White Paper

Latency Imperatives and Implications

Capital Markets today face unprecedented challenges around their scope and function. The purpose that capital markets serve—part complementary and part competitive to the banking industry— has come under the increasing scrutiny of regulators across the globe. While markets evolve, both in response to regulations and the circumspect economic climate, the key imperative in navigating through the present scenario is to look at differentiators and competitive advantages which will sustain the business, as it stands today, and as it incrementally evolves.

The differentiators around various product offerings are increasingly blurred while competitive advantage around financial innovation is on a tight rein forcing focus on inward looking innovations. These inward looking innovations include exploiting efficiencies and tapping resources in imaginative ways to create operational differentiators. The key operational differentiators in today's environment are the **'speed edge'** provided by the low latency paradigm and the **'lean and business focused operating model'** provided by the cloud computing paradigm.

Radically reducing time scales - the low latency paradigm:

The time gap between perceiving an opportunity and exploiting it for business gain is at the heart of trading effectiveness in capital markets. Data networks interconnecting institutions and markets have evolved radically to provide 'order of magnitude' increases in both speed and capacity. Computer processors and information system architectures are also evolving in step to correspondingly increase information processing speeds. Increasing use of computerized algorithms, enabled by both computing and networking advancements, is shifting focus from human reflex based trading advantages to computerized trading battles in the microsecond realm.

This paper is an in-depth commentary on the context, adoption and implications of the latency driven operational differentiator. From an adoption perspective, it covers approaches spanning (a) stretching existing paradigms which are technology stack innovation based, (b) paradigm shifts which are the shifts seen from instruction to data centric computing and the shift to imperative style programming and (c) industry specific solutions which cover proximity, co-location and lean, proprietary messaging formats. From an implications perspective, developments such as globalization and diversification of trading, new technology focused offerings and increased emphasis on smart order routing are foreseen.

Through a holistic perspective of the trading value chain, this paper strives to address the challenges and priorities of trading business managers and operations managers involved in trade processing as well as IT managers supporting the dynamic trade processing environments.

Part 1: Deals with the context of low latency in financial markets, specifically focusing on the trading chain

Part 2: Deals with the sources of latency in the trading chain, from the perspective of market participants

Part 3: Deals with the approaches for dealing with low latency

Part 4: Summarizes latency imperatives and discusses possible evolutionary aspects

About the Author

Avinash Patil

Avinash Patil is a member of the Capital Markets Practice Group at TCS. He has 16+ years of experience in the IT industry, most of which is with global banking and financial services firms. His areas of interest include low latency, cloud computing and innovation challenges in Capital Markets. He holds a B.Tech(Hons) degree from IIT Kharagpur.

Table of Contents

PART 1: Low Latency Context – Trading Value Chain in Capital Markets

Context

At the very core, Capital Markets are about capital transfer and risk sharing, driven by differing perspectives of participants, who interact with each other in market places. It is this interaction in the market place that drives the bulk of operational activities in capital markets, from portfolio analytics, trade execution to fulfillment. Given the developments in the market place over the last few years, there is a clear and significant shift in the proportion of transactions from investing to pure trading.

In the US, Morgan Stanley's findings1 highlight that trading by 'real' investors accounts for the smallest share of US stock market volumes (since Morgan Stanley started tracking 10 years ago). The findings highlight how fast turnover of shares by independent firms and market making desks of brokerages is increasingly fueling US trading activity. To put it into perspective, the proportion of US trading activity represented by buy and sell orders from mutual funds, hedge funds, pensions and brokerages, referred to as 'real money' or institutional investors, accounted for just 16% of total market volume in the form of buying, and 13% via selling, in the final quarter of last year .

The story is similar in the UK, Europe and to some extent in the Asia-Pacific (APAC) markets too. In the UK, the TABB Group estimates pure trading in equity markets to be around 77% while AFM estimates peg it at 30-40% for European markets. The estimates for APAC and Australian markets2 range from 15 to 25%.

