmic trading?

Many algorithmic trading firms are market makers. This is a firm that stands ready to buy and sell a stock on a regular and continuous basis at a publicly quoted price. Customers use them to place large trades.

Large quantitative funds (also called investment or hedge funds) and banks have an algorithmic trading department and benefit from having large funds to invest along with the ability to correlate their own internal data with the algorithm.

High-frequency trading firms are leveraging low-latency technologies to make profits. They often place trades on behalf of other financial institutions.

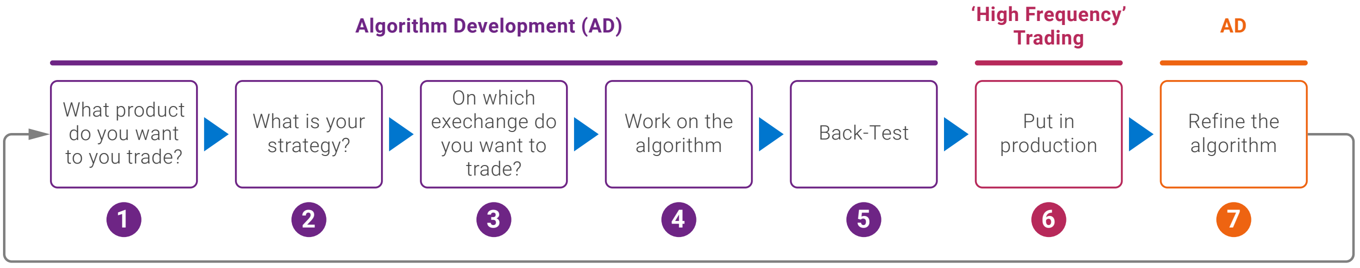
vi. What is the workflow of algorithmic trading?

First, the algorithm has to be developed. This is being done by a mix of data science, statistics, risk analysis and DevOps. Secondly, the algorithm will be used for back-testing, trying it against past data. This gets repeated until the algorithm is refined (i.e. produces the expected profits). Once results are satisfactory, the algorithm is put in production and trades on behalf of the firm. The yields produced by the algorithm trading in the real-world markets will produce data and further feed the algorithm in the backend.

This process puts a lot of stress on the infrastructure, since continuous data must be fed to the algorithm, so that it produces best results.

Additionally, the high-frequency trading platform places thousands of trades within a very short amount of time, before anyone, or the market, can react. Therefore, it must have the lowest latency possible.

Each sub-part will be further explored in the paper.



Algorithm and strategy optimisation

1. Products

Term	Definition
Equities	Shares of a company (also called stock)
Currencies (FX)	Dollar, Euro, Pound…
Futures	Betting on the future value of an asset
Bonds	Governments or companies' debt
Derivatives	Betting on artificial values (e.g. whether the whole NASDAQ's value will increase/decrease)

2. Strategy: the trader will calibrate the algorithm according to a strategy i.e. risk level, long vs. short term.

3. Exchanges: places where products are being bought and sold, exchanges are split by product and geography (e.g. Deutsche Börse is an exchange for equities and securities in Germany) – see part viii, Market rules.

4. Work on the algorithm: run risk analysis, time-series, artificial intelligence techniques and create your own algorithm. The algorithm will be based on the product chosen (how liquid it is), the market it trades (data available) and strategy (level of risk willing to take). Then analysis will be done to find which patterns in the market influence have most effects on these factors.

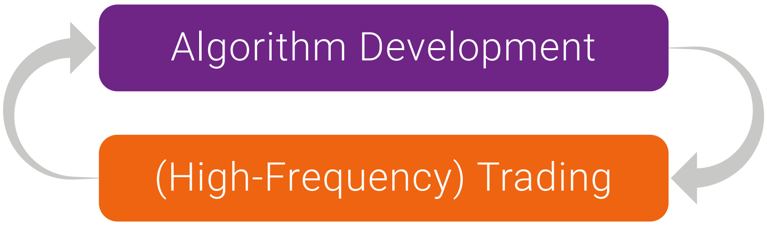
5. Back-test, will then take this algorithm and test it on historical and simulated data.

6. Production: trade with the algorithm. In the past high-frequency trading was the norm, i.e. executing many trades (thousands or millions) in a fraction of second before the competition can react. The problem is that there is only space for so many traders operating such a strategy. Therefore, the new key competitive advantage is how much data and how well one can compute and leverage.

7. Refine the algorithm: see how the market reacts and change your algorithm to get even better profits.

As a result, we can think of algorithmic trading as a never-ending loop between refining the algorithm and trading with it.

Algorithmic Trading Loop



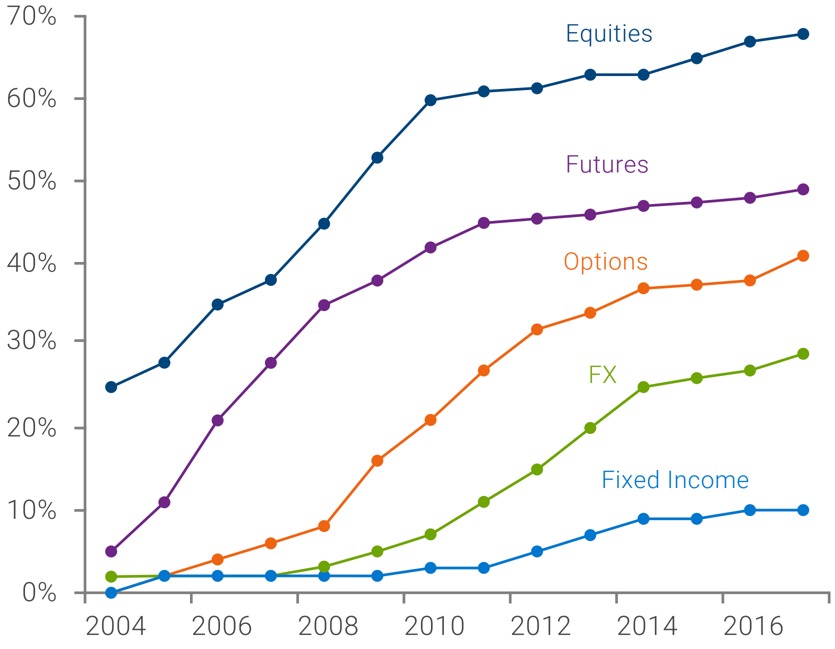
Algorithm Development

(High-Frequency) Trading

vii. How big is algorithmic trading versus 'standard' trading?

The below graph indicates the share in percentage of algorithmic trading against traditional trading amongst financial products.

Market Share of Algorithmic Trading by Asset Class



As of 2017
Source: Goldman Sachs, Aite Group

I. Industry

1. Information technology as a competitive advantage in financial services

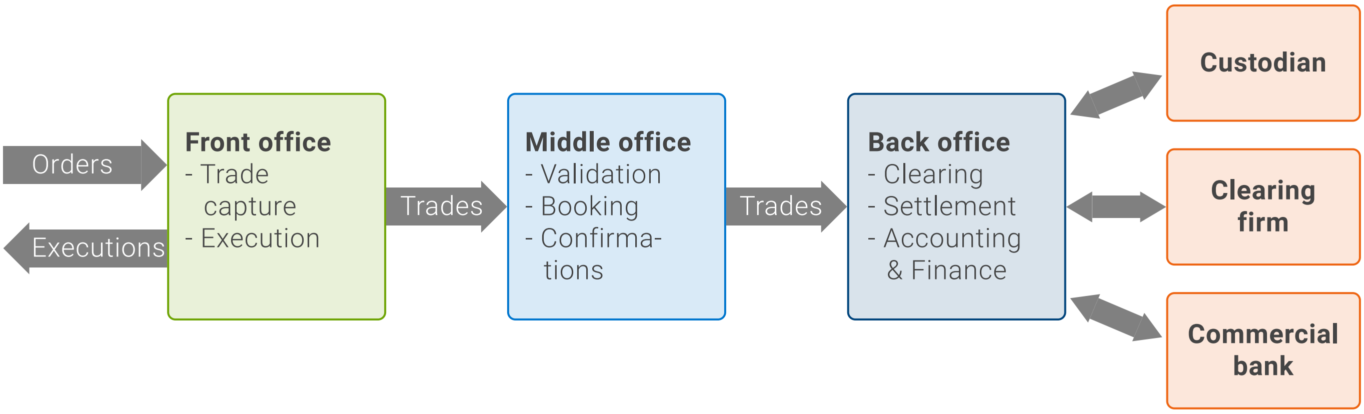
The financial industry is a heavy user of information technology in all its forms: from the basic administration and accounting software for managing its clients and its own assets and loans to the online availability of accounts, transactions and cyber security.

In the investments business, information technology is used for front office, which is where the trader sits, accesses information and places the trade, middle office when the trade is made both parties send electronic confirmation, and back office where the legal details and reporting is done. All these steps occur electronically without human interaction.

1. **Front office:** electronic trading of securities in
   - Price exploration
   - Transacting on the exchange
   - Over the counter with a market maker

2. **Middle office:** confirming trades to the mutual satisfaction of the counter-parties managing the portfolios through service providers such as custodians and clearing houses

3. **Back office:** settlement and for valuation, performance measurement and quantitative risk management.



Securities trade processing flow

The financial industry is heavily regulated and owes many different routine reports to its regulators, clients and partners, such as custodians and exchanges. Capital is a peculiarity to the financial industry: sound institutions require adequate capital availability and a robust capital management process, which is an input into their license to operate.

Capital models are computationally heavy and require up-to-date positions and valuations from their firm's own administration system and from the capital markets. That means that data must be accessed and retrieved from exchanges to update internal systems.

Automated tools have also picked up a role in compliance monitoring, where they automatically update the legal and reporting institutions so that traders only focus on trading.

Information technology – computation and data – are a critical ingredient for a successful financial services company and sometimes its prime, or even only source, of competitive advantage.

New platforms now permit live or quasi-live computation where ten years ago, computationally demanding tasks such as capital computations, credit or capital value-adjustments (CVA/KVA/XVA) to derivatives books and variable annuity hedging programs involved overnight batch runs or weekend runs with complicated planning and inevitable frustration when errors triggered reruns of critical and long-awaited outcomes. Expansion into alternative data and use of machine learning also trigger more demand for scalable computation.

2. Electronic trading, algorithmic trading and high-frequency trading

This paper will focus on high performance computational needs in electronic trading and investment management as a particularly demanding client audience in financial services. Let's introduce some terminology.

i. Electronic trading

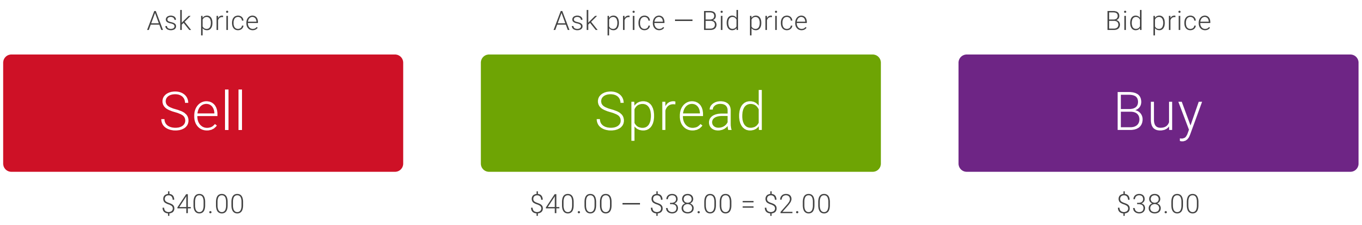
Investment managers use information on the valuations of different products (e.g. securities, money markets and currencies, commodities and derivatives) to buy low and sell high. Transactions position the portfolio where the investment manager expects returns. This is the first step of the process defined in the introduction; choosing the right product, or here the portfolio, a mix of products.

Many portfolios are "buy and hold" or "buy and maintain", where the investment manager is patient for the investment bet to play out, but some are actively managed. Rebalancing the portfolio requires a trade.

Electronic trading has enabled cheap and efficient transactions at low bid-ask spreads in many common assets and called into existence algorithmic trading and even high frequency trading as an investment strategy.

A bid-ask spread is essentially the difference between the highest price that a buyer is willing to pay for an asset and the lowest price that a seller is willing to accept. A low spread means getting the lowest-price available.

Bid, ask and spread



Money-zine.com

ii. Algorithmic trading

Algorithmic trading is a highly automated investment process in equities, currencies, futures and exchange traded bonds and derivatives, where an algorithm selects the investments and implements the trades to achieve the desired portfolios.

iii. High-frequency trading

The ultimate evolution of algorithmic trading is high-frequency trading. You could say that in high-frequency trading, the trade itself has become the profitable investment. The information used to design investments is the trading process itself.

High-frequency trading has led to competition in computational speed and automated decision making, connectivity to the execution venue and even co-location at the execution venue to save in microseconds and beat the next trader in timely execution on an opportunity.

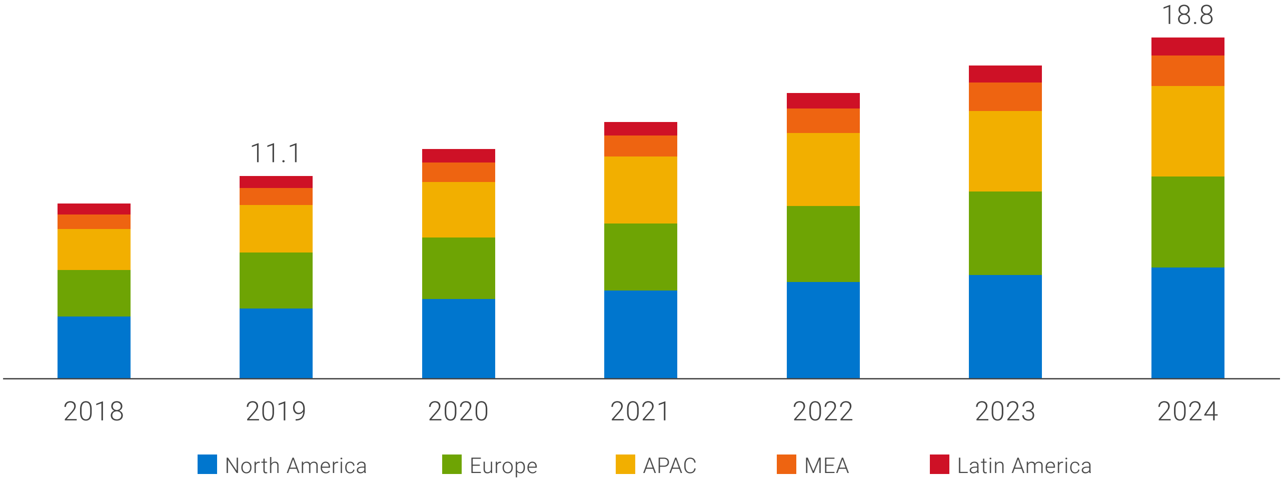
Indeed, some companies are so dependent on how fast their trade is being executed, i.e. before anyone else, that they purchase buildings that are just next to the exchange so that they reduce latency to its lowest.

High-frequency trading has moved more and more into the space of market making, which is when securities broker offer two-way markets, buying and selling, from their inventory of securities at their own profit and loss.

iv. Algorithmic trading market

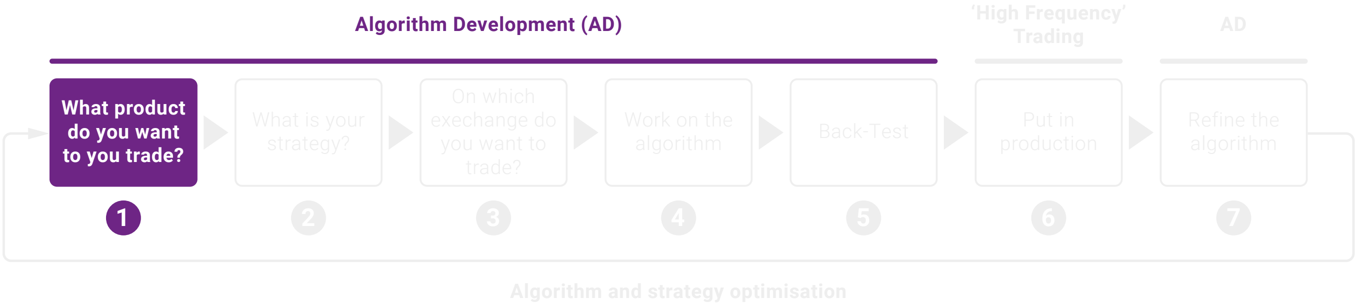
Algorithmic trading is a growing market overall and is becoming more the norm than the exception. What is important to understand is that the IT investment necessary to be able to do algorithmic trading is significant. Therefore, a growing market means further investment in the necessary technologies.

Algorithmic Trading Market, by Region (USD Billion)



Source: MarketsandMarkets Analysis

v. Products traded in algorithmic trading


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exchange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

Equity reflects the ownership of a company. Equity is embodied in shares, which reflect pro-rata ownership. Private equity is unlisted and illiquid whereas public equity is publicly listed and traded on stock exchanges.

Fixed income or **bonds** are issued by a variety of issuers. The most prominent are government, corporate, covered bonds and securitisations. The fixed income market has been heavily impacted by quantitative easing, since the credit crisis triggered by mortgage markets in 2008, the sovereign crisis affecting Greece and the Euro in 2011 and now the pandemic.

Since the pandemic, we are presumably in a permanent low interest rate environment. Bonds have credit quality features which bond ratings attempt to reflect. The worst outcome for a bond investor is a default of the issuer, leading to a loss of principal and coupon payments. Collateral mitigates this risk in covered bonds and securitisations.

The foreign exchange market trades **currencies** such as US dollar (US $, USD), Euro (€, EUR), Yen (¥, JPY), Pound Sterling (£, GBP) and Chinese Yuan or Renmimbi (¥, CNY or RMB). Prices are expressed for currency pairs: EUR/USD 1.1830 expresses that 1 € will purchase 1.1830 US $. The foreign exchange market trades over the counter with the main international banks making a market in currencies.

Commodities are physical resources such as oil, gold and agricultural commodities. Commodities are often traded based on standardised product with defined features and sometimes settled with a net payment if not physically delivered at an agreed warehouse.

For the purpose of hedging or for speculative investment, **derivatives** pay a cash settlement based on the movement in an underlying financial asset or commodity. Futures, swaps and put and call options are the most common derivatives. Futures and swaps are agreed contracts which bind both counterparties without discretionary choice. The final settlement could move either way. Options are the right but not the obligation to buy (call) or sell (put) an asset to the option writer at a fixed price known as the strike. Derivatives are nowadays routinely collateralised at a clearing house for securing counterparty default risk.

vi. Characteristics of trading

Risk is integral to any financial instrument. **Volatility** has been adopted as a metric although it is a very imperfect measure of risk. Volatility is the standard deviation of financial returns expressed as an annual percentage, assuming a normal or Gaussian distribution. Volatility ignores tail risk, liquidity risk, settlement risk, operational risks etc.

Investors typically hold **portfolios** of upward of thirty securities. They are bought and periodically rebalanced through trading. Trading uses two approaches, depending on the asset: trading on exchange or over the counter.

Trading on exchange is the most **transparent** trade. An exchange publishes buy and sell orders in an equity or bond. The trade settles where the cheapest sell price and the most expensive buy price meet. Outside this range expensive sell prices and cheap buy prices which do not overlap are published in the sell and buy books waiting for future execution. Exchanges are fully automated and accessed electronically.

Trading over the counter is off exchange. **Over the counter** (OTC) market makers establish an inventory of securities and trade the inventory for their own account. They may buy for or sell from inventory and make a profit or loss doing so. Even over the counter has now moved away from voice-trading over the phone and operates through execution venues, electronic platforms where market makers post availability and prices in select securities for both buy and sell transactions. The difference between buy and sell prices is called bid-ask spread.

vii. Characteristics of algorithmic trading

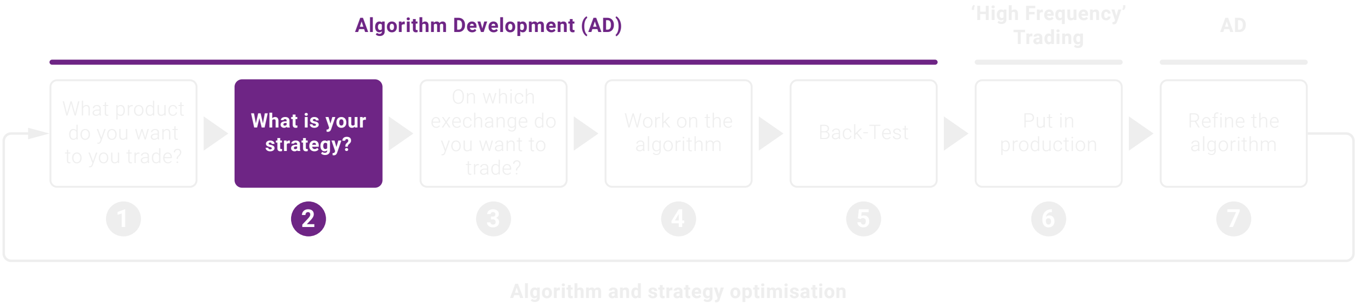
Algorithms can trade electronically on exchange or on execution venues. Algorithmic or quantitative investment management establishes quantitative strategies for investing money in the markets. The selection of securities and construction of a portfolio are based on quantitative rules (i.e. mathematical algorithms). Algorithms attempt to establish systematically successful investment strategies through investigation and back testing of quantitative strategies.