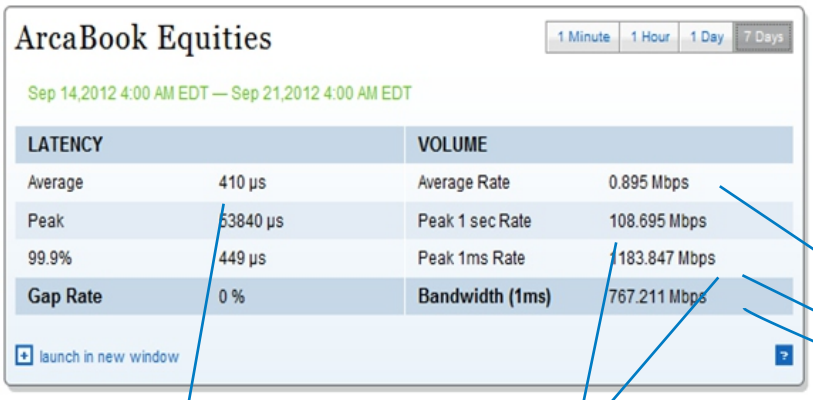
While this trend is most pronounced in the equity markets, the other instrument class where it is observed is in the foreign exchange (forex) markets – it is estimated3 that computerized, high speed trades drive about 40% of spot forex volumes.

Let us put the speeds into perspective – in the time it takes for a human eye to blink (~250 milliseconds), a high speed trader can execute over 5000 transactions! An Aite Group estimate puts the value of a one millisecond latency advantage for a brokerage firm at over $100 million annually. In forex markets, a high speed trader can execute at less than a millisecond, whereas a comparative execution by a conventional, non-high speed trader takes around 10-30 milliseconds.

While debate rages across the world on the pernicious effects of a form of high speed trading, called HFT , the use of computerized algorithms for trading is here to stay and grow. Increasingly lower latencies, which are the new normal in the trading world today, are dictating the efficiency and outcome of a trading strategy. An Oct 2011 survey by Interxion reveals that over 70% of market participants indicated that their trading strategies are latency dependent.

A deeper look at the pure trading transaction volumes reveals that most of it comprises peaks of orders as well as their related cancellations and resubmissions at milli-second intervals - indicating algorithm driven computer traders talking to each other. In 'real money' transactions too, an institutional trade is 'sliced' and executed at 'second' and 'sub-second' intervals, driven by the need to minimize price impacts and thus the total transaction cost.

Fig 1 shows statistics on the impact of latency on transaction volumes (Source: NYSE Arca – LatencyStats.com) that can help in putting this into perspective:



Average latency is <0.5 ms, compared to 10s of seconds in 2006!

Peak 1ms rate is an order of magnitude higher than 1s rate - indicating the extent of intra-second bursts driven by automated strategies!

These three figures indicate a regularity on the border of a second or within a second – which indicates extent of automated strategy use by a "real money" trading strategy

This implies that though the objectives of pure trades and real trades are different, the means are similar- through high speed, low-latency technology infrastructure.

Low latency trading also has significant ramifications on other operations in capital markets. Market data volumes are exploding exponentially and post trade processing, involving clearing, settlement and portfolio accounting, is registering significantly higher volumes.

It is clear from the above that low latency capability is no longer a matter of isolated choice, it is real, it is pervasive and it impacts almost everyone – it is the 'new normal' today.

Let us examine how these imperatives play out in the markets:

At a simplistic level, the following considerations based on credit and market risk assessments as well as risk appetite drive trading in capital markets:

1. Fast turnover with minimal to no over-night 'holdings' or
2. Longer term 'value investing' based on the belief that business fundamentals will drive asset value or
3. Hedging with counter-cyclical / loss mitigating instruments, across asset markets and geographies

In the current uncertain and unstable business environment with high information sensitivity, small fluctuations predominantly drive volumes in (1) and large fluctuations predominantly drive volumes in (3). However, in all these trading scenarios, the 'relative time advantage' imperative clearly stands out:

For 'real money[(1)]' traders, achieving desired price with minimal price distortions is key, (this typically plays out in cases where agency traders execute large block orders on behalf of institutional investors). Achieving desired price and controlling price distortions have counter-balancing influences – the traditional way to counteract the influence of a trade on price distortion is to distribute executions over a

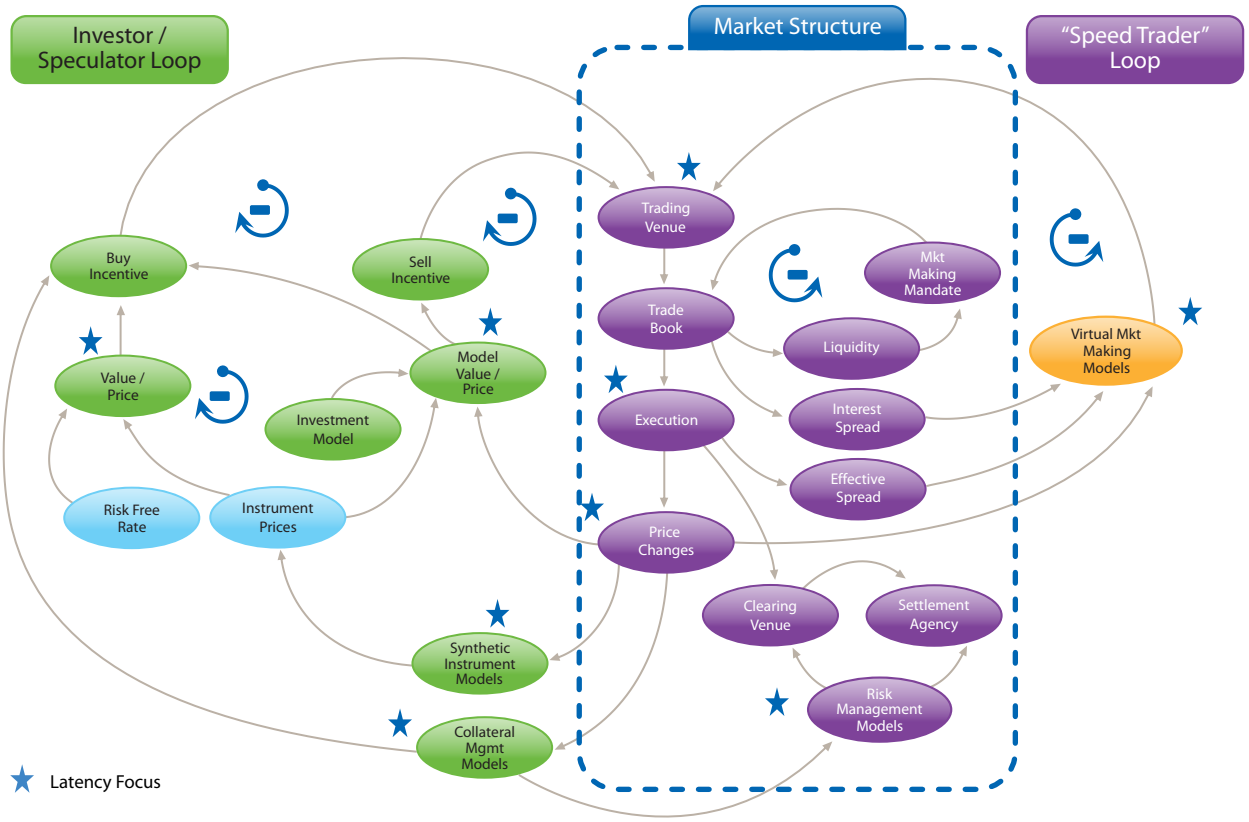
number of venues to ensure liquidity. However, in volatile business environments where liquidity fluctuates, achieving a 'relative timing advantage' on trade executions at the execution venues becomes an imperative.

For 'pure trade[(ii)]' players, capitalizing on pricing distortions is key. With increased access to information in the market, the size of the distortion is becoming increasingly small. This is driving a race to capture the window of opportunity ('relative timing advantage') through better automation and faster execution.