In the past mostly financial time-series of prices, balance sheet data of companies and macroeconomic data on global trade and domestic economies were used, but today there is an increasing interest in alternative data, which could be web scraping, spatiotemporal data on the real economy, credit card data as a directional signal on the economy or sectors etc.

Rebalancing of portfolios is routinely done through algorithmic or program trading in the equity market, where algorithms seek to minimise a transaction’s impact cost. Buys and sells are executed in a coherent manner, slicing and dicing a trade anonymously over many small trades, each easily absorbed by the markets.

High frequency traders are ultra-high-speed execution algorithms where in a rapid sequence of buying and selling, the trading itself is the source of profitability. High frequency traders do not keep large net exposure to the markets for any longer period of time. High frequency traders are tolerated since they support liquidity in the financial markets.

3. Strategy development process in algorithmic computing

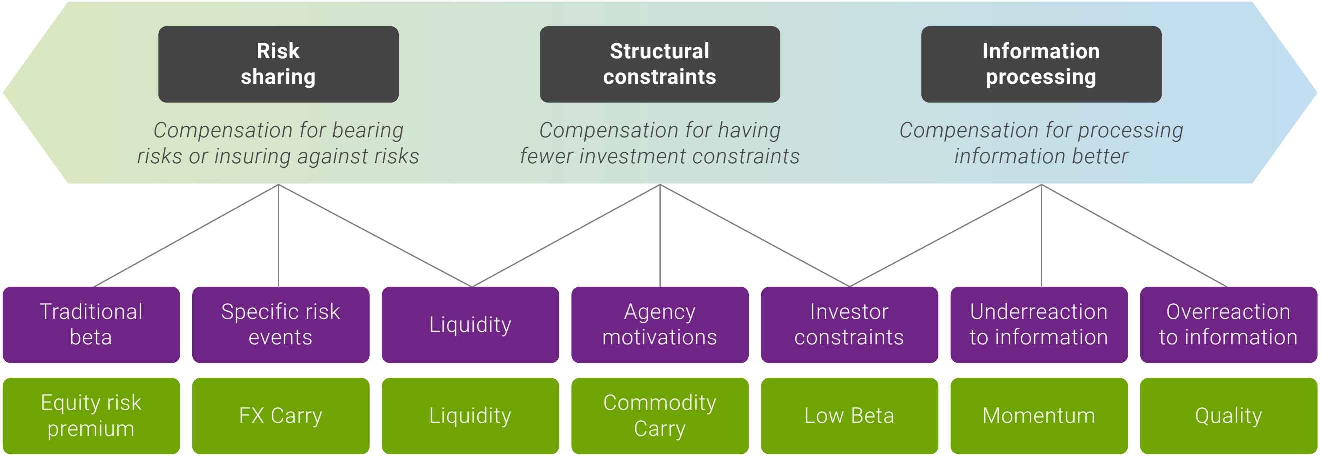

Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exchange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

Strategies for investing and trading are both data-intensive and computationally demanding and may have a limited successful shelf life.

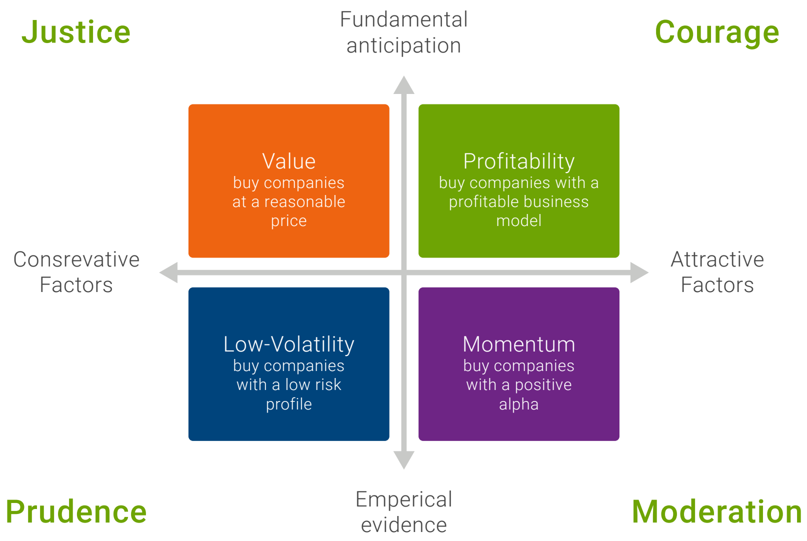
Consequently, quantitative funds continually develop, implement and perfect algorithmic trading strategies on the go. Quantitative researchers and developers, portfolio managers and risk managers compete with the next fund through innovation in investment and trading but also through their speed at implementing a successful strategy. Developing code and back-testing are an ongoing activity and in an ideal scenario, a successfully back-tested algorithm can be implemented in trading without delay. Ideally, the research code will serve the live trades as well. Meaning the code being developed during the algorithm development will be used for trading on the exchange.

Having larger databases for more securities and longer histories can enable stronger back-tests. Having the computational capacity to run such back-tests quickly speeds up the development process and enables more insight and confidence in the trading idea. A quicker launch of a trading strategy is a desirable outcome.

Here are some common strategies used in algorithmic trading



Here are some common principles used to design trading strategies


Justice
Fundamental anticipation
Courage
Value
buy companies at a reasonable price
Profitability
buy companies with a profitable business model
Consrevative Factors
Attractive Factors
Low-Volatility
buy companies with a low risk profile
Momentum
buy companies with a positive alpha
Prudence
Emperical evidence
Moderation

A strategy is simply a behaviour that an algorithm will follow – as a human would – should it be more risk taking? Prudent? Passive and quantitative approaches have become the dominant paradigm in investment management. The approach is favoured by investors for being cheap to implement, low in management fees and transparent to investors. Let's consider those in turn.

i. Passive management

Quantitative investing started off with passive investing, which is the mere efficient replication of a benchmark. Passive has become a very popular, low cost investment approach and the assets under passive management are starting to dominate the markets. Its major attraction is minimal investment management fees and low execution costs.

Passive management is a low-cost quantitative investment strategy tracking well known market capitalisation benchmarks (such as S&P 500 for US equities, FTSE All Share for UK equities or MSCI for global and regional indices) by replication.

Passive mandates acquire exposure to the equity markets cheaply and transparently. Periodical rebalances of the indices with promotions and demotions of index constituents are replicated through an algorithmic program trade. The approach is rule-based and performed for a small management fee, one of its attractions.

Market capitalisation benchmarks tend to be over-invested in the most liquid but also most highly valued and potentially overvalued companies. Passive management has sought alternative selection and weighting schemes, while yet sticking to rules-based approaches. Algorithms implement the rules and at regular intervals, say monthly or quarterly, the target portfolio is rebalanced to follow the rules, selling or buying securities or adjusting weights. A trade can then be run algorithmically to realise the target portfolio

ii. Systematic risk

As an evolution of passive investing, systematic risk factors or risk premiums have been identified. These are repetitive characteristics in the capital markets. Risks here are simply patterns that one can find when looking at market data, how did the market react? What events caused the reaction?

These find their origin in institutional constraints on investors and in psychological patterns of behaviour which have been observed for long stretches of time. The factors received names such as growth versus value, momentum, carry, liquidity, the low beta anomaly... Identification of risk premiums is a quantitative search and validation process. These are all names for strategies resulting from the identification of the risks.

The quantitative investment strategies have generated the search, back testing, definition and implementation of systematic risk factors. Risk factors are long standing trends or anomalies in the markets which have received names such as growth versus value, carry, merger arbitrage, momentum, term structure. They are anomalies from simplistic capital market models. Many of these anomalies have been observed for a long period of time, but some are typical of certain capital market regimes.

Successful implementation of risk factor strategies depends on rigorous quantitative methodology and controlling transaction costs. Meaning, to confirm a strategy, one will have to run multiple analysis and correlation methods to make sure that the algorithm will perform well by following this strategy.

The positive return on risk factors is believed to be a reward for the intrinsic risk the strategies face. Investors capable of investing across the cycle are remunerated for accepting the risk. Since not everyone can pursue these strategies, a positive remuneration is persistently gained over the cycle.

The successes of academics in risk factor research have led to sophisticated investment managers constantly searching for and refining systematic risk factors from market data in the past and now also relying on alternative data sets.

Mathematical and statistical skills as well as data gathering and cleaning are critical for success in identifying and running money in systematic risk factors. The search is more effective with a solid IT development platform and computational expertise. In this way, quantitative approaches can be credibly back tested and sometimes explained with an economic or financial rationale as to why the approach might have worked in the past and plausibly hold up in the future.

iii. Active management

At the other end of the spectrum, active management of portfolios based on qualitative approaches driven by macro-economic research and company analysis result in portfolios that are not per se quantitatively defined. Active investment managers will structure concentrated portfolios with a small number of intensely researched bets. Yet even these portfolios require effective trading with a focus on execution cost and timely execution.

iv. ETFs

ETFs - exchanged traded funds - wrap a passive management strategy into a single shareholding all the benchmark constituents in one. The ETF share can conveniently be traded on exchange as any other share.

v. Arbitrage Trading

Algorithms might attempt to identify arbitrage opportunities and trade on those. An arbitrage is defined in theory as a risk-free opportunity, such as selling the same currency high to one client while buying it low from another client at the very same instant. The spread is pocketed immediately without having incurred an open risk on the currency. In practice, opportunities nearly free of risk are also called arbitrage opportunities. Such as buying a future on one exchange and selling the same future on another exchange. The risk is limited here to the credit quality of the two clearing houses backing the futures of the two exchanges.

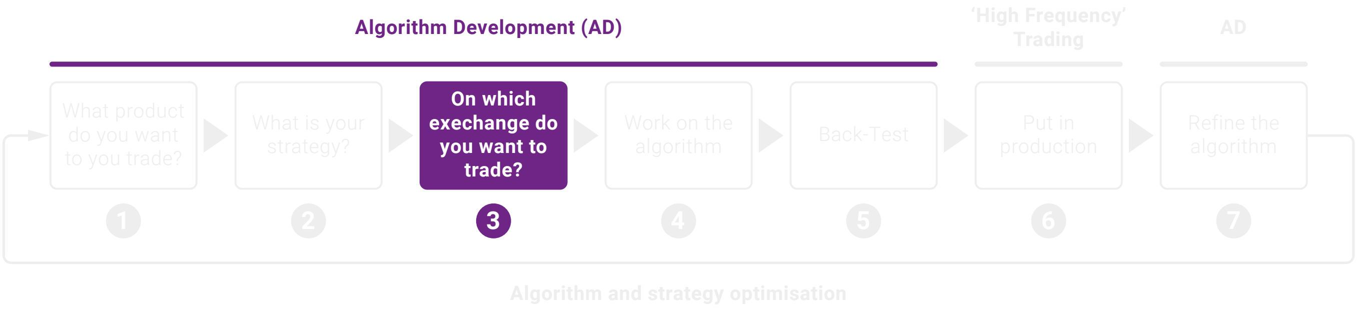
vi. Factors to consider

Regulations require quasi instant reporting of trades for transparency and fairness. In the EU the latest standard for trade reporting has been embedded in MiFID II, an EU and EEA wide directive. Trade reporting following MiFID II has commenced in 2018. The US has long-standing regulations on equity market trading formulated and enforced by the SEC. What results is better investor protection against insider trading.

An increasing variety of **automated support tools** cover every aspect of trading. Just an excerpt of software tools that have seen the light: automated supervision for insider trading based on voice and text recognition of suspicious words or conversations, reporting and publication of transactions within tight deadlines, implementation of a variety of trading enhancements such as randomisation of transaction size and timing for hiding a large trade, pacing the execution to a trading benchmark and synchronising buys and sells to keep a trade market neutral.

More recently quantitative and algorithmic investment have taken up **machine learning**, where structural models are not in use for lack of theoretical understanding or apparent lack of goodness of fit. Deep neural networks find positive trends to buy into and negative trends to sell. The successes of some machine models have triggered a renewed interest into the basic question as to why strategies have worked, why they should work in the future and when they might stop working. Interpretability of machine models is not yet obvious. Crowding out of strategies with too many investors might signal when a strategy is becoming stretched.

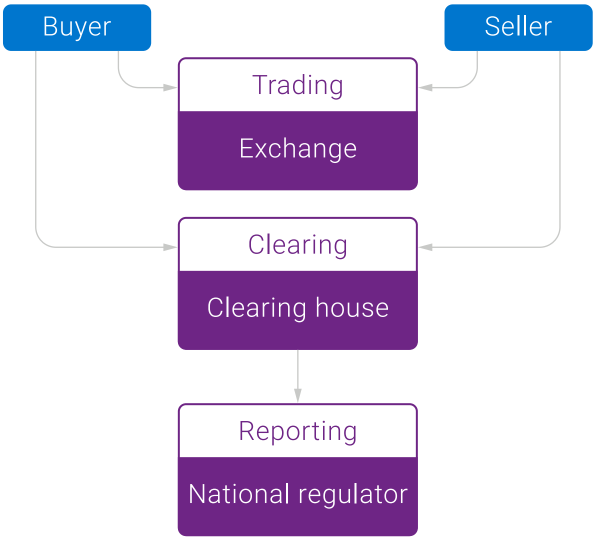
4. Markets and market structure


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exechange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

What type of data may be available and how a trade can be implemented is deeply influenced by market structure. So, what is market structure about?

A market structure provides the structure for a trade to be executed. Here is the flow a trade generally follows:
1. Both the buyer and seller agree on a price
2. The trade is being 'cleared' checking that the buyer actually paid
3. The clearing house reports the trade to the regulator

All of this is being done automatically and electronically in the backend. The differences in structure in markets is defined by what product (see financial products in the introduction) one trades. This section will cover each product with its respective market structure and regulation.



i. Exchanges

Exchanges promote transparency of market information and better liquidity in trading. They are the trading venue of choice for publicly traded securities. Exchanges are well-developed for equity, trading shares in companies, and for the exchange traded funds, which are unambiguously defined portfolios, wrapped together to trade as a single share. A bid book and an ask book are maintained by the exchange, containing all live orders at their offered prices. Where the bid book and the ask book meet at the highest bid and the lowest ask, is the market price at which all orders are executed.

Algorithmic and high-frequency traders routinely access exchange trading through executing programs. Program trading slices and dices a larger order into small, bite-sized chunks which execute a small portion of the intended trade over multiple orders. The small orders are anonymously drip fed into the market, to limit the adverse price impact a single large block order could trigger. Portfolio trading executes program trades in many securities on multiple exchanges at once. Buy and sell orders of the client are executed simultaneously in the portfolio trade, which renders the trading neutral to market direction.

Price variations in the market level the impact of buy and sell orders in opposite directions and roughly even out. Randomisation of sizes and timings are applied to hide the actual full order from the market. The randomisation is an effective strategy against front-running of a trade by opportunistic traders or market makers.

Popular benchmarks for program trading have been TWAP or better still VWAP. A time-weighted average price (TWAP) over the course of a trading day seems a sensible benchmark for execution a trade through the day. Since the size of trades is published, a volume-weighted benchmark (VWAP) became more popular.

Algorithmic execution can then trade around the VWAP benchmark. In a downward trending market, sales could be sped up and buys slowed down to benefit from the price trend. The same rule could be applied to individual shares rather than the market as a whole. Unexpected liquidity injected into the buy book could be met with an increased rhythm of sales for the trade. Further rules can be devised and executed algorithmically. An experienced trader can make judicious choices as to which algorithm to apply to which trade.

ii. Where futures trade

Futures also typically trade on futures exchanges. Futures are standardised derivatives with a clear and unambiguous underlying security, often an equity index, and a well-defined pay-off on set maturity dates. Unlike equity, futures have a short lifespan until maturity and final settlement, but they can be traded like equities. A future might mature say on the third Friday of every month, settle and expire, after which a trade can roll into the next month’s future.

There are liquid futures in equity indices, government bonds of standardised maturities for the more important sovereigns, such as the 10Y US Treasury Bond, and for the major currencies pairs such as EURUSD and commodities such as West Texas Intermediate crude oil. Consequently, futures are widely traded both for speculative and hedging purposes and are extremely liquid as all trades are collapsed and transparently executed in just a few standard instruments.

Futures are cleared through clearing houses, which eliminates the direct counterparty risk between the original two trading counterparties.

iii. Where Forex trade - Foreign Exchange Market

Currency trading executes off exchange, through market makers who commit their balance sheet to manage an inventory of the main currencies and at their own profit and loss take positions in those currencies. Most large banks engage as market makers in currency trading as a service to clients but also a source of profits for the bank. Irrespective of the market structure, electronic execution venues post prices from the banks in the major currency pairs and permit automated electronic execution, confirmation, reporting and settlement. Consequently, algorithmic trading has also found its entry into currency trading.

iv. Where corporate bonds trade

Corporate bond trading follows the pattern of currency trading, with market makers, typically investment banks, selling from or buying for their inventory. Corporate bonds are a very large universe, with multiple securities for most issuers and with regular redemptions and fresh issuance of corporate bonds.

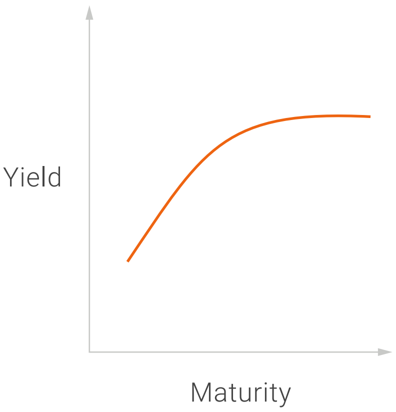
Liquidity is not guaranteed as no single investment bank may have the appetite or the inventory to make a market in each title. Yet, here as well, electronic execution venues supplement and progressively replace voice trading over the phone. Market makers chose to advertise their activity in certain titles on the platform, without for that matter entirely disclosing their inventory on the electronic trading venues. On occasion, a trader will have a “trading axe”, such as trying to clear a large inventory on behalf of a strategic seller. Such a trader may announce bid and ask prices but the ask price will be attractive and the bid price uncompetitive. But at the venue, they do commit to certain volumes and prices in select titles. This enables algorithmic trading to take place.

Trading in sovereign bonds is supported by market makers as well. Sovereigns tend to set high standards on the commitment to transparency and liquidity of market makers in government debt. The sovereigns typically require support in placing primary issues and in secondary trading of the government bonds from the market makers. Sovereign bonds are issued in large sizes and often tapped, meaning additional fresh issuance in existing maturities adding to size and liquidity of the bonds.

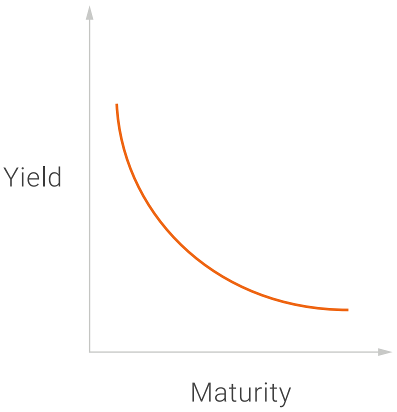
The yield curve, stipulating yields for different maturities, is normally well populated across the entire range of maturities from discounted Treasury bills below a year to coupon bearing Treasury bonds maturing 30Y to even 50Y in the future. Electronic trading in sovereign markets is standardised, sophisticated and very liquid even for large transactions. Trades in bonds are associated with trades in government bond futures on futures exchanges, delivering at maturity the cheapest-to-deliver 2Y, 5Y and 10Y bonds.

Inverted yield curve

The inverted yield curve is a graph that depicts, long term debt instruments yielding fewer returns than the short term.



Normal yield curve



Inverted yield curve

v. Where OTCs – over the counter - trade

Over the counter derivatives, such as interest rate swaps, currency forwards, credit default swaps and cross currency swaps, have also moved towards electronic venues. The over the counter derivative is in principle a freely negotiated agreement between the two counterparties. Thus, it can be formatted as best fits the purpose of a transaction, such as hedging a position on the balance sheet. The derivative should closely match the underlying instrument, maturity and other features to provide the optimal hedge.

Hedging explained with an example

1 A farmer prepares a wheat crop

A farmer purchaseees fertiliser, fuel, seed and everything else necessary to grow a wheat crop.

2 Farmer sets target price for harvest

Based on all his costs, the farmer determines a price he'd like to get for the wheat when he sells it to a local bakery at harvest time.

3 Farmer and baker consider price fluctuations

The farmer is concerned that wheat prices will go down, and he won't make enough to cover his costs. The baker is concerned that wheat prices will go up, and he'll have to raise prices.

4 Farmer and baker use a hedge to reduce risk

The farmer and baker agree in advance to a set price for the wheat, regardless of the market price at harvest time.

5 The hedge manages risk

By creating a hedge, the farmer and baker managed the risk of fluctuating wheat prices. If the market price at harvest is higher than the set price, the baker benefits from the hedge. If the price is lower, the farmer benefits. In either case, the hedge protected both against the potential for serious losses.

Nonetheless, standardisation of the common contracts has been helped by the International Swap Dealers Association (ISDA) with template ISDA agreements defining the most common derivatives transactions, leaving only a few specific terms to be filled out in the ISDA template. Once those terms are filled out, the derivatives contract is legally perfect. Counterparty risk is minimised with Credit Support Annexes or CSAs.

vi. Repos - repurchase agreements

To settle transactions requires cash. Cash in the age of quantitative easing is at best low yielding and in the Eurozone loses you money with a negative interest rate. So, while you need cash on a daily basis, you want to hold the least inventory. Management of ready cash for the bank, the insurer and the asset manager called the repo market into existence. In essence repos, short for repurchase agreements, are short-term loans of cash at very cheap interest rates. To secure the cheapest rate, the borrower sells the lender a liquid asset. At the end of the loan, the asset is bought back when the borrower returns the cash loan. Repurchase agreements settle immediately, or in transaction plus zero days, to enable smooth cash management with immediate availability of the loan. Liquid assets such as government bonds or corporate bonds of the highest credit quality are typically the easiest and cheapest source of access to instant cash.

vii. Clearing

Clearing is today common for most contracts, with an initial margin up-front and variation margin through the life of the derivative adjusting the initial margin depending on market moves. The margins form collateral which secures that both counterparties live up to the contract, even in case of a default. The margins are nowadays collected and held as collateral by clearing houses for most derivatives. Electronic execution is now also well underway in many over the counter derivatives.

viii. Market rules - regulators

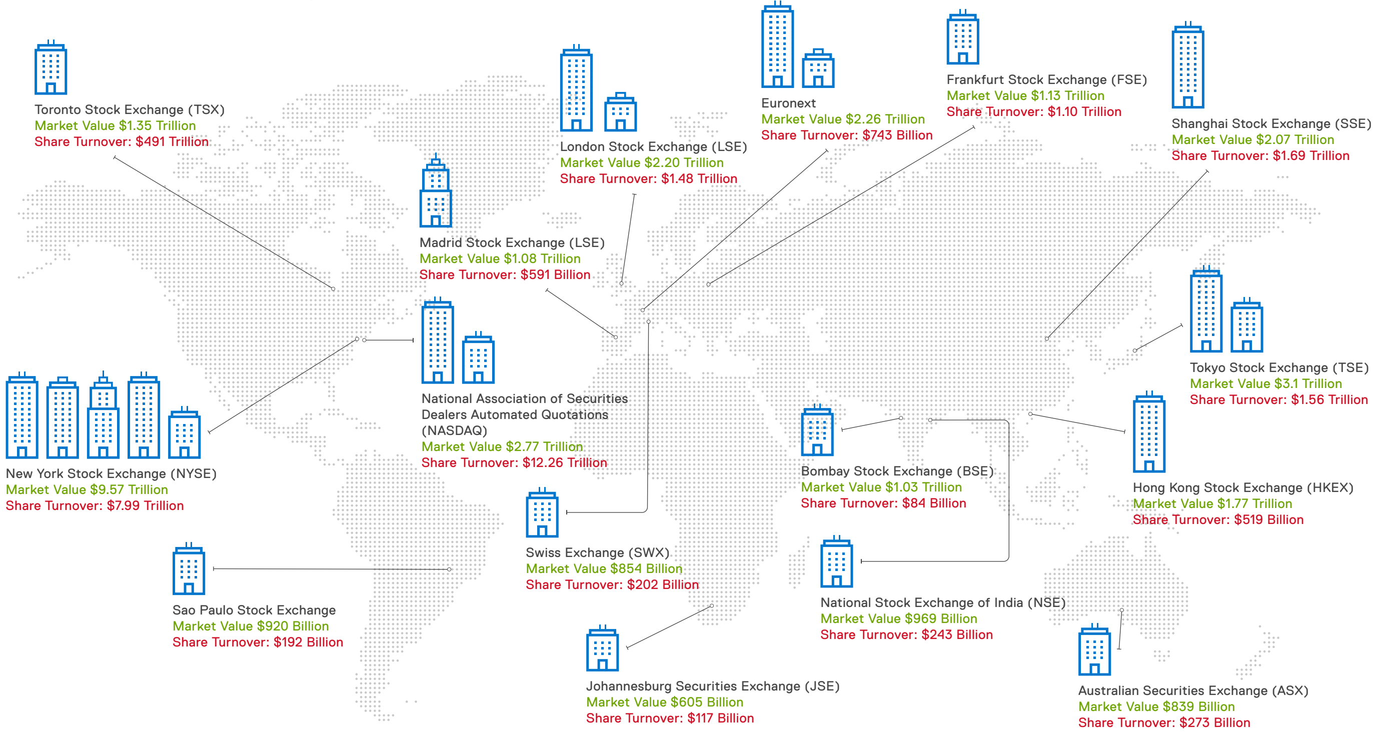
In the most recent incarnation of market rules in the European Union (MiFID II, a rewrite of original MiFID), market makers are referred to as systematic internalisers (SI). Between electronic exchanges with order books, where buyers and sellers trade directly with the order book, i.e. with each other, and supported liquidity from market making systematic internalisers are the electronic venues. The multi-lateral (MTF) and organised trading facilities (OTF) offer transparent liquidity from professional participants such as investment banks. The MTF and OTF can be consumed and transacted electronically.

Note this MiFID terminology (SI, MTF, OTF) is not in use in North America though, certainly for the time being it is the largest capital market, and by far the largest market for equities in publicly traded companies. The Securities and Exchanges Commission (SEC) regulates the securities markets and the Commodity and Futures Trading Commission (CFTC) regulates derivatives transactions in the United States.

World Stock Exchanges

Billions have been lost in the stock market since the recession started.
Where is all that money being traded?

Large buildings represent approximately $2 Trillion



Toronto Stock Exchange (TSX)
Market Value $1.35 Trillion
Share Turnover: $491 Trillion

London Stock Exchange (LSE)
Market Value $2.20 Trillion
Share Turnover: $1.48 Trillion

Euronext
Market Value $2.26 Trillion
Share Turnover: $743 Billion

Frankfurt Stock Exchange (FSE)
Market Value $1.13 Trillion
Share Turnover: $1.10 Trillion

Shanghai Stock Exchange (SSE)
Market Value $2.07 Trillion
Share Turnover: $1.69 Trillion

Madrid Stock Exchange (LSE)
Market Value $1.08 Trillion
Share Turnover: $591 Billion

Tokyo Stock Exchange (TSE)
Market Value $3.1 Trillion
Share Turnover: $1.56 Trillion

National Association of Securities Dealers Automated Quotations (NASDAQ)
Market Value $2.77 Trillion
Share Turnover: $12.26 Trillion

New York Stock Exchange (NYSE)
Market Value $9.57 Trillion
Share Turnover: $7.99 Trillion

Bombay Stock Exchange (BSE)
Market Value $1.03 Trillion
Share Turnover: $84 Billion

Hong Kong Stock Exchange (HKEX)
Market Value $1.77 Trillion
Share Turnover: $519 Billion

Swiss Exchange (SWX)
Market Value $854 Billion
Share Turnover: $202 Billion

Sao Paulo Stock Exchange
Market Value $920 Billion
Share Turnover: $192 Billion

National Stock Exchange of India (NSE)
Market Value $969 Billion
Share Turnover: $243 Billion

Johannesburg Securities Exchange (JSE)
Market Value $605 Billion
Share Turnover: $117 Billion

Australian Securities Exchange (ASX)
Market Value $839 Billion
Share Turnover: $273 Billion

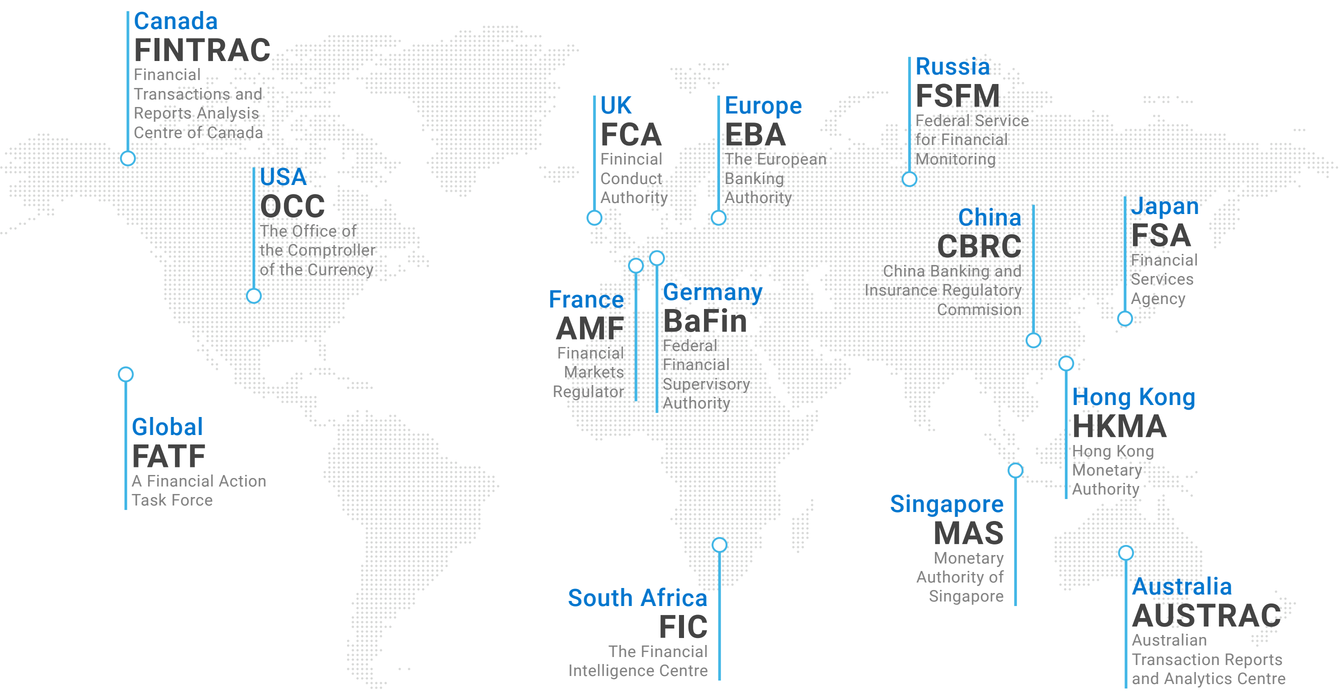
This infographic depicts 16 of the world's largest stock exchanges. There are approximately 100 major exchanges in the world in total.
The data in this graphic is current as of May 2009. Sources include world-exchanges.org and sec.gov.

5. Regulations in the capital markets

Trading in financial instruments is heavily regulated and requires prompt reporting, for investor protection, for creating transparent markets and for regulatory monitoring and statistical purposes. In the US, the SEC and CFTC regulate trading in securities and derivatives whereas in Europe the European Union has set standards in the second Markets in Financial Instruments Directive (MiFID II) implemented in practice by the national regulators such as the UK PRA and the German BaFin. Prompt publication of transactions within tight deadlines aims to promote market transparency. Yet the volumes of transaction reporting quickly reach the “big data” threshold, such that the reporting deadlines can only be met through extensive automation.

List of Regulators

Around the world



Canada
FINTRAC
Financial Transactions and Reports Analysis Centre of Canada

USA
OCC
The Office of the Comptroller of the Currency

Global
FATF
A Financial Action Task Force

UK
FCA
Finicial Conduct Authority

Europe
EBA
The European Banking Authority

France
AMF
Financial Markets Regulator

Germany
BaFin
Federal Financial Supervisory Authority

South Africa
FIC
The Financial Intelligence Centre

Russia
FSFM
Federal Service for Financial Monitoring

China
CBRC
China Banking and Insurance Regulatory Commision

Japan
FSA
Financial Services Agency

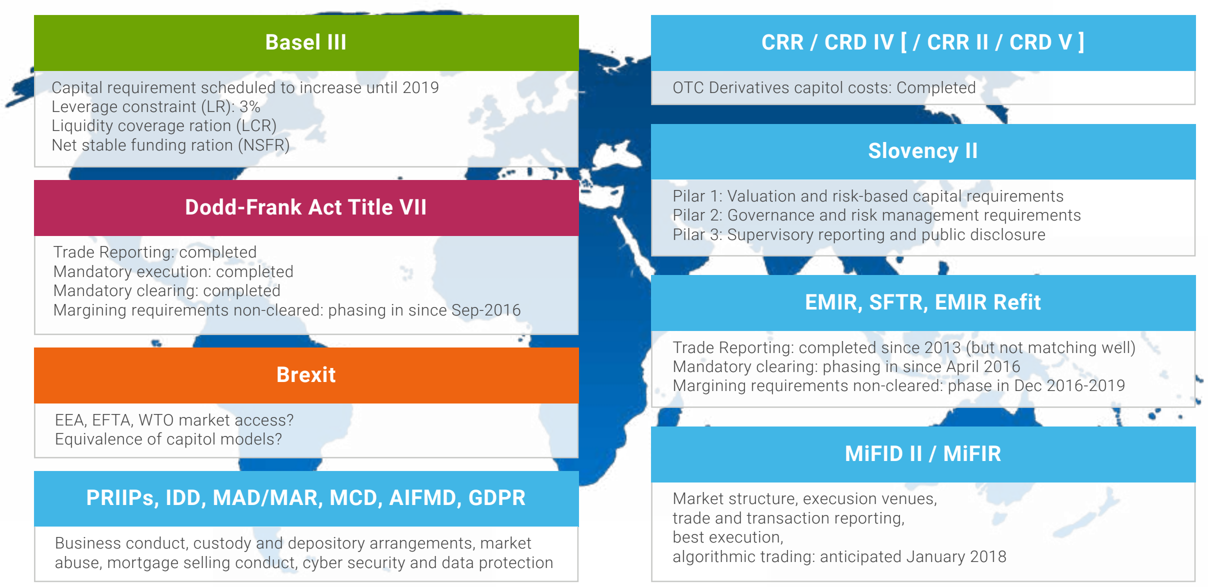
Hong Kong
HKMA
Hong Kong Monetary Authority

Singapore
MAS
Monetary Authority of Singapore

Australia
AUSTRAC
Australian Transaction Reports and Analytics Centre

List of Regulations

Around the world



Basel III

Capital requirement scheduled to increase until 2019
Leverage constraint (LR): 3%
Liquidity coverage ration (LCR)
Net stable funding ration (NSFR)

Dodd-Frank Act Title VII

Trade Reporting: completed
Mandatory execution: completed
Mandatory clearing: completed
Margining requirements non-cleared: phasing in since Sep-2016

Brexit

EEA, EFTA, WTO market access?
Equivalence of capitol models?

PRIIPs, IDD, MAD/MAR, MCD, AIFMD, GDPR

Business conduct, custody and depository arrangements, market abuse, mortgage selling conduct, cyber security and data protection

CRR / CRD IV [/ CRR II / CRD V]

OTC Derivatives capitol costs: Completed

Slovency II

Pilar 1: Valuation and risk-based capital requirements
Pilar 2: Governance and risk management requirements
Pilar 3: Supervisory reporting and public disclosure

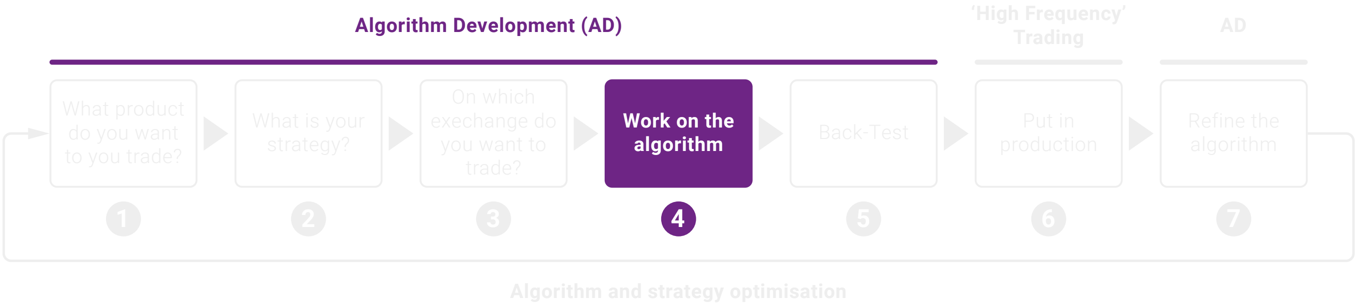
EMIR, SFTR, EMIR Refit

Trade Reporting: completed since 2013 (but not matching well)
Mandatory clearing: phasing in since April 2016
Margining requirements non-cleared: phase in Dec 2016-2019

MiFID II / MiFIR

Market structure, execution venues,
trade and transaction reporting,
best execution,
algorithmic trading: anticipated January 2018

6. Financial computation


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exechange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

Let's briefly review the most common mathematics employed by quantitative researchers and developers in the financial markets.

Fixed income mathematics involves yield curve models implied from markets through bootstrapping. Swap markets or government bond markets can provide the risk-free yield curve in a given currency for assets with minimal default risk. Credit-risky bonds are then priced using an additional credit spread over the risk-free yield. Bootstrapping is an efficient mathematical approach to determine risk free and credit-risky curves, starting with the lowest maturities and progressively stepping up to the higher maturities, until a full curve covering all maturities has been built.

i. Monte Carlo simulation

Monte Carlo simulation is widely utilised for options pricing, for risk management and for capital computation. This provides an immense opportunity to apply parallel power and mathematical sophistication. Alternative approaches to options pricing are partial differential equations (one or multi-dimensional depending on the number of risk factors) or Fast Fourier Transforms, however, these need and can benefit from high performance computation to a similar degree.

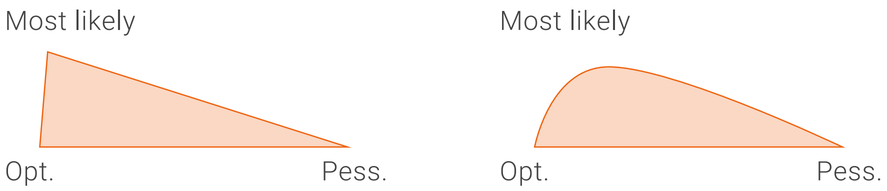
Invariably computational power is in high demand and for some applications, live updating from market prices is important so that computing will go on throughout the trading day. Monte Carlo scenarios sketch out possible future paths for the markets step by step, for all relevant risk factors. Risk is then inferred from the envisioned adverse scenarios mapped out.

The scenarios are constructed for the main parameters such as yield curves, credit spreads, equity returns, currency quotes for a range of future time steps. At each projected time step, the portfolio is evaluated using these parameters and its likely evolution is mapped and classified in statistical percentiles.

Applying Monte Carlo scenarios and revaluations to the entire balance sheet of a financial institution performs a capital computation for the bank or insurer. Applying Monte Carlo scenarios to the trading book of an options trader determines prices and risk of the trading desk.

Since Monte Carlo simulations are simply computing many random scenarios (called paths) to then average results, it can be easily parallelised. Using CPUs to generate 1,000,000 of these paths (the standard minimum to get acceptable results) is simply not efficient enough. GPUs on the other hands are able to parallelise these paths, reducing computational time.

How Monte Carlo works in MSP



- Apply a distribution model to ‘risky’ activities.
- Enter multi point estimates for time and/or cost.
- Run an ‘iteration’ where randomly generated estimates are selected for each activity based on the distribution model.
- Record projected finish date or budget and repeat many times (500-5000 iterations).
- Analyse the resulting finish date or budget distribution curves to determine high confidence schedules or budgets. (TypicallY 90%)

With an extreme degree of parallelism in computing the multiple scenarios, Monte Carlo fits ideally to high-performance computing platforms. Basic Monte Carlo problems are embarrassingly parallel, but some variations such as least-squares Monte Carlo can fit from information exchange between the scenarios. Best performance is achieved if the cores can be flooded with computational demand to the highest degree. Communication and memory need to be thought through for the best performance.

ii. Mathematical optimisation

Mathematical optimisation is deeply embedded in financial computing. Two major applications are calibration of volatility surfaces of assets and asset classes aiming for the best fit between model and market and improving risk-return characteristics of portfolios.

Calibration of options models requires setting up a model for implied volatility, the main pricing parameter for options. The volatility models are two-dimensional: maturity of the options plays a role but so does exercise price of the option, called strike.

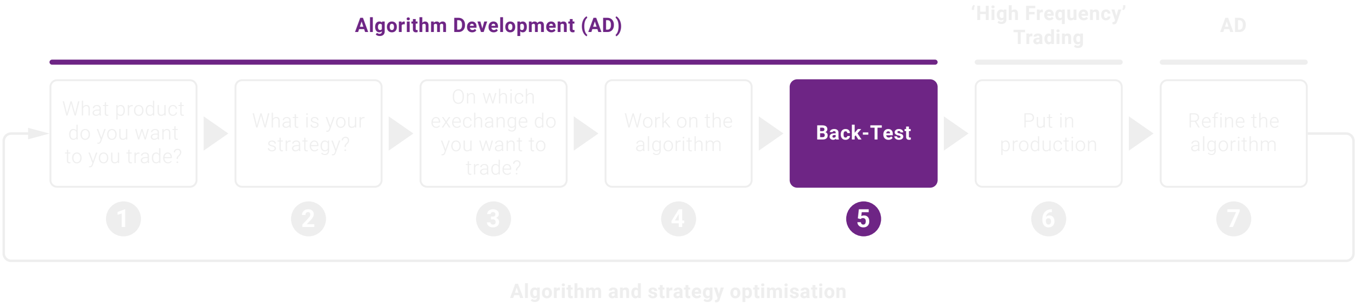
iii. Portfolio optimisation

Sequential quadratic programming and dynamic programming are applied to portfolio optimisation. A risk return model needs to exist for the portfolio: the risks could be imported from volatilities implied by the options markets whereas the return model could be a time-series, factor analytical or machine learning model.

Portfolio optimisation is a large dimensional problem. Selecting which securities to include and selecting the weights for a balanced portfolio are the two basic dimensions to decide. More advanced approaches to portfolio construction identify not only composition of the portfolio but also evolution across multiple time steps, with trading allowed. Dynamic programming leads to longer compute times and more demand on the sophistication of the software and hardware.

The choice has always been mathematical sophistication or brute force computation. A combination of both may be the way forward. The route of mathematical sophistication can be highly effective but could port across from platform to platform with difficulty, as compromises are embedded in the implementation, potentially leading to subtle dependencies. Therefore, explicit awareness of the compromises made should be part of the development plan of a financial software library.

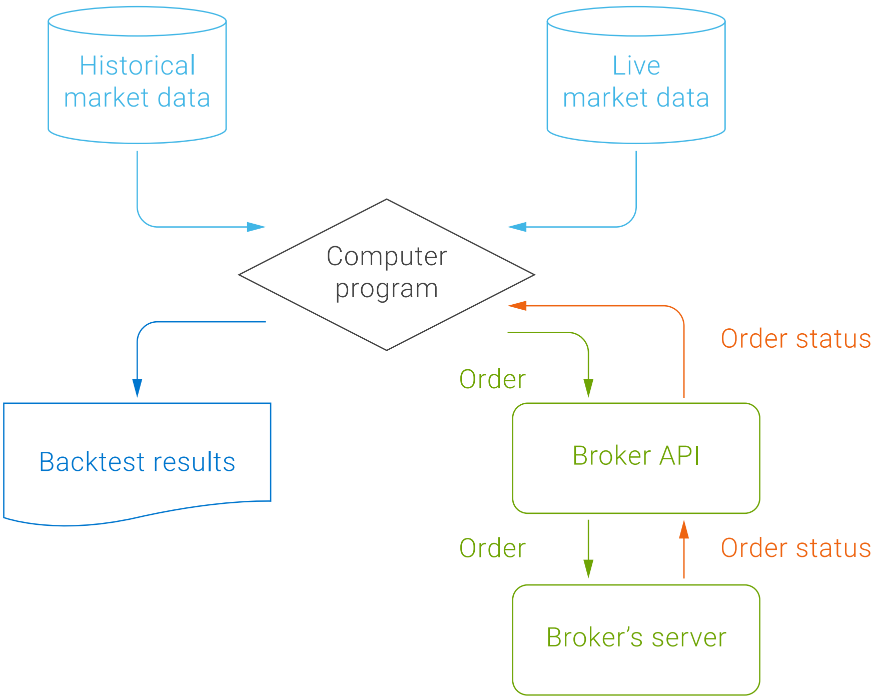
iv. Back-testing


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exechange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

As previously introduced, back-testing is key in algorithmic trading to confirm the strategy on historical data. This is quite simple on a higher level, but in reality, it involves aggregating a lot of data to test the algorithm accuracy.

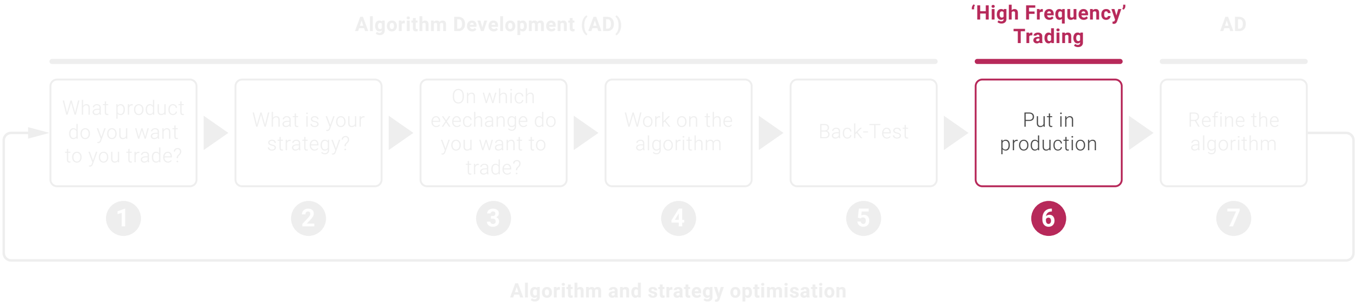
We could make the comparison of a machine learning (ML) algorithm being trained, back-testing being the training phase. Whilst the high-frequency trading part is when the ML model is put in production.

A description of the process of back-testing



With newer deep-learning techniques such as VAEs (Variational Autoencoders) and GANs (Generative Adversarial Networks), back-testing can be done by simulating scenarios. This gives the added advantage of running the model on future prediction rather than past market conditions.

v. Trading


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exechange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

Trading speed used to be the main differentiator for traders, reaching the lowest latency meant placing trades before the market updated its information, this is commonly referred to High-Frequency Trading.

Accessing data ahead of market participants meant being able to see price changes and reacting accordingly before anyone could do so. This created much controversy towards high-frequency trading.

Lowest level of latency could be achieved by;

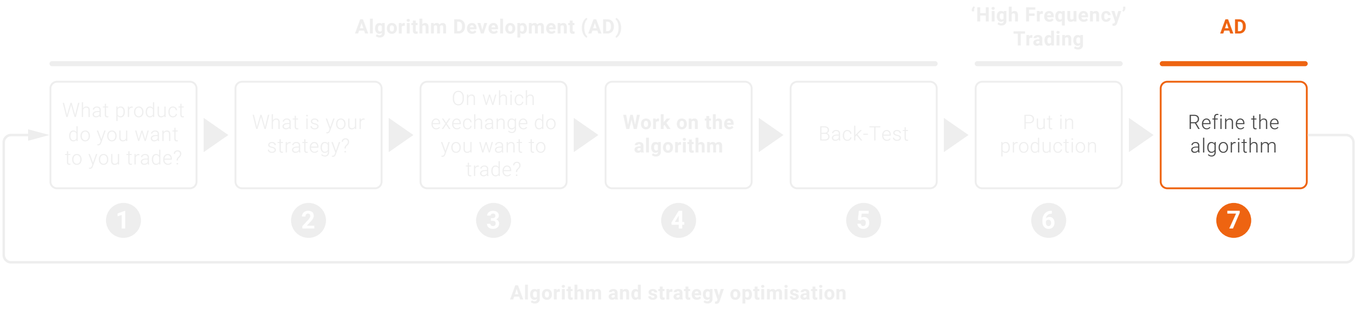
- Reducing the distance between the exchange and trading platform through co-location: housing the trading platform in the same building of the exchange
- Direct connections: using cables between exchange and trading location without intermediaries
- Specific hardware such as;
  - Networking Cables such as InfiniBand
  - NIC cards that bypass server kernel such as SolarFlare
  - FPGAs

High frequency traders are ultra-high-speed execution algorithms where in a rapid sequence of buying and selling, the trading itself is the source of profitability. High frequency traders do not keep large net exposure to the markets for any longer period of time. High frequency traders are tolerated since they support liquidity in the financial markets.

The rise of high-frequency trading has gained much interest and has consequently resulted in higher entry barriers. Indeed, reaching the lowest latency can only be attained by a few, not a majority.

As a result, the current trend is shifting towards gaining a competitive advantage through the ability of crunching more data and using smarter algorithms. Yet, trading platforms will remain as a collection of hardware and techniques that aim for lowest latency.

vi. Short Shelf Life


Algorithm Development (AD)
'High Frequency' Trading
AD
What product do you want to you trade?
What is your strategy?
On which exchange do you want to trade?
Work on the algorithm
Back-Test
Put in production
Refine the algorithm
1
2
3
4
5
6
7
Algorithm and strategy optimisation

Once the algorithm has been put in production and starts executing trades, there will be a reaction from the competition. As a result, algorithms have a short shelf life, i.e. they can be put in use for so long.

This significantly put stress on the hardware infrastructure, since the algorithm cannot be static anymore and must adapt continuously to new input to remain relevant.

As such, the backend infrastructure must accommodate for live-data feed and quick processing of large amount of data. Databases must be able to live – or in near real-time – feed the compute engine to update the algorithm.

This creates a loop of data movement and processing that must remain secure and reliable, i.e. experiencing near to zero down time.

7. Trends - where's the industry heading?

i. Artificial intelligence

This has hit mainstream financial services as well as FinTech. Initial applications have been in customer relationship management where, the financial sector could piggyback more innovative sectors, and in credit analysis, credit scoring and lending, where data has always been critical in decision making. Investment management, trading and risk management have now followed. The FinTech sector heavily relies on artificial intelligence to find new ways of distributing, transacting and managing finance.

Artificial intelligence has opened the door to utilising new, unstructured data without having to develop theoretical approaches before being able to make sense of the data. Text recognition for treating information flows and deep neural networks replacing explicit mathematical models are part of the quantitative and risk management toolkit.

ii. Cryptocurrency

Cryptocurrency arrived in financial services as an alternative to central bank supported fiduciary money. Its use has so far been speculative rather than transactional. A banking system has not developed around any of the cryptocurrencies, payment tokens mutually accepted in certain online communities, although they have been instrumental in the growth of electronic wallets which are now also used for regular currency.

The regulated financial services industry has picked up blockchain technology from cryptocurrencies, to establish distributed ledgers. The R-3 consortium has enabled standardisation across firms, an essential ingredient for scaling up and determining robust applications in real-life finance. Experiments in creating purely electronic assets tradable based on electronic certificates, in ultra-rapid settlement of transactions and in immediate clearing and collateral operations, cutting the counterparty risk out of bilateral transactions, are also being attempted.

While the insurance and reinsurance industry are experimenting with “smart contracts”, which pay insurance claims automatically, once the contract terms trigger.

iii. Quantum computing

As of yet, quantum computing is still in its research infancy and the problem of mature application in financial services is not for now. Yet the promise that currently unsolvable NP-hard problems might be cracked quickly with quantum computing is, at least, attracting interest.

The most immediate financial problem follows from cyber security. Credit card PIN-codes are routinely secured with the RSA public and private key mechanism. In classical RSA is pragmatically safe as the solution requires polynomial time for longer keys. Quantum computing makes this a trivial task for the quantum Fourier Transform, as Shor proved with Shor’s algorithm. Consequently, the search for better computer security is no longer just a problem of theoretical interest. Applications in credit modelling and portfolio optimisation have also been investigated with quantum computing.

II. Technical

1. Importance of technology for success

Strategies for investing and trading are data-intensive and very computational and may have a limited successful shelf life. Consequently, these funds continually develop, implement and perfect algorithmic trading strategies on the go.

Quantitative researchers and developers, portfolio managers and risk managers compete with the next fund through innovation in investment and trading but also through speed of implementation. Developing code and back-testing are an ongoing activity and in the ideal case, a successful back test can be implemented in trading without delay. Ideally, the research code may serve for selecting and sizing the live trades as well.

Having larger databases enables more robust back-tests with higher statistical significance while having the computational capacity to run the back-tests quickly can speed up the development process and enable more confidence and a quicker launch of a trading strategy.

High-frequency trading comes down to a competition between technologies for speed of identification of trading opportunities, and once opportunities have been identified, speed of execution. Simply put, without a technological advantage, do not attempt high-frequency trading.

2. Algorithm development

Where do algorithms for trading come from? What investigations have been run and have generated useful trading or investment ideas?

Time-series analysis for market investigations has a long tradition. Typically, maximum likelihood estimations (MLE) fit parameters to attempted models. Price trends have been investigated using auto-regressive and moving average modelling (ARMA) for tracing autocorrelation over time. Volatility regimes of the market prices have been successfully fitted with auto-regressive conditional heteroskedasticity (GARCH). The two combined are an effective econometric description for many assets in capital markets.

Linear and non-linear regressions have underpinned systematic risk factor or risk premium analysis. Historically linear regressions attempted to make sense of market returns. Although there has always been an understanding that there is no logical reason as to why markets should behave linearly, non-linear analysis was more limited in mathematical and computational tools, until the advent of deep neural networks.

Machine learning and deep neural networks have made their entry into capital market investigations. The rationale being twofold: non-linear modelling is easily accepted in machine learning; furthermore, the model is self-building and does not depend on an explicit mathematical equation which has to be guessed successfully to get reasonable results. Artificial intelligence and even alternative data sets have made their entry into market investigations.

3. Technology review software

i. Programming languages

Languages of choice have been C++ for quantitative and trading software and increasingly Python for machine learning.

C++ was for a long time the only real contender for quantitative development, which many financial services companies did overwhelmingly in-house. Many of the quantitative packages from specialist financial software vendors, which in the past five years have progressively replaced in-house development, are also in C++. With the advent of .NET, some C# trickled into quantitative packages as well, but it is fair to say that virtually all quantitative developers see C++ as the cornerstone, with other programming languages needing to justify themselves against C++. The entire development environment of C++, including object orientation, operator overloading, templates and functional libraries, such as Boost, are routinely in use.

Python has been introduced into financial programming as a rapid development tool based on scripting, with more computationally demanding tasks either sourced through libraries or coded up in earnest in compiled and linked C++. With data science becoming very prominent in finance over the last five years, Python's role has increased dramatically, particularly because of the availability of tools and libraries in artificial intelligence and machine learning. Python has become the language of choice for data science.

R is in use where the methodology is primarily statistical, as R comes with extensive statistical libraries and accommodates large datasets. R is trained and promoted heavily in education to graduates with a statistical option, hence it's increasing its footprint in finance.

The data science world has also engaged in early experiments with Julia, but the development environment is still considered weak by the leading practitioners.

Besides freeware and manufacturer sponsored libraries, there are also some commercial libraries. In numerical computations, Numerical Algorithms Group (NAG) has dominated with ongoing supported development of new routines. NAG has moved from FORTRAN to C and C++ and provides additional support in CUDA and in Automatic Differentiation specifically for the financial industry with quality libraries and skilled consultants. For data science, there are mature implementations of the MapReduce programming model.

ii. Parallel programming languages

Languages of choice are C/C++ with sometimes MATLAB and Python as wrappers for compiled routines. Open MP is used with C/C++ and Fortran to automate the creation of parallel threads, releasing less expert programmers from the overhead. More expert programmers can resort to used of MPI to communicate between computing unites and distributed or shared memory.

iii. How to choose the right programming language

Choosing a preferred programming language is always a compromise, but the following criteria may help.

On the pragmatic level, is the developer base sufficiently large? This need to include the risk of employees moving on after having delivered their "masterpiece". Can the software be maintained when accounting for staff turnover? Documentation standards need to contribute to handle this risk.

Certainly, support from numerical, statistical and machine learning libraries is a leading criterion and may determine the selection. A commercial vendor such as NAG provides excellent numerical libraries in FORTRAN and C++ which through long and widespread use have become tremendously efficient.

Financial engineering libraries are available from a number of software vendors. QuantLib is the best known open source library. Commercial libraries such as FiNCAD, Numerix and Quantifi Solutions come with development toolkits making the libraries extensible. Some libraries such as Murex are locked systems with deep integration into the risk, portfolio management and trading environment.

In this camp, it may also be investigated whether automatic differentiation (AD) makes sense for the application. In AD, the speed-up is realised through judicious mathematical sophistication and cannot just be achieved with brute force hardware availability. Its use has increased materially in the last decade. Does the chosen language support AD and are there toolkits to turn to? Preferably check whether there are good interfaces with databases and good live access from and to electronic exchanges. Note that the exchanges will by themselves provide tools in support of distribution of their data and connection to the exchange for executing trades.

When thinking about the longevity of the software, can the piece of software move to the next best-performing hardware? How tight is the interaction between software, hardware and databases? Very tight interaction without the necessary level of abstraction or object orientation, may stoke up trouble in versioning to the next platform?

Finally, is a mix of programming languages the best option, as often proves the case today? Python scripting has expanded beyond its strengths in data science to the top-level business logic. The software is organised at a high level in a straightforward, easily understood and easily maintained script with the hard work being done at higher efficiency in a more explicit and compiled programming language. Compiled C++ where it matters delivers outcomes at high speed.

CUDA for GPU acceleration in the computationally heavy parts can be built into the package. Deep support with numerical and machine learning libraries with a large professional user base in academia and industry is valuable as you benefit from other folks' prior investment in the library.

4. Databases

Databases of choice have been SQL RMDBS for structured financial data for a very long time. SQL servers function well as multi-purpose databases as long as some logical structure – the SQL schema – fits the data well. Hadoop has been applied for alternative data sets such as online scraping, as unstructured data are hard to squeeze effectively into an SQL database. For financial time-series which are very prevalent in trading and securities data, KX has become the industry leader with an adapted time-series solution called Kdb+.

Adding to the typical market data, such as time series covering prices, volumes and individual trade "tick" data, the financial industry has expanded to alternative data sets. Investor sentiment applies text recognition to the worldwide online press. Other data are scraped from the web, such as complementing collection of prices from online stores to front run inflation releases. Aerial data supplement estimates of economic activity in shops, sectors or regions. Credit card data is subjected to data science to figure out trends in consumer behaviour and economic activity.

5. Compute

Hardware has traditionally been on-site, within the large investment banks and grid computing on demand. So far Intel®-based processors have predominated but accelerators are in use for specific purposes. GPUs have been used for computational enhancement through parallelism. In the GPU space, CUDA has been accepted as the de facto standard, with few people attempting development in Open CL. In high frequency trading FPGAs have furnished quick trade computation and quick connections to electronic exchanges. FPGAs have been hindered by the limited number of software developers calling FPGAs their home, although there is a Java-flavoured compiler for FPGAs.

Yet the cloud is definitely gaining traction and acceptance. Accelerators have also been made available in the cloud, with GPU cloud capacity easiest to find. The broad application of GPUs in a variety of scientific and engineering applications, by academia and industry, have led cloud providers to include GPUs in their offering. Clouds are a common approach to bursting with extra computational capacity and for some, the preferred approach to computing on GPUs, offloading hardware investment, operations and maintenance to external providers.

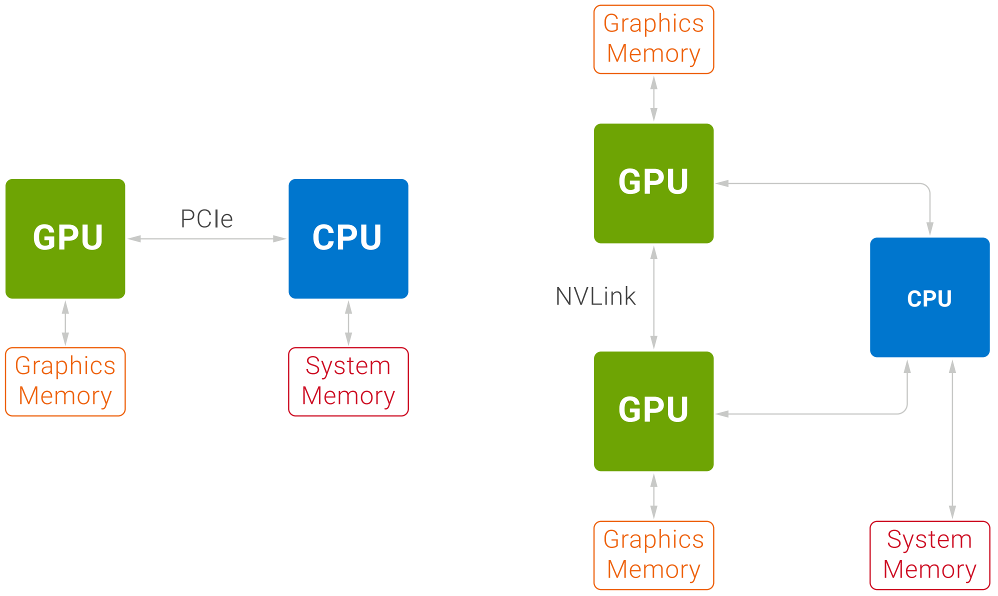
i. GPUs

Since the slowdown of Moore's law, computational throughput has sought new optimisations. Some optimisations follow mathematical methodology, but a more general principle is multi-core parallelism. Even CPUs come with multiple cores now. GPUs make the most of the parallelism by providing a massive number of cores in close contact with the CPU. For successful development the multi-core CPUs require multi-threaded software. Hence moving to the GPU CUDA programming environment is not necessarily a bigger cost than redoing CPU C++ software with more attention to threading.

GPUs will provide massive accelerations if the GPU cores can be saturated with computational demand. This requires the data and the results to be flushed to and from the GPU with adequate speed. In first line, investigate the PCIe bus to this end. A further optimisation is the more recent intra-core GPU communication for some applications according to a process called NVLink. Some algorithms in the data science field can make good use of intra-core communication and implementations are available.

NVIDIA® NVLink® is a high-speed point-to-point peer transfer connection, where one GPU can transfer data to and receive data from one other GPU. It is particularly beneficial when the application runs on multiple GPUs at the same time. Otherwise NVLINK doesn't bring the full performance enhancement.

CPU-GPU Connection via NVLink



CPU-GPU Systems Connected via PCI-e

NVLink Enables Fast Unified Memory Access between CPU & GPU Memories

i. FPGAs programming

Verilog/VHDL is close to the metal with a good control of the actual electronic instructions running but is notoriously difficult to code up. Fewer skilled FPGA programmers are around, and most are not familiar with financial markets, applications or programming. It is hard to code higher algorithms effectively in low level VHDL. Recently, a Java-flavoured compiler for dataflow programming in FPGA has become available from Maxeler with a number of investment banks and exchanges having applied FPGAs in financial computation.

Also, there are FPGA implementations of trading algorithms in high-frequency trading from investment banks and high-frequency traders, where the critical need for speed has convinced the bank to make this step into somewhat uncharted territory. Compilation of an FPGA program involves arraying gates mapping out the logic in semiconductor whitespace, which is a lengthy process compared to compiling for established microcircuits as exist on CPUs and GPUs. FPGAs have however remained a niche platform in financial programming nonetheless, mostly in specific applications in high-frequency trading where every microsecond counts.

ii. Virtualisation/ cloud

GPUs are now routinely available in the cloud from several well-known cloud providers, accepting CUDA code. FPGAs are sparingly available.

For moving to the cloud, it is important to not only think about the software and the computation but also the interaction with the database and speed to execute transactions on exchange, if this is where you compete. Clouds can be called flexibly for bursting demand and for independent failover back-ups.

Cloud computing is nevertheless reserved for the algorithm development phase. Indeed, the latency penalty that occurs when using cloud technology becomes infeasible for the trading or high-frequency trading phase.

iii. VDI Virtual desktop infrastructure

Virtual desktops built onto a common infrastructure have the benefit of providing an integrated hardware and software environment for teams of quantitative researchers, quantitative developers, portfolio managers and traders. In the absence of an integrated infrastructure with shared resources, there is a risk of building "private" environments which are locally optimised but do not permit collaboration across the firm. Over time, things invariably go wrong with staff turnover, poor versioning of successive efforts along with inconsistencies between desks and versions. The virtual infrastructure brings such conflicts into the open and eliminates private, ad hoc optimisations.

6. Networking

Whereas Ethernet is common for tying processes on multiple servers together and for passing data around, for critical high-performance applications, InfiniBand by Mellanox can be applied. This improves the speed of communication and data transfer. Data transfer from and to exchanges is critical for up-to-date prices and volume information. High-frequency traders have moved as far as co-locating their production servers running the trades at the exchange itself for minimal latency. Cost from the right licenses with the exchanges also has to be thought through as exchanges see themselves as data providers and charge for the service, with speedier live access costing more than delayed access.

Financial information exchange with exchanges or over the counter with trading counterparty has converged on the FIX protocol, now in release 5.0, for passing information. The physical transport layer is TCP/IP, IP or UDP.

7. Factors to take into consideration

i. Speed versus flexibility

These criteria are always present, but the respective stress varies across different parts of the industry. In development, flexibility is worthwhile, in production speed dominates, but at the end of life of software, flexibility and the capacity to version or transfer to the next platform come to the front again.

There is no true solution to this conundrum, other than raising awareness and being explicit about the conflicts that might appear through the life of a software package on a platform. An explicit register or map will help the Chief Information Officer and the development and service teams. Planning scalability and versioning ahead is best practice.

ii. High availability

Critical operations for financial services are trading, portfolio management, risk management, collateral operations and regulatory aspects such as reporting and capital management. For algorithmic and high frequency trading, an interruption of trading leads to a temporary suspension of the profit generation. But some of the other failures are potentially even more devastating. Risk management could be interrupted, leading to losses accumulating without monitoring or without the capacity to mitigate, such as acting on a stop loss or executing a hedge. An interruption of collateral posting when due, could lead to the clearing house closing out an essential transaction for you. Failure to report in line with regulations can trigger regulatory investigations, fines and losing a license to operate in a certain market. Clients might ask for compensation, retrieve funds or discontinue business altogether.

It speaks for itself that financial services organisations expect high availability from information technology and play through war games ahead of an actual emergency. Offering a failover solution by duplication at a second location or through diligent planning of the cloud is considered a plus in financial services.

It is mostly a poor development process and poor testing of a new service or new version that cause problems in financial services: preventable mistakes by the user. Protocols for moving between research, testing and production environment for moving software over and reversibility of the steps taken are part of best practices. Engagement of the Chief Risk Officer (2nd line) and internal audit (3rd line) for teams with a poor record can help. For technical breakdowns, such as power outages, interruptions of telecommunications networks and failing equipment, a failover facility in a separate location or in the cloud should also be considered.

8. Algorithmic trading environment

The research platforms often allow for more experimentation and are less rigorous in terms of testing and redundancy for handling fails. Optimising research productivity leads to differentiated environments from trading efficiency. Trading hardware and software will have passed through rigorous development and testing before going live. It is still beneficial for a quick roll-out of a strategy to have as much alignment as possible in hardware and software practices moving from research over development to portfolio management and trading. Younger researchers may be more accustomed to Windows NT, R and less to Linux, C++ than quantitative developers.

High-frequency trading rests on very high speed and high confidence in the accuracy of the code as there will be no human intervention before a trade executes. Co-location and use of FPGA pipelines is common in high-frequency trading.

III. How to size the best solution?

Dell Technologies and NVIDIA have over the years served the algorithmic trading industry by developing bespoke solutions, this section will provide a generic overview of their respective technology stack applied in the industry. To configure the best solution to your needs please get in touch with your local Dell Technology and NVIDIA representative.

1. Overview of the Algorithm Trading infrastructure

Before going into more details in the infrastructure, we will review the challenges, trading process and high-level overview of the IT infrastructure.

i. Challenges identification

The solution will attempt to tackle the following challenges;

- Aggregate a lot of data from multiple sources (live/ batch)
- Compute high amount of new data – to keep feeding an algorithmitic loop
- Deal with Unstructured and structured data
- Develop more complex algorithms over time
- Aim for lowest latency
- Have application failovers (high availability)

ii. The algorithmic trading process

Firstly, it is key to understand where and how the algorithm is being developed since there are several steps in terms of data;

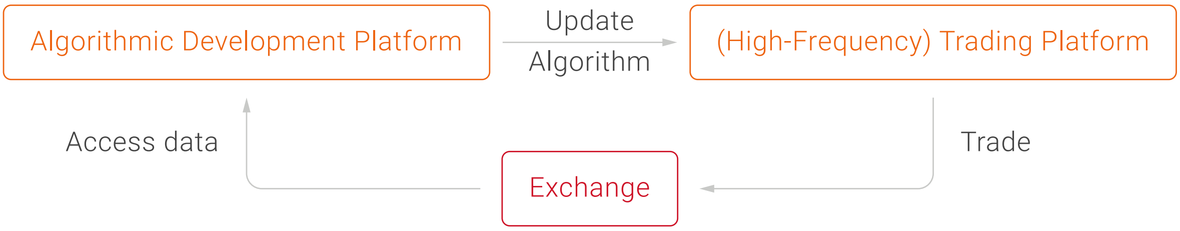
1. Design: designing or redesigning the algorithm if the simulation doesn't prove satisfying
2. Simulation on past-data: training the algorithm on historical data, the more data the better
3. Simulation on real time data: using a test environment, the algorithm is put against live data
4. Production: if the algorithm has passed all tests, it will be used and making profits/ loss

Hence, putting a trading strategy into production means to first run it against historical data and then against real data coming from exchanges. This simulation ensures that the algorithm behaves as expected – making profit – here back-testing.

This process can be done on one or several classes of assets which could yield different results accordingly. Regardless of the step or strategy, a massive amount of data must be used in a very low latency environment.

Additionally, this process is far from linear, it is most of the time – if not always – circular, i.e. it forms a loop and each slight change must update the algorithm, rerun and push for production again.

iii. Two environments: algorithmic development and (high-frequency) trading


Algorithmic Development Platform
Update
Algorithm
(High-Frequency) Trading Platform
Access data
Trade
Exchange

Let's beak-down the computational needs of platform. Similarly, to an artificial-intelligence process, the algorithm trading flow follows a training and production (inference) phase. Let's look at their own specificities;

The Algorithmic Development Platform needs to aggregate massive amount of data in parallel as quickly as possible. Therefore, databases that allow fast IOPS, ingest various data type and make it easy to query for languages such as Python, Spark, SQL or R are preferred.

The risk analysis or data-science models must be supported by powerful in parallel compute power, especially in the back-testing phase. Models' size can commonly fit on one or two graphic cards. Therefore, GPUs and CPUs with large number of cores and high-memory bandwidth are preferred.

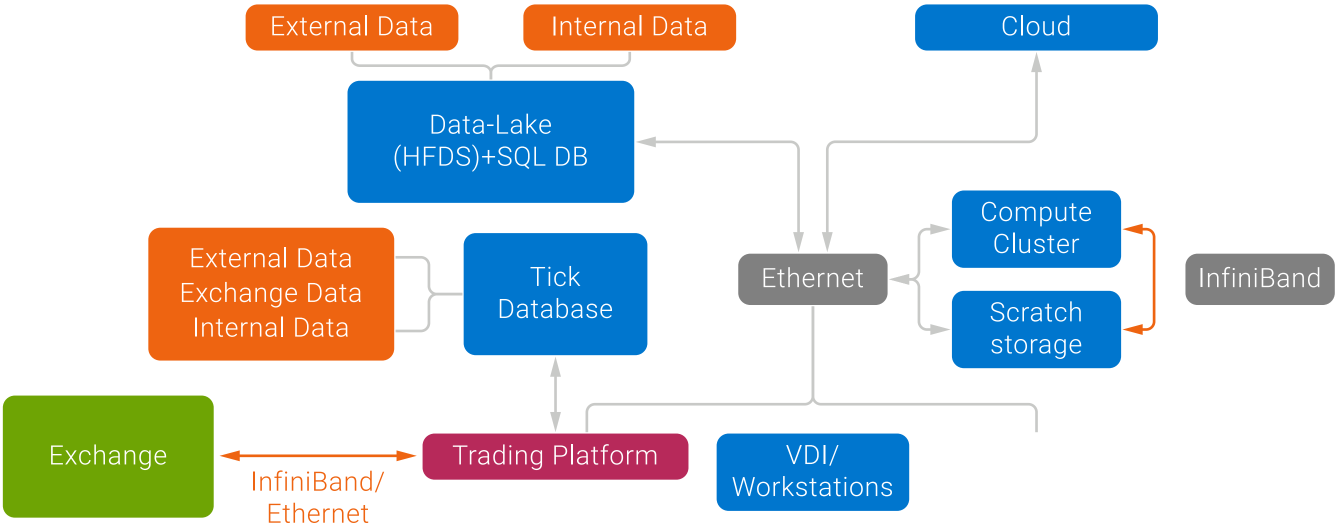
Since application downtime translates into pure profit loss, high availability of the platform is a prerequisite.

On the other-hand, for the trading platform, the algorithm is simply put in use, latency is key in this case since it must connect to exchanges as fast as possible, ideally before market and competition reaction. Since the algorithm has been optimised, it is similar to the inference phase of an AI model.

In this case, FPGAs or high-frequency over-clocked CPUs are commonly deployed for best performance.

Nevertheless, the need of a second environment for the trading part is not always necessary. Indeed, as some firms find their competitive advantage from the amount/ variety of data and their ability to process and develop better algorithms, latency has become less of a determinant factor As a result, the trading platform can be optional for some firms.

Overview of both platforms



2. Data sources types and movement

Data is the fuel of a well-performing algorithm, in the past having more data was enough, but now variety – of data type – and speed (ingestion and computation) are essential. Before moving on to designing a robust data-infrastructure, one must understand what kind of data is fed into models.

i. Exchange data

Tick-data and time series coming from the exchanges, they are structured but must be updated as quickly as possible to update the model. The most used data is by far price data, which is highly structured, it is most commonly called tick-data due to its 'tick' format.

In terms of amount, tick data created from exchanges has only been increasing. In 2013, the NYSE averaged half a billion trades and quotes per day, in 2018, there were approximately 4 billion trades and quotes per day with peaks going over 8 billion trades per day. As a result, the financial services industry now receives multiple terabytes of streaming tick data per day which further stresses the data-base infrastructure.

ii. External data

Comes in different forms, it can be historical data extracted from exchanges themselves or from third party vendors such as Bloomberg, Reuters and Trading-economics. Alternatively, news data is very important to be fed to the model, FED (Federal Reserve Board) announcement - or more currently Twitter feeds - greatly affect price movements, nevertheless this data often comes as unstructured.
Alternative data has made its entry as a differentiator, though due to its amount and diversity, leveraging its value still remains a challenge.

iii. Internal data

The ability to correlate data is key to outperform the competition, as such, having data that no one else has is truly the differentiating factor in an algorithm. As a result, large institutions such as banks are looking at correlating sources like customer, credit, fraud and other types of data with trading and market behaviours.

This enables patterns identification in market prices, especially with the advent of machine-learning techniques, analysing uncorrelated data can yield a winning trading strategy.

3. Storage and databases

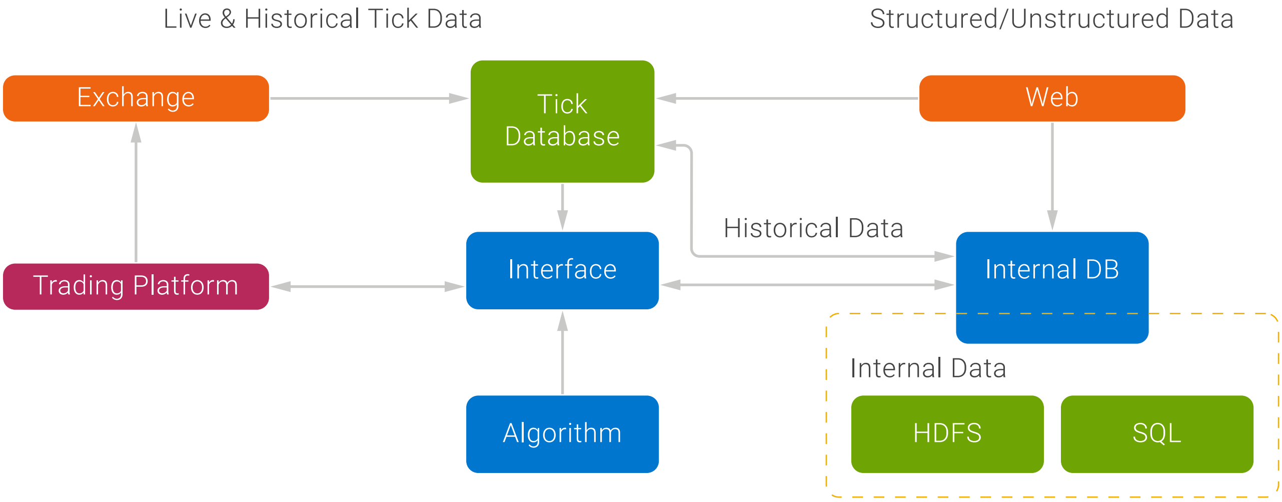
Resulting from the complexity and variety of data, databases must possess the ability to get the data from various sources simultaneously. Databases traditionally have the following requirements for algorithmic trading;

- Fast ingest of a large number of events into durable storage
- Real-time analytics, including aggregations
- Ability to process vast amounts of historical data for patterns and trends

Choosing the right database is often a choice between;

- Consistency
- Durability
- Performance

Overview of databases for algorithmic trading



i. Time series - tick databases

This is where the tick-data first lands and the biggest challenge is latency (speed) and bandwidth (volume). Exchanges now publish data at a much more granular scale because receiving it and feeding the algorithm in time is often one of the main differentiators between profit and loss. Most customers also store additional tags and proprietary data along with the market data. With the introduction and advancements of edge analytics and high-performance processing of data at the edge, customers are beginning to adopt this idea into the algorithm. In this case subsets of the larger dataset are fed into the “hot edge” for higher real-time performance and later moving that data into the much larger historical database for back-tests.

The second main challenge for the algorithm is being able to correlate the new data with historical data. To cope with this challenge, the Kdb+ (Kx System kdb+) database is the best suited as it is a high-performance, high-volume database focused on lowest latency and in-memory analytics.

With the Kdb+ 3.6 release Kx provides rapid access to unstructured data. Kx now works with Spark and Hive, which is the best use case for Kx/Hadoop interoperation. So, runtime data being generated and stored in Spark/HBase or Spark can be interoperated with Kx.

Ultimately, storage features required for a Kdb+ database include;

- Highly scalable storage solution to efficiently store billions of files
- High read/write performance
- High user concurrency capability
- Fast interconnect between compute resources and the storage platform

The Dell PowerScale platform is especially suited to meet the needs of these challenges with its scale-out architecture enabling high bandwidth, high concurrency, and high performance with all flash options.

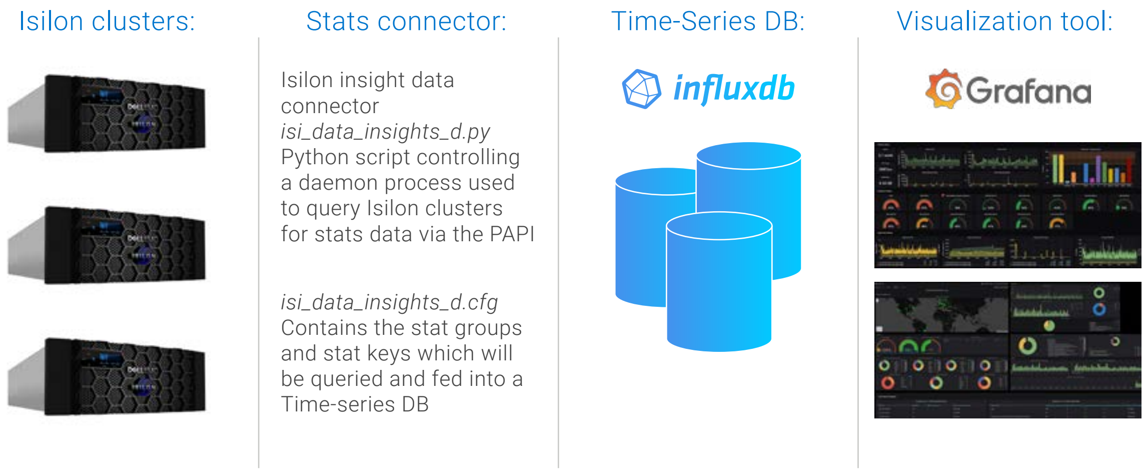
Recently Dell Technologies has published two benchmarks with the Securities Technology Analysis Center - STAC Benchmark Council and the STAC-M3 benchmark suite for “tick database” stacks. The benchmarks show performance in a test environment with a Kdb+ database for a very large database and the latest demonstrating a smaller subset of the same dataset for the “hot edge” use case.

Both of these reports can be found on the STAC Council website here:
stacresearch.com/KDB190430 - stacresearch.com/KDB200914

Where the Kdb+ database sits in comparison to the Hadoop environment and SQL database is a choice that can be determined by two factors; latency or availability. Having all, Kdb+, Hadoop and SQL, in one location is best fit for the latter factor otherwise if lowest latency is preferred, then the Kdb+ database should be as close to the exchange as possible.

In the case of having both databases apart, features such as SynIQ which is an application that enables the flexible management and automation of data replication enables seamless and automated movement of data according to pre-set policies.
Read more on the Dell EMC PowerScale SyncIQ: Architecture, Configuration, and Considerations.

Other tick-data databases include InfluxDB which is open-source and NoSQL-like. It is mostly used conjointly with Grafana, a visualisation and analytics tool directly plugged to InfluxDB.
Dell Technologies integrates InfluxDB and Grafana thanks to its Isilon Insight Data Connector which is opensource.



Isilon clusters:

Stats connector:

Isilon insight data connector
isi_data_insights_d.py
Python script controlling a daemon process used to query Isilon clusters for stats data via the PAPI

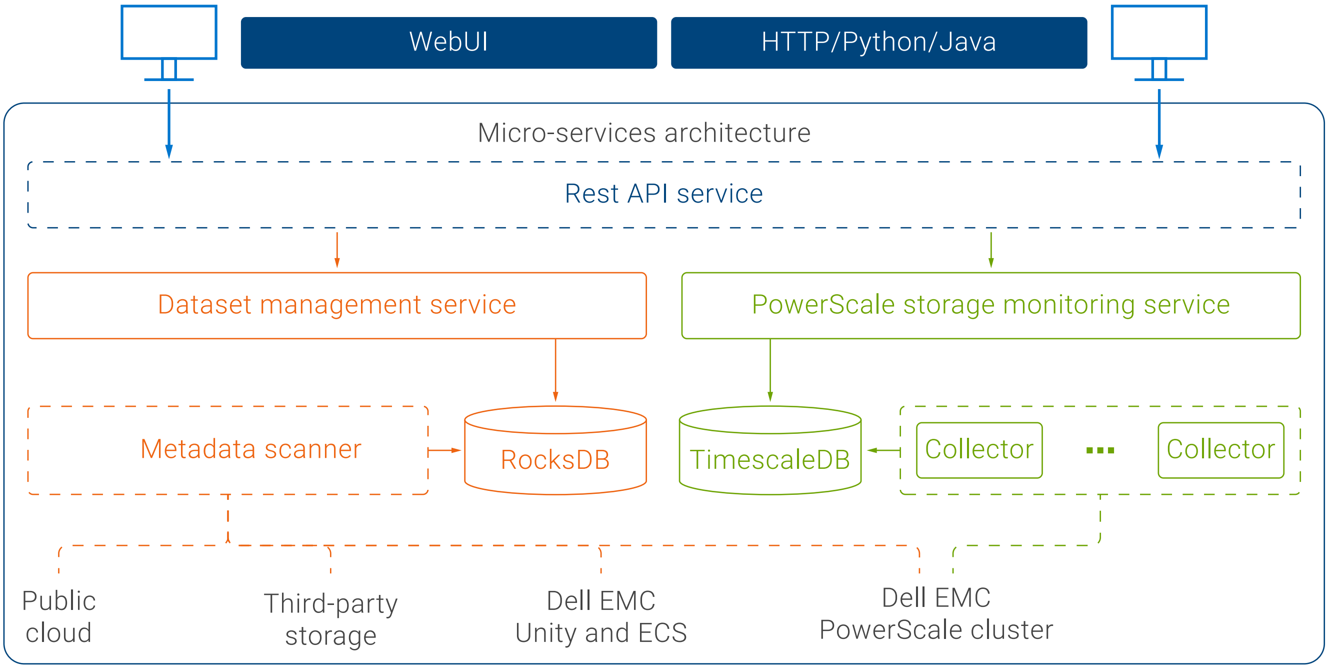
isi_data_insights_d.cfg
Contains the stat groups and stat keys which will be queried and fed into a Time-series DB

Time-Series DB:

Visualization tool:

TimescaleDB is another database that performs well with tick-data, it supports SQL making it easier to adopt. Dell Technologies now integrates the TimescaleDB with SyncIQ a dataset management software which began with ClarityNow that works together with Dell Technologies hardware.

Architecture



Finally, other popular tick-databases include Snowflake and Yellowbrick, cloud-based data-warehouses, homegrown databases or ExtremeDB - an option for high-performance, low-latency ACID-compliant embedded database management system using an in-memory database system architecture.

ii. RDBMS

Tick-databases as their name implies work great for tick-data but there are better options for overall structured data; SQL databases. Despite its performance penalty, it can be used in concert with a Hadoop environment. One obvious advantage is that it can be directly queried using SQL.

Despite not having the same level of performance than tick-databases, SQL options have the advantages of being free, easy to install and use, have a community and enterprise features. Financial data can be treated as "objects" (such as exchanges, data sources, prices) and split into tables with relationships defined between them.

Check the Dell extensive portfolio to cover your SQL data-base

iii. NoSQL

Document stores/NoSQL databases, unlike DBMS systems, don't have table schemas which makes it much easier – potentially faster – to use.

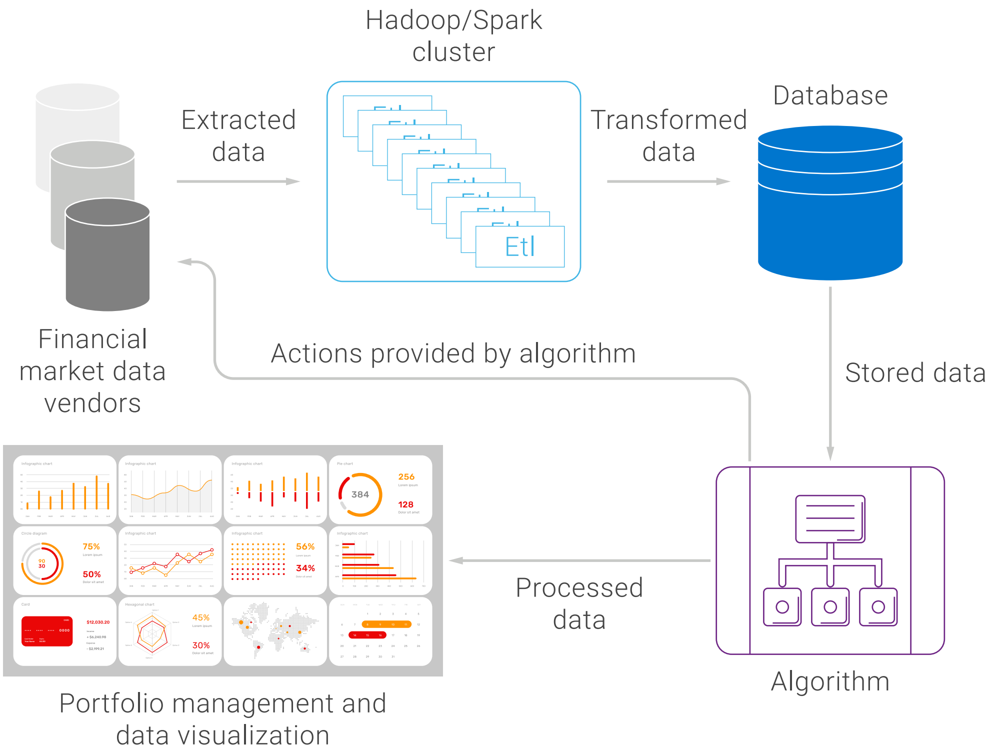
Most popular options are MongoDB, Cassandra and CouchDB. Document stores, in financial applications, are mostly suited to fundamental or meta data. Fundamental data for financial assets comes in many forms, such as corporate actions, earnings statements, SEC filings etc. Thus, the schema-less nature of NoSQL DBs is well-suited. However, NoSQL databases are not well designed for time-series such as high-resolution pricing data.

Check how to Optimize MongoDB on Dell EMC PowerStore

Move your private cloud to Dell EMC PowerEdge C6420 server nodes and boost Apache Cassandra database analysis

iv. Data-lake

MapReduce performs very well for large volume of data – both structured and unstructured. For algorithmic trading, data-lakes are great options to store historical data.


Hadoop/Spark cluster
Extracted data
Etl
Transformed data
Database
Financial market data vendors
Actions provided by algorithm
Stored data
Processed data
Algorithm
Portfolio management and data visualization

Because of the tremendous amount of data that could potentially affect markets, Hadoop and Spark bring several advantages which include:

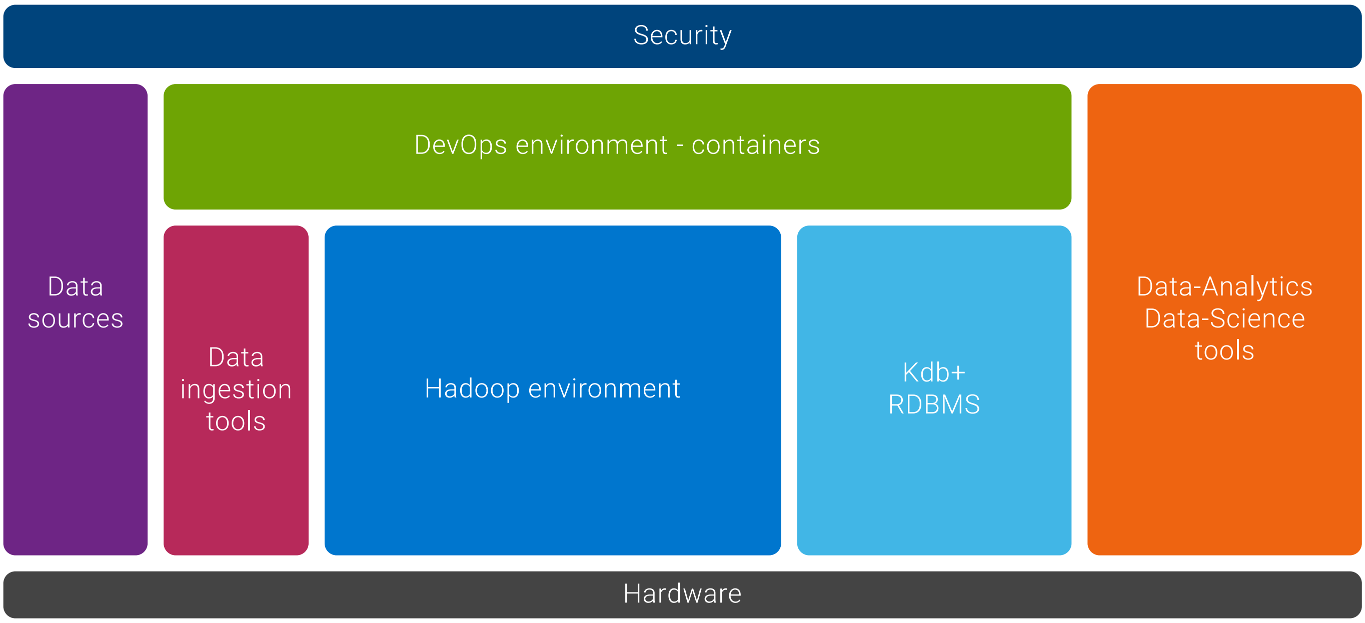
Hadoop features

- HFDS replicates your data three times which prevents data-loss
- Hadoop ability to store raw data enables the algorithm to come back to previous models without compromising performance
- Hadoop is primarily used for its ability to ingest data from various formats and convert into one
- Scalability enables you to grow your data-lake

Spark features

- An in-memory platform, enables data-processing much faster than SQL
- It is supported in Python, Java, Scala and others

Data-Lake components



To help you best size a data-lake environment, Dell Technologies has built, tested and documented several Ready Solutions for data-analytics

- **Data-Ingestion:** streaming tools such as Boomi, Kafka and StreamSets aimed for low-latency real-time data feed

Ready Solutions for Data Analytics Real-Time Data Streaming Architecture Guide

- **Hadoop Environment:** Cloudera and Greenplum are the currently most common platforms – both supported by Dell Technologies solutions

Dell EMC Ready Solutions for Hadoop
Dell EMC Ready Solutions for AI & Data Analytics - Cloudera CDP Data Center on Dell EMC Infrastructure
Dell Greenplum Reference Architecture | VMware Tanzu

- **Containers for Data-Analytics:** Kubernetes are the founding blocks of VMware Tanzu enabling fast-deployment of data-analytics models and applications

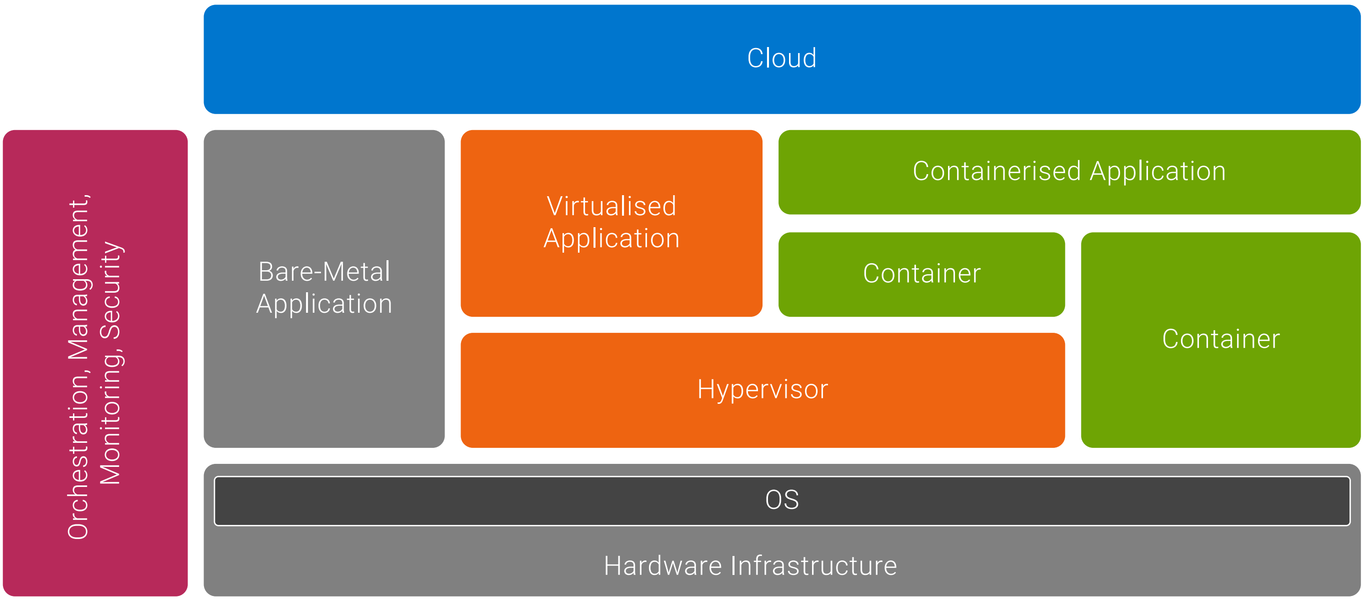
Dell EMC Ready Solutions for Data Analytics - Spark on Kubernetes
Data Science & Advanced Analytics | VMware Tanzu

4. Compute cluster

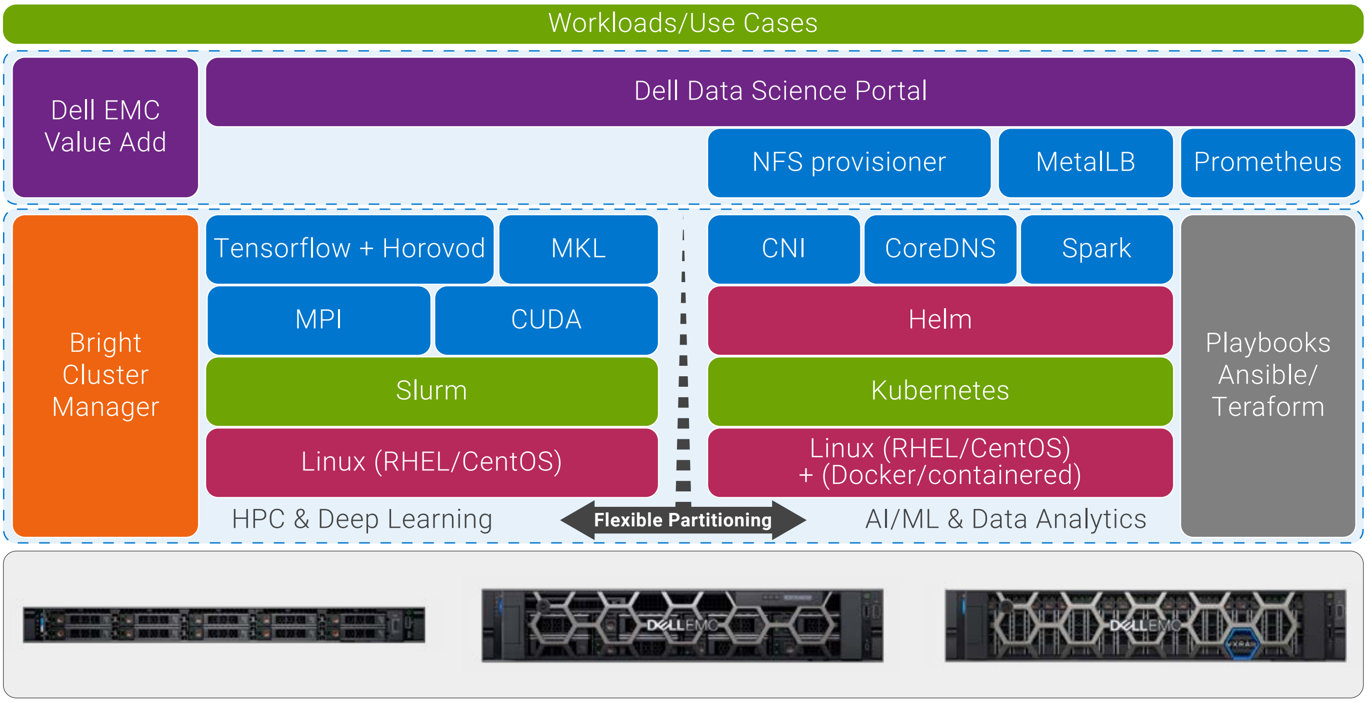
i. Overview

Algorithms are getting more and more complex and the amount of data only greater, this calls for higher computational needs that can only be met with high-performance computing clusters. Dell Technologies is the current lead in this space and has created several ready-solutions in this space.

A typical compute environment should accommodate workloads seamlessly and thus provide flexibility, performance and scaling capabilities.



Therefore, Dell Technologies has recently released a new Architecture Guide—Dell EMC HPC Ready Architecture for AI and Data Analytics – which provides high-level guidance for building a converged architecture that allows customers to run HPC, AI, and data analytics workloads on a single infrastructure.



ii. Where to run which application

A successful compute infrastructure is made of bare-metal, virtualised and containerised in order to meet changing user requirements. This allows for flexibility, performance optimisation, increased security and workload isolation.

Bare-metal is really where performance matters most, traditional HPC applications are run there as well as computational-intensive AI or data-analytics tasks such as training models or high-frequency data-analytics.

On the other hand, **virtualised environments** provide interesting features and enable having different OS environments, so if a developer wants to develop an application on Windows, then a virtual-machine can be spun-up. Virtualisation includes – but is not limited to – many advantages such as heterogeneity, multi-tenant data security, fault isolation, reproducibility, fault resiliency, dynamic load balancing. It should be noted that performance might be degraded when virtualising application.

Read this blog post to see what best OS to choose for Quant Trading

Finally, **containerised applications** are the new trend, they allow for portability, which is very convenient for developing application and reproducing the same with all dependencies. A container can port about anything, a database, code or an application. A containerised application is packaged, it can be easily movable and doesn't need to be recompiled.

iii. Hardware infrastructure

Overall, computation is the key determinant of trading strategy, the more powerful the machine, the faster it detects market movement and can place trades accordingly.

Processors

Dell Technologies will build with Intel® or AMD processors. While currently AMD processors are competing with Intel® in terms of performance and price, some operators have invested and adapted to the Intel® processors and continue to prefer to work on Intel® platforms. Dell will supply either.

Increasing voltage and frequency enables servers to attain faster speeds by **overclocking CPUs**. This generates extra-heat which in turn must be cooled. Whilst traditionally done with air, liquid-cooling has now become a necessity to reach higher overclocked frequencies.

Choosing the best CPU for your application is mostly done through testing. One key element is to understand whether the application can scale on multiple cores/ nodes. Obviously, a mix of performance/ price determines the final choice.

- Core count: if the application can scale to multiple cores, then core-count is key
- Frequency: if the application doesn't scale to multiple cores, then frequency is key
- RAM: each CPU type comes with a recommended and determined amount of memory. The larger the memory capability, the larger model the CPU can accommodate. This becomes especially important for GPUs.
- Optimised instructions: such as FMA or optimised CPU vendor libraries can speed-up your application if your code has taken advantage of these possibilities

Dell Technologies offers **liquid-cooling** with CoolIT Systems and other techniques (such as rear-door cooling) that can dramatically improve power and cooling efficiency. Check out the Data Centre Power and Cooling Solutions here

Dell Processor Acceleration Technology (DPAT) is a feature on Dell EMC PowerEdge servers that allows the iDRAC to **control turbo mode and allows for lower latency**.

Read more about the DPAT here

Overall BIOS settings do make a difference in performance; therefore, it is advised to get familiarised with how to tweak it to get best results.

Check the Dell Technologies Innovation Lab Page to optimise your BIOS settings

Graphic Cards

Even though CPUs have greatly increased their core counts, added threads, larger instruction set and other software optimisation, they remain limited in their ability to process massive parallel computational tasks.

This is where graphic cards (GPUs) excel, indeed GPUs architectures are highly parallelised allowing them to compute millions of operations simultaneously.

Dell accommodates graphical cards in its technology stack and has a direct collaboration with NVIDIA to provide you with an integrated and successful build. The GPU is the accelerator of choice since it has obvious logic for parallelising with a straightforward code development and mature numerical libraries. The GPU benefits from an extensive library and development support. Niche applications of FPGA are supported as well.

Computationally focused applications may find the “C” series of Dell servers the best choice. For database-heavy applications the “R” series of Dell servers offer comprehensive, scalable and industry-tested approaches. Finally, DSS8440 servers can accommodate 10 GPUs (V100) or up to 16 NVIDIA T4.

In terms of accelerators, NVIDIA has been and remain the choice for advancing compute capabilities. As data-models grow larger and more complex, parallel computing capabilities are necessary for performance.

When models are too large to fit on one GPU card, GPU-to-GPU communication can become the bottleneck, therefore NVIDIA has developed NVLINK – a bridge that allows every card in the server to directly speak to another, resulting in higher in bandwidth and lower in latency.

One of the many interesting features of the new A100 card is the MIG partition with which a single NVIDIA A100 GPU can turn into as many as seven independent GPU instances.

The newest NVIDIA GPU card - A100 - is pushing the limits of performance to another level. Thanks to its massive amount of cores, TensorCore, CUDA environment, and optimised libraries (see appendix) NVIDIA cards are the de-facto standard for accelerated computing.

In algorithmic trading, the development of models is very often limited by computation. NVIDIA GPU cards excel at parallel computing and is particularly efficient for back-testing.

Indeed, back-testing is a key step for the development of an algorithm and NVIDIA platforms have benchmarked performances of 6000X speed-up for back-testing in algorithmic trading. In fact, an NVIDIA system running accelerated Python libraries was able to run 20 million simulations versus the previous STAC-A3 record of 3,200 simulations in 60 minutes.

Furthermore, for deep learning trading models developed in Tensorflow or PyTorch, NVIDIA TensorRT™ software optimises trained deep learning networks. TensorRT takes the carefully trained network, once all the parameters and weights are known, and effectively compiles the model into an equivalent but more efficient version.

You can download the NVIDIA study on Algorithm Trading here

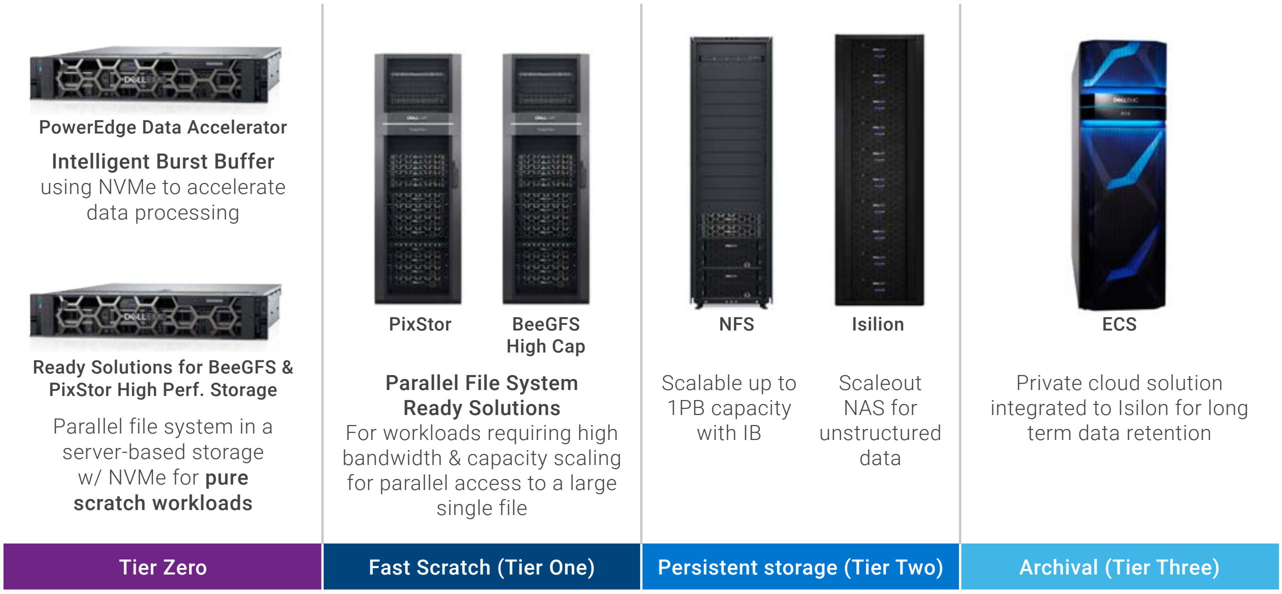
Scratch storage
Even though this not a compute component, having a scratch-storage infrastructure as part of the HPC cluster is key for performance. Indeed, this latter has several advantages such as:

- Front-end Infiniband connectivity
- Multiple concurrent reads and writes to the same file (i.e. “massively parallel”)
- Data needs can be constrained for a few days

Incumbent is a parallel file system like Lustre, GPFS/Spectrum Scale or BeeGFS. Dell Technologies provides several scratch options that include:

- Dell EMC Ready Solution for HPC PixStor Storage
- Dell EMC Ready Solutions for HPC BeeGFS Storage

Dell EMC HPC Storage Portfolio



PowerEdge Data Accelerator

Intelligent Burst Buffer using NVMe to accelerate data processing

Ready Solutions for BeeGFS & PixStor High Perf. Storage

Parallel file system in a server-based storage w/ NVMe for **pure scratch workloads**

PixStor **BeeGFS High Cap**

Parallel File System Ready Solutions
For workloads requiring high bandwidth & capacity scaling for parallel access to a large single file

NFS

Scalable up to 1PB capacity with IB

Isilon

Scaleout NAS for unstructured data

ECS

Private cloud solution integrated to Isilon for long term data retention

Tier Zero | Fast Scratch (Tier One) | Persistent storage (Tier Two) | Archival (Tier Three)

iv. Hypervisor

What are core challenges with Bare Metal High Performance Computing environments?

- No capability for **prioritising workloads** for multiple tenants: Every group or department has its own dedicated HPC environment
- Lack of **Load Balancing** Bare metal clusters are bound by the physical limits of the HW and is prone to load imbalances
- No **High Availability** for critical components: Bare Metal has no inbuilt mechanisms to deal single points of failure
- Workload **state reproducibility**: Bare metal environments do not have the capability to capture and reproduce the state of workloads

Why choosing an enterprise solution for virtualisation?

OpenStack is as good as the platform you run it on. There are literally thousands of ways to build an OpenStack cloud. This leads to a lot of snowflake clouds where the combinations are loosely integrated and poorly tested together. VMware provides a much more opinionated approach, with a single stream since the compute, network platforms, tools and even storage form only one single system.

What is the potential level of latency penalty when using virtualisation?

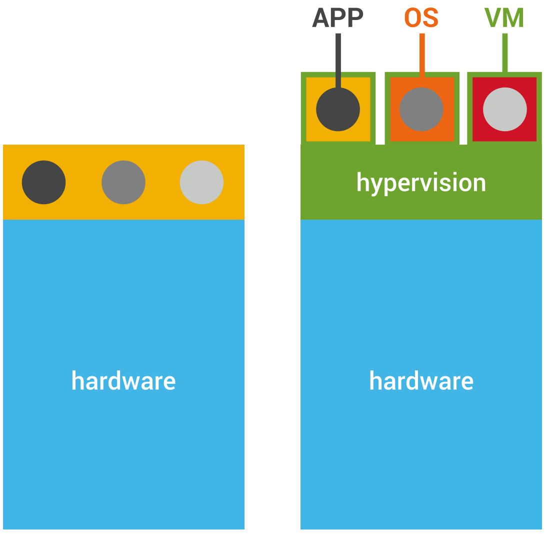
- NIC virtualisation and virtual switching is directly added to the response time and affects its variance due to extra processing
- Virtual CPUs (VCPU) supported by the virtual machine hardware result in higher latency
- Power management in both the BIOS and vSphere can negatively affect latency

Deploying Extremely Latency-Sensitive Applications in vSphere

Enabling the Latency-Sensitivity Feature in VMware; latency Sensitivity should be set to be High. The High setting is designed for extremely latency-sensitive applications and all the functionalities and tunings that this feature provides are applied. Further turnings in VMware to bypass Virtualisation Layers and other techniques can be made.

Virtualisation overhead may incur increased processing time and its variability. VMware vSphere® ensures that this overhead induced by virtualisation is minimized so that it is not noticeable for a wide range of applications including most business-critical applications such as database systems, Web applications, and messaging systems. vSphere also supports well applications with millisecond-level latency constraints such as VoIP streaming applications

With VMware, you can capture the benefits of virtualisation for HPC workloads while delivering performance. Our approach to virtualising HPC and to virtualising AI workloads adds a level of flexibility, operational efficiency, agility and security that cannot be achieved in bare-metal environments—enabling faster time to insights and discovery.



Virtual machines offer:
- Heterogeneity
- Multi-tenant data security
- Fault isolation
- Reproducibility
- Fault resiliency
- Dynamic load balancing
- Performance

Here are VMware's main components:

- vSphere: The platform that runs workloads in your data centre

- ESXi acts as the hypervisor, it is directly deployed onto servers:
  - All virtual machines (VM) run on top of this hypervisor
  - VMs run within carefully controller and managed containers
  - ESXI controls all access to the server hardware, including compute, memory, storage and networking

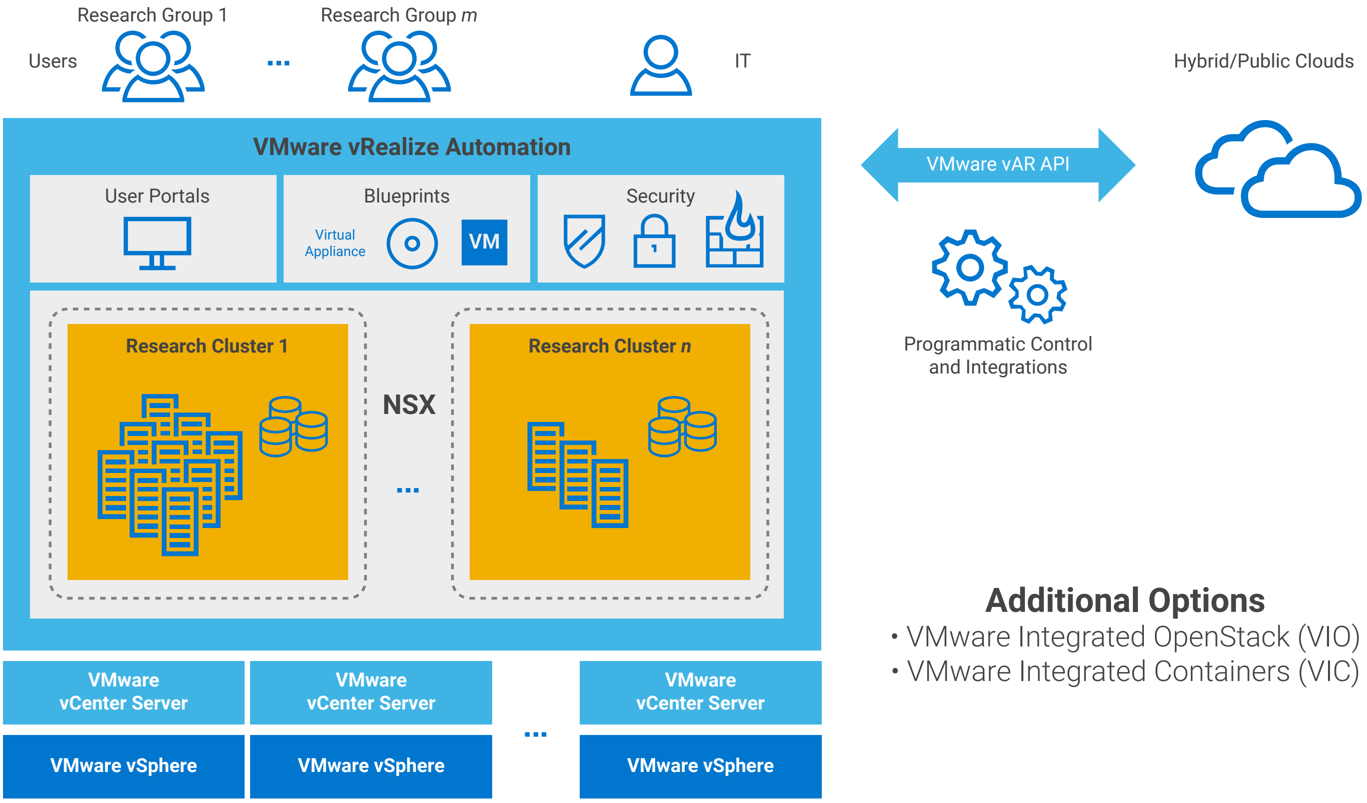
- VMware vCenter Server
  - An application itself deployed as one or several VM
  - Manages the operations of the virtual data centre
  - Takes care of tasks such as host configuration, VM deployment, operation monitoring

- vSan: Enterprise-class, storage virtualisation software that, when combined with vSphere, allows you to manage compute and storage with a single platform

- Bitfusion: Creates pool of accelerator resources, which can be shared or allocated

- VMware Tanzu: Centralised management and security of Kubernetes infrastructure and modern apps

- VMware Pivotal Labs enables the creation and management of applications, while simplifying operations across multi-cloud infrastructure: on-premises, public cloud, and edge



Additional Options

- VMware Integrated OpenStack (VIO)
- VMware Integrated Containers (VIC)

v. Containers

Containers plays an important role in the transformation and Kubernetes provides the best container orchestration. As more and more organisations are adapting and embracing a cloud-native journey, Dell EMC PowerProtect can help customers protect their Kubernetes environments.

VMware Tanzu helps you deploying Kubernetes containers on the cloud or on-prem using VMware vSphere.

Video Introduction to VMware Tanzu

VMware Tanzu is much more than containers, it is a family of products and services for building, running and managing modern apps on any cloud. Here are some key features that will help the development of algorithm:

- **Tanzu Application Service** to accelerate your business with full-stack modernisation

An intrinsically secure, scalable platform that automates the release and operation of software, optimized for Spring, .NET, Go, Node.js, and more.

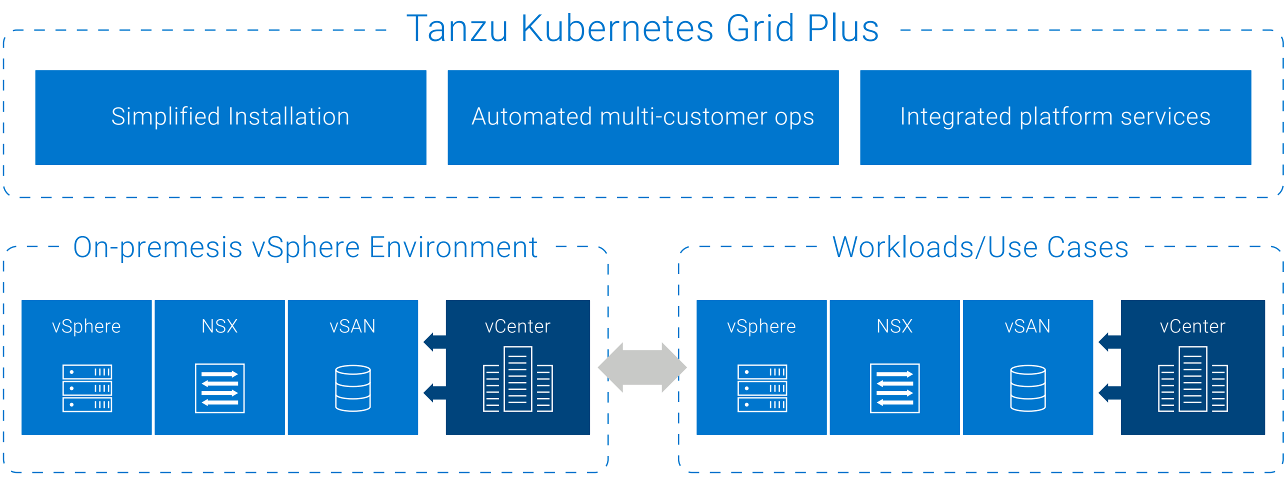
- Spring

The world's most popular Java framework that accelerates cloud-native development.

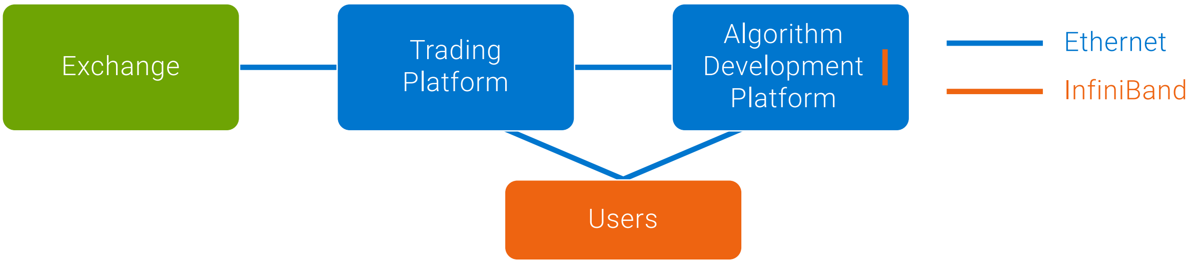
- Tanzu Kubernetes Grid

An enterprise-ready Kubernetes runtime that streamlines operations across multi-cloud infrastructure.

Read the full set of features on this solution brief


Tanzu Kubernetes Grid Plus
Simplified Installation
Automated multi-customer ops
Integrated platform services
On-premesis vSphere Environment
vSphere
NSX
vSAN
vCenter
Workloads/Use Cases
vSphere
NSX
vSAN
vCenter

5. Networking



Within an IT infrastructure for algorithmic trading, there are several networks with different requirements, we typically see this configuration:

1. Network to the exchange: lowest latency possible - InfiniBand or Ethernet
2. Internal data network: high bandwidth to move as much data as possible - Ethernet
3. Admin network: high-availability - Ethernet
4. RDMA HPC network: low-latency, high-bandwidth connection between the nodes that speak between CPUs (or GPU) memory directly without going through the CPU (or GPU) - InfiniBand

Networking options offered by Dell Technologies include Ethernet (Dell) and Mellanox (integrated package delivered in Dell builds).

For lowest latency Mellanox InfiniBand technologies provides the lowest and more deterministic latency than any other network products for Low Latency Trading – check out Mellanox Financial & Trading Solutions.

Remote direct memory access (RDMA) permits high-throughput, low-latency networking, which is especially useful in massively parallel computer clusters. Though Mellanox InfiniBand is the most common choice, RDMA over Converged Ethernet (RoCE) is a network protocol that allows remote direct memory access (RDMA) over an Ethernet network.

Ethernet still remains the main choice for the other part of the network, with switches that go up to 400Gb it allows for high-bandwidth which perfect for moving large amount of data.

Dell Technologies data center switching solutions are cost-effective and easy to deploy at any scale. From 1 GbE to 100 GbE, and 400 GbE multi-rate options, Dell EMC PowerSwitch switches provide optimum connectivity within the rack, between multiple racks, and modular compute chassis solutions. PowerSwitch switches feature a choice of software options, including Dell EMC SmartFabric OS10, Enterprise SONiC Distribution by Dell Technologies, and several options from the Dell Technologies Open Networking software ecosystem and open-source communities, to address virtually any enterprise or service provider use-case at scale.

The Dell EMC standards-based networking solutions are interoperable with leading virtualisation environments, serving as a foundation for scale-out and hyper converged infrastructure through deep integration for VMware NSX-T, vCenter, vSphere, and vSAN and deployments. At the top of rack, our latest PowerSwitch S series 25 GbE switches help customers unlock the high-speed I/O capabilities inherent in today's server and storage elements, boosting performance 2.5x over legacy 10 GbE environments.

All Dell Technologies latest PowerSwitch S series platforms include 100 GbE uplinks to facilitate high-speed interrack connectivity with our PowerSwitch Z series family of 100 GbE and 400 GbE fabric switches. In addition to the migration towards open networking solutions within the data centre, customers are also looking to reduce operating expenses.

Dell EMC SmartFabric Services enables autonomous fabric interconnects for VMware-based software-defined infrastructure. Its tight integration with VMware vCenter enables administrators to easily manage the entire infrastructure from a single pane of glass.

The Dell EMC SmartFabric Director enables data center operators to build, operate and monitor an open network underlay fabric based on Dell Open Networking PowerSwitch Series switches. SmartFabric Director automates and simplifies the provisioning and monitoring of the fabric using Openconfig based models and protocols. Tight integration with VMware vSphere and NSX-T allows SmartFabric Director to dramatically simplify fabric provisioning for dynamic virtualised workloads and overlays.

Ethernet adapters, Solarflare has been acquired by the FPGA company Xilinx. It provides lower-latency by by-passing the kernel, it does so using its proprietary technology OpenOnload, a high performance network stack. It gets its performance improvement by bypassing the OS kernel entirely on the data path.

Here are the SolarFlare cards that are supported in Dell EMC PowerEdge servers

Form Factor	Brand	Model	Speed/Ports
PCIe	SolarFlare	SFN8522 Onload	2 x 10GbE SFP+
PCIe	SolarFlare	SFN8522	2 x 10GbE SFP+
PCIe	SolarFlare (ESI Intake)	X2522 PLUS	2 x 10/25GbE
PCIe	SolarFlare (ESI Intake)	X2522	2 x 10/25GbE

Additionally, Dell Technologies just launched Solarflare's XtremeScale X2562-25G dual-port 10/25G OCP NIC3.0 Ethernet network adapters. The launch includes two variants of this NIC with and without the PLUS license that includes Onload and Precision Time Stamping.

These new NICs are fully integrated with Dell Systems Management. With Solarflare's entrance into the OCP 3.0 slot, customers no longer need to manage two NIC vendor's drivers in their OS image to take advantage of Shared LOM (iDRAC port redirect).

Going forward, Dell Technologies will be targeting to follow Solarflare firmware releases within 4 weeks with a Dell.com web-post. Solarflare firmware release will be followed by full Dell systems management, full Dell support, and the DUP (Dell Update Package) through iDRAC.

6. Users

The financial industry has quickly adopted desktop virtualisation because of its many advantages such as flexibility, mobility, ease of scalability and meeting tight regulatory and security standards.

Therefore, virtualisation has quickly become the norm, its many advantages include flexibility, security, ability to fully optimise compute and storage resources.

Dell Technologies Virtual Desktop Infrastructure offering includes:

- Wyse Thin Client family: Flexibility, form factors and several mounting options
- Dell EMC QuickStart Bundles for VDI: Enable work anywhere, anytime, on any device, for any application

Additionally, performance requirements must be kept so that no downtime for applications occur, which could lead to revenue loss. As a result, graphic cards are key to keep compute performance high whilst enabling graphic-intensive applications to run. Additionally, the ability to share powerful graphic cards optimises the total return on investment by ensuring it is fully utilised.

NVIDIA virtualisation options will enable virtualised use-cases for the financial industry:

- Virtualisation of GPUs with NVIDIA GRID® for Windows 10, Microsoft Office 365, modern browsers, advisory and analysis software, proprietary and custom applications

- The NVIDIA Quadro® Virtual Data Center Workstation (Quadro vDWS) is positioned for high-frequency, super traders. Optimised for Bloomberg, Eikon, Reuters, other electronic trading platforms

- The NVIDIA Virtual Compute Server is ideal for running compute intensive workloads including AI, data science, and high-performance computing (HPC). Perfect for applications such as RAPID framework, TensorFlow, MXNet

7. Cloud computing

Cloud computing (i.e. public cloud) enables somewhat cheap and easy access to compute and storage resources. Nevertheless, it comes with a few caveats such as vendor lockdown and difficulty to move or take-out data. Additionally, latency is still higher than similar on-prem solutions.

Ultimately, using cloud resources should be according to the use-case, a hybrid model (on-prem & cloud) is therefore advisable. Finally, the ability to leverage several cloud-vendors is key to keep optimal price/ performance.

Especially designed for AI, HPC and high-performance data-analytics, the Dell Technologies Hybrid Cloud offers agility, reduce complexity whilst staying affordable. The model enables any organisation to run several cloud models at the same time. Dell Technologies helps you to move HPC resources to the cloud, protect your intellectual property and maintain compliance.
Mix and match the options below to create a hybrid cloud HPC environment that best suits your needs

	Purchse options	**Deployment options**	**Management options**
Private on- or off-premise cloud	Buy or lease hybrid cloud solutions using Dell Technologies Financial Services	Use internal resource or Dell Deployment services for on-premises. Use Dell or hosting provider deployment services for off-premises	Use internal resources or outsource with Dell or partner managed services
Public cloud	Pay-per-use for HPC resources (IaaS or PaaS)	No deployment required	Managed by the cloud providers
Hosted and managed public or private cloud	Consume hosted and managed HPC services in a PaaS/HPCaaS model	No deployment required for PaaS/HPCaaaS	Managed by the cloud providers
Public/private cloud partnerships	pay-per-use for supercomputing resources offered by academic institutions	Typically, no deployment reguired	Managed by the organisation or institution

8. Software

The Dell Technologies OpenManage portfolio provides systems management solutions designed to simplify, automate and optimise your IT operations:

- Deploy as a secure virtual appliance
- Intuitive dashboard and elastic search engine
- One to many intelligent automations with user defined policy, template, and baseline
- Comprehensive RESTful API enables customisable automation and solution integration

The Integrated Dell Remote Access Controller (iDRAC) is designed for secure local and remote server management and helps IT administrators deploy, update and monitor Dell EMC PowerEdge servers anywhere, anytime.

Bright Cluster Manager is a software that automates the process of building, managing and monitoring clusters to reduce risk, increase productivity and accelerate time to value.

Bright and Dell Technologies have been partners since 2011 with Bright Software optimised and integrated in Dell hardware. In fact, Bright can be used to configure/ tune and managed Dell EMC server BIOS which results in faster deployment and quicker BIOS changes in each node. Additionally, Bright can be used to monitor specific system health related metrics from iDRAC, providing much more comprehensive overall system monitoring than anyone else. Integration of iDRAC and Bright for health metrics:

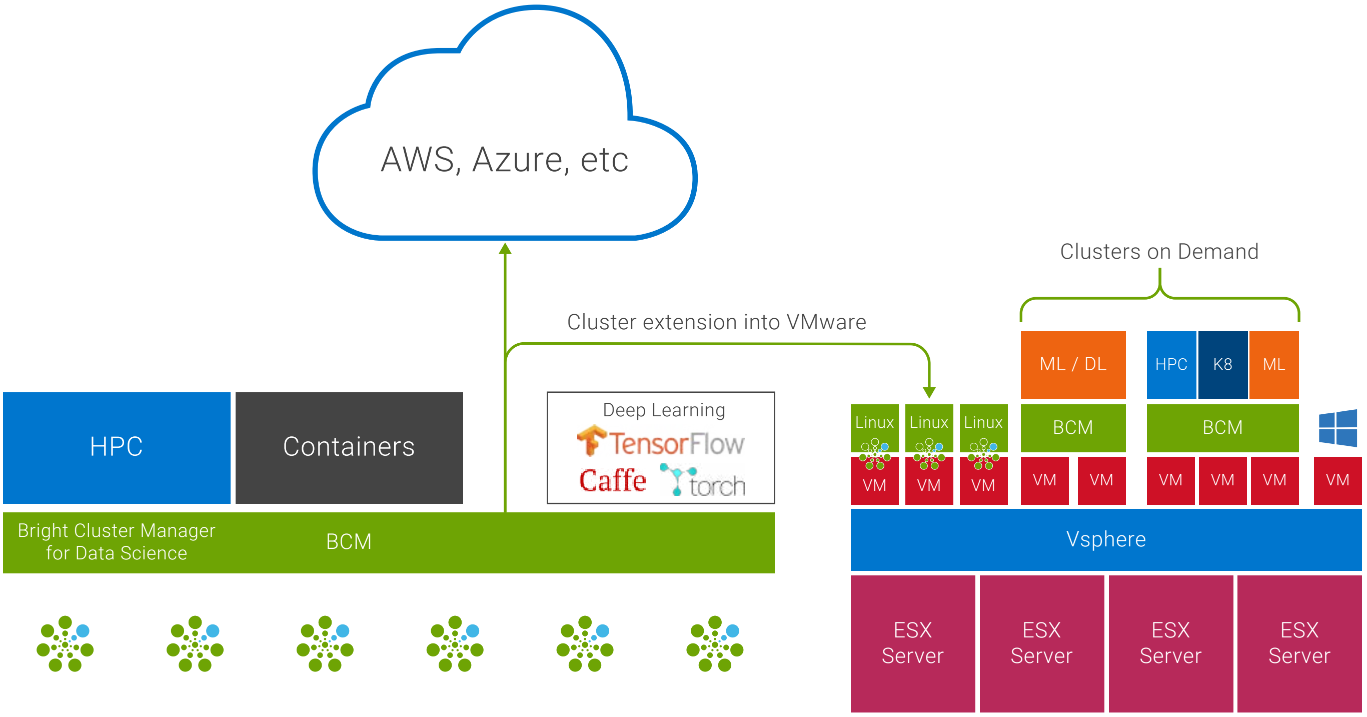
- Parameters from iDRAC are gathered to monitor specific system health-related metrics
- Over 40-50 out-of-the-box metrics are available in Bright today, including:
  - Power measurements and Usage in Amps and Volt
  - PSU status and motherboard sensor and voltage regulator info
  - In-Let, outlet and Exhaust Temp of the system,
  - Different Fan speeds and Fan redundancy status
  - GPU temp
  - IO usage
  - Memory usage
- Custom metrics can be easily added

BIOS integration in Bright:

- Dell APIs talk to the CMD in Bright and initiate commands using RACADM to modify certain tunable options to help optimise or modify settings.
- Some of the BIOS tunable options available in BrightView are:
  - System Profile Settings: Custom, DAPC, OS DenseOptimized, PerfOptimized
  - Logical Processors: Enabled/Disabled
  - Node interleaving: Enabled/Disabled
  - Processor pre-fetches
  - Network latency parameters: IONP Prefetch, C-states etc.

Bright Computing integration with VMware:

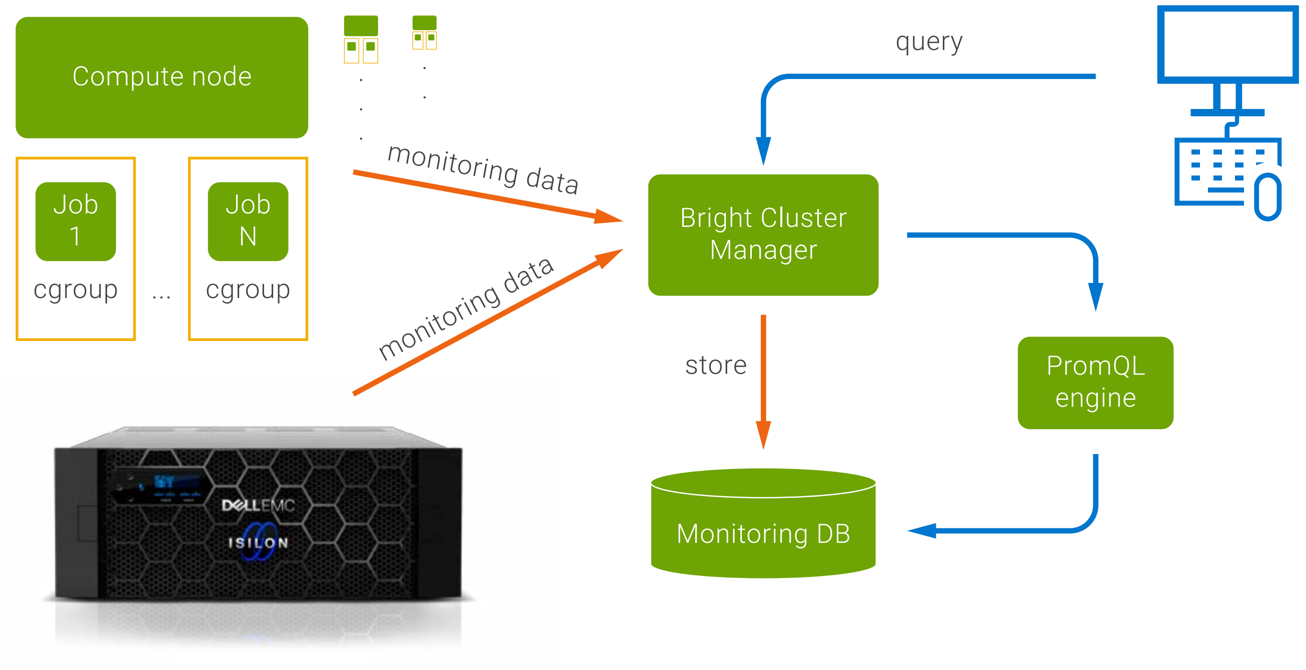
- Integrate with VMware to provide HPCaaS private cloud
- Customers administrators can create fully functional Bright clusters within VMware (cod-VMware)
- Cluster administrators can manually extend Bright cluster into VMware (Bright-managed instances)
- End user jobs can burst into VMware (cex-VMware)
- Works with existing VMware clusters



Bright Computing integration with Isilon (WAR – Workload Accounting and Reporting)

Administrators can create reports which show (for example):

- Which job is consuming all the storage capacity right now?
- Which users'/applications/jobs/projects generated the most I/O over the last year?
- Which users' jobs caused the storage to be under-performing last night?



IV. Decision process

1. Creating a solution

Choosing the right HPC/ AI solution is a lengthy and complicated process. Thankfully, Dell Technologies can guide you through your journey with five main steps:

1. Consulting
2. Testing
3. Financing
4. Deployment, support, management
5. Community

i. Consulting

Dell Technologies has a dedicated team of consultant as well as HPC, AI, HPDA and workload specialists that will provide you with the right advice. Feel free to contact your local HPC specialist:

- **North America** - HPC_NA_Sales_Team@Dell.com
- **EMEA** - EMEA_HPC_Team@Dell.com
- **APJ** - APJ_HPC_Team@Dell.com
- **LATAM** - HPC_Latam@Dell.com

You can read more on Dell Technologies Service Here

Alternatively, our innovation lab publishes white papers on application and hardware performance.

ii. Testing

How an application behaves in real-life is key to understand real performance levels. Dell Technologies Innovation Lab helps you just with that; it is a 13,000 square foot data centre which houses thousands of servers, a TOP500 cluster, sophisticated storage and network system for HPC.

Additionally, you can test your solutions locally at Dell Technologies Customer Solution Centres for POCs, design workshop or trainings.

Finally, you might as well just visit our customers, the centres of excellence (COEs) are Dell Technologies users with which you can organise workshops.

iii. Financing

Dell financing options include:

- Dell Preferred Account (DPA): A revolving line of credit tailored for individual and home purchases of Dell equipment.

Small Businesses

- Dell Business Credit: Helps companies of all sizes build their business credit and continue to grow.
- Dell Business Lease: Powered for Business leasing from Dell Financial Services helps you acquire the technology you need today.

Medium to Large Businesses

- Flexible payment solutions: Whether your business is working under constrained budgets, limited cash flow, or facing an uncertain business climate, our Payment Flexibility Program can help.

iv. Deployment, Support, Management

Dell Technologies services for HPC are a dedicated offering for clusters:

- Deploy: Dell EMC’s HPC deployment model provides our customers with comprehensive, proven cluster implementation at the right price, that scales regardless of cluster size.

- Support: The ProSupport Add-on for HPC provides solution-aware support with specific entitlements for HPC Ready Bundles, including access to dedicated HPC solution experts to help manage the complexities of supporting a multiple-vendor cluster.

- Manage: Dell EMC’s Remote Cluster Management (RCM) service provides highly skilled experts to pro-actively manage and maintain your HPC cluster and applications so that you can focus on your core business.

v. Community

The Dell Technologies HPC Community is starting a new online program, with weekly presentations and conversations featuring leaders from Dell Technologies, HPC technology partners in the Dell Technologies HPC ecosystem, and technology thought leaders from industry, government, education and research.

Join the Dell Technologies HPC Community

2. Dell Technologies Advantage

i. Supercomputers at customers' disposal

The Dell Technologies HPC & AI Innovation Lab in Austin, Texas is a 13,000 square foot data centre that houses thousands of servers, a TOP500 cluster, a wide range of storage and network systems. The lab is made of Intel®, AMD, NVIDIA, Mellanox and other technologies, which customers can use to test their applications and get access to the latest technology.

Explore the Dell Technologies HPC & AI Innovation Lab

Additionally, Dell Technologies partners with Supercomputing centres all over the world which are called Centres of Excellence. Customers can schedule a visit to them to have a 'user to user' conversation.

HPC & AI Centers of Excellence Hubs for innovation and expertise

ii. Software optimisation

Making sure that the applications will be working on the hardware is crucial, therefore Dell Technologies has partnered with software optimisation service providers to make sure that you get best performances out your infrastructure.

iii. Green and social responsibility

Dell Technologies is embracing a greener and more sociable future, we are tracking our goals with our 2030 Moon-shot Goals.

Renewable electricity

At Dell We are committed to transitioning to renewable power and have a moonshot goal of sourcing 75% of electricity from renewable sources across all Dell Technologies facilities by 2030 — and 100% by 2040.

Dell Technologies champions Energy Efficiency

Since FY12, we have reduced the energy intensity across all product categories by 69.9%. To put that into perspective: FY20 customers will spend approximately $182 million less in electricity costs over the life of their products compared to the FY19 portfolio in order to do the same amount of computing work. Our server portfolio alone has seen an 82.5% reduction in its energy intensity since FY12.

Better Design for the Environment

Dell follows the ISO 140001 standard for environmental management, Dell helps commercial customers resell, recycle, or return to lease their excess hardware in a secure and environmentally conscious manner that complies with local regulatory guidelines.

Dell Technologies stands for Diversity & Inclusion

The STEMAspire programme connects female students in higher education within science, technology, engineering and mathematics (STEM) disciplines with female mentors at Dell.
Releasing Female Potential is a 12-month development programme that connects high performing early to mid-career women professionals and under-represented minority women professionals with senior leaders at Dell Technologies.

Achievements

- More than 2 billion pounds of electronics recovered since 2007
- 99% of all manufacturing waste diverted from landfills in FY20
- Industry's first ocean-plastic packaging, with commitment to scale 10x by 2025
- Founding member of NextWave, a consortium to address ocean plastics at scale
- 2020 World's Most Ethical Company Award from the Ethisphere Institute for the seventh consecutive year
- 2019 Responsible Business Alliance's Compass Award for Leadership
- Ranked #1 for environmental management of our supply chain across all industries by IPE's Corporate Information Transparency Index in 2019
- Ranked #3 America's Most Responsible Companies List 2020 by Newsweek and Statistica
- Honoree: FastCompany's 2019 World's Most Innovative Companies for Consumer Electronics
- Ranked #2 in the Computer Industry for Fortune's Most Admired Companies according to Universum in 2019
- Many of our servers are EPEAT Bronze certified, with 26 server categories registered in various European countries
- Our factories in Poland and Ireland run 100% on renewable energy.

i. Appendix

1. Industry terminologies

CFTC Commodity Futures Trading Commission
ISDA International Swaps and Derivatives Association
MiFID II Markets in Financial Instruments Directive
MTF - Multilateral trading facility
OTF – Organised trading facility

2. External Links

Dell EMC High Performance Computing Solutions

Dell EMC Isilon: Using Isilon F810 with SAS Analytics for Financial Services

Dell EMC Ready Solutions for AI, HPC and DA

Dell Technologies Legacy of Good

NVIDIA Portfolio

Three main SW platforms:

- Accelerated Computational Finance/HPC (CUDA C++)
- Accelerated Data Science for Python (Rapids)
- Accelerated Deep Learning
  - Tools: TensorFlow, PyTorch, Keras (via cuDNN)
  - Accelerated Inference (TRT, Triton)
  - Containers: Optimized solutions and selected trained models (via NGC)
- www.nvidia.com/en-us/gpu-cloud/ also has CUDA and Rapids containers

Two generations of GPUs for compute in the market today

- Volta (Broadly available) www.nvidia.com/en-us/data-center/v100/
- Ampere (New/Ramping) www.nvidia.com/en-us/data-center/a100/

Accelerated Computational Finance/HPC -- CUDA C++

- Programmer's Guide: docs.nvidia.com/cuda/cuda-c-programming-guide/index.html
- Online training: courses.nvidia.com/courses/course-v1:DLI+C-AC-01+V1/about
- C++ Template programming: docs.nvidia.com/cuda/thrust/index.html
- STAC A2 Market Risk Benchmark stacresearch.com/a2
- Library solutions: (umbrella developer.nvidia.com/hpc-sdk)
  - Linear algebra: docs.nvidia.com/cuda/cublas/index.html
  - Sparse linear algebra: docs.nvidia.com/cuda/cusparse/index.html
  - Some linear algebra solvers: docs.nvidia.com/cuda/cusolver/index.html
  - Random Numbers: docs.nvidia.com/cuda/curand/index.html
  - FFT: docs.nvidia.com/cuda/cufft/index.html
- CUTLASS: Write your own matrix routines
  - developer.nvidia.com/blog/cutlass-linear-algebra-cuda/
- Samples – educational, not fully optimised
  - General: docs.nvidia.com/cuda/index.html
  - Financial: docs.nvidia.com/cuda/cuda-samples/index.html#finance
- Write your own compiler: docs.nvidia.com/cuda/libnvvm-api/index.html
- Some firms:
  - www.hanweck.com/
  - www.flink.ai/
  - www.murex.com/

Accelerated Data Science for Python – Rapids

- Open Source HQ: rapids.ai/
- STAC A3 Backtesting Benchmark: stacresearch.com/a3
- Subsites for
  - Pandas-like data frames: cuDF
  - Scikit-Learn like routines: cuML
  - Focus on XGBOOST
  - SciPy Signal-like: cuSignal
  - NetworkX like: cuGraph (very early days)
  - Spatial analytics: cuSpatial
  - Strings: nvStrings
- Numba for custom JIT code numba.pydata.org/numba-doc/0.13/CUDAJit.html
  - Some examples at gQuant: github.com/rapidsai/gQuant
- Multi-GPU/Multi-Node via Dask rapids.ai/dask.html
- SQL Analytics via blazingSQL rapids.ai/blazingsql.html
- Spark 3.0 support nvidia.github.io/spark-rapids/
- In Memory Databases
  - Kinetica www.kinetica.com/
  - Omni-Sci www.omnisci.com/
  - Brytlyt www.brytlyt.com/
  - Sqream sqream.com/

Accelerated Deep Learning

- MLPerf Benchmark
  - NVIDIA Site www.nvidia.com/en-us/data-center/mlperf/
  - MLPerf Site mlperf.org/
- Deep Learning Frameworks Details with GPUs support docs.nvidia.com/deeplearning/frameworks/index.html
  - (also check NGC Container Repo for prebuilt/tuned/tested version)
- TensorFlow (Overview and various links)
  - www.nvidia.com/en-sg/data-center/gpu-accelerated-applications/tensorflow/
  - www.tensorflow.org/install/gpu
  - ngc.nvidia.com/catalog/containers/nvidia:tensorflow
  - docs.nvidia.com/deeplearning/frameworks/tensorflow-user-guide/index.html
- PyTorch
  - docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/running.html
  - pytorch.org/ will literally let you pick your CUDA version to download the correct version
- Inference
  - TRT optimises your trained model developer.nvidia.com/tensorrt

TRITON serves optimised models with lower latency and higher throughput developer.nvidia.com/nvidia-triton-inference-server

Model parallelism

www.stacresearch.com/news/2018/10/23/INTC181012
www.stacresearch.com/news/2017/08/01/NVDA170718
www.stacresearch.com/news/2016/10/27/INTC161016

3. Some literature

Trevor Hastie, Robert Tibshirani, Jerome Friedman: “The Elements of Statistical Learning – Data Mining, Inference, and Prediction”, 2nd Edition, Springer, 2017

Matthew F. Dixon, Igor Halperin, Paul Bilokon: “Machine Learning in Finance – From Theory to Practice”, Springer, 2020

Alexander Denev and Saeed Amen: “The Book of Alternative Data – A guide for Investors, Traders and Risk Managers”, John Wiley & Sons, 2020

Michael Dempster, Juho Kanniainen, John Kean, Erik Vynckier: “High Performance Computing in Finance – Problems, Methods and Solutions”, Chapman & Hall/CRC Financial Mathematics Series, 2018

Gabriel Pirastru - gabriel_pirastru@dell.com

- - HPC_NA_Sales_Team@Dell.com
- - EMEA_HPC_Team@Dell.com
- - APJ_HPC_Team@Dell.com
- - HPC_Latam@Dell.com

www.delltechnologies.com/en-us/solutions/high-performance-computing/index.htm

www.delltechnologies.com/en-us/solutions/data-analytics/index.htm

www.delltechnologies.com/en-gb/service-providers/edge-computing.htm#scroll=off

www.delltechnologies.com/en-us/precision/index.htm

Dell Technologies Global Headquarters is located at One Dell Way, Round Rock, TX, 78682