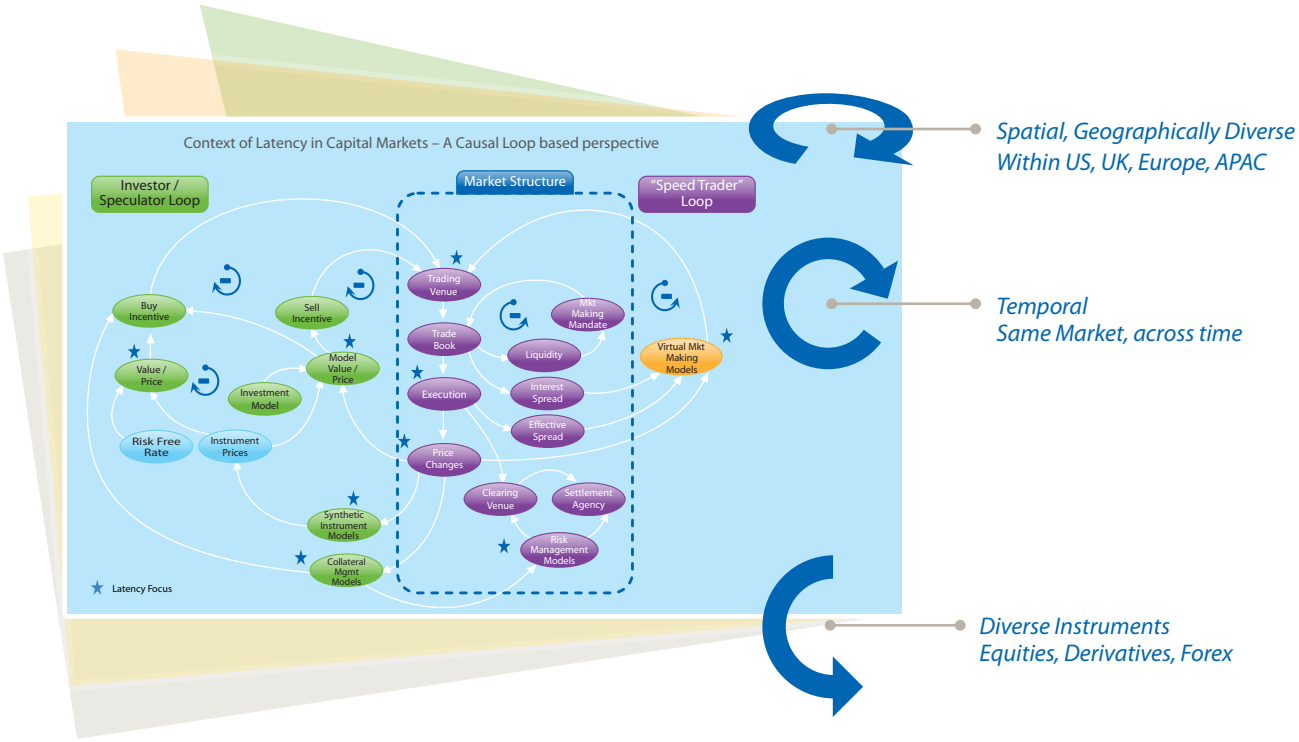
Latency also correspondingly impacts other value chains involved in the trade life cycle depending on the nature and extent of their interaction. Index service providers and Exchange Traded Fund (ETF) vendors whose 'product values' depend on the prices of underlying primary stocks, have to churn out values faster. Collateral management models used by stock lenders as well as clearing and settlement organizations have to deal with volatility risk by computing exposures at increasingly frequent intervals, almost down to real-time computations.

Another way of looking at how these imperatives are playing out in the market is to look at the causal loops driving speeds in the trading lifecycle itself:

Context of Latency in Capital Markets – A Causal Loop based perspective



(i) 'Real Money Trading' refers to trading strategies based on value investing, with a medium to long term view
(ii) 'Pure Trading' refers to trading strategies which are extremely short term, typically to exploit arbitrage opportunities



As seen in Figure 2, the key 'battlegrounds' for latency and 'first order gains' are at:

- **Trading Venue Access,** which determines how fast an order can reach a trading venue and gain a timing advantage compared to other participants. Proximity to trading locations, co-location strategies and faster networks are the foundations of such competitive advantages.
- **Market Data Access,** which determines how fast data on stock prices, executions, corporate actions, economic indicators, can be accessed by businesses to 'identify trading opportunities', compared to other participants. Market data vendors such are Reuters, Bloomberg and SIX are increasingly investing in high performance, low latency infrastructures, both from a data sourcing and data distribution perspective.

While network speeds drive 'first order' imperatives, a realization that network speed can provide diminishing rate of return on investment is driving **'second order'** imperatives. Such imperatives include faster and differentiated information processing, which is at the core of the investment/trading models. High Performance Computing (HPC) infrastructures and systems architectures coupled with market specific differentiated processing strategies are driving increasing efficiencies in the latency race, both for participants (traders) as well as market venues (exchanges).

As a consequence of latency battles centered on network speeds and faster trading/investment models, a **'third order'** imperative focused on regulatory compliance is emerging. Latency battles are creating increasing imbalances from the perspectives of 'accessing liquidity and cornering the market' as well as the 'role and risk of market volatility'. European regulations aimed at restricting the 'ratio of orders to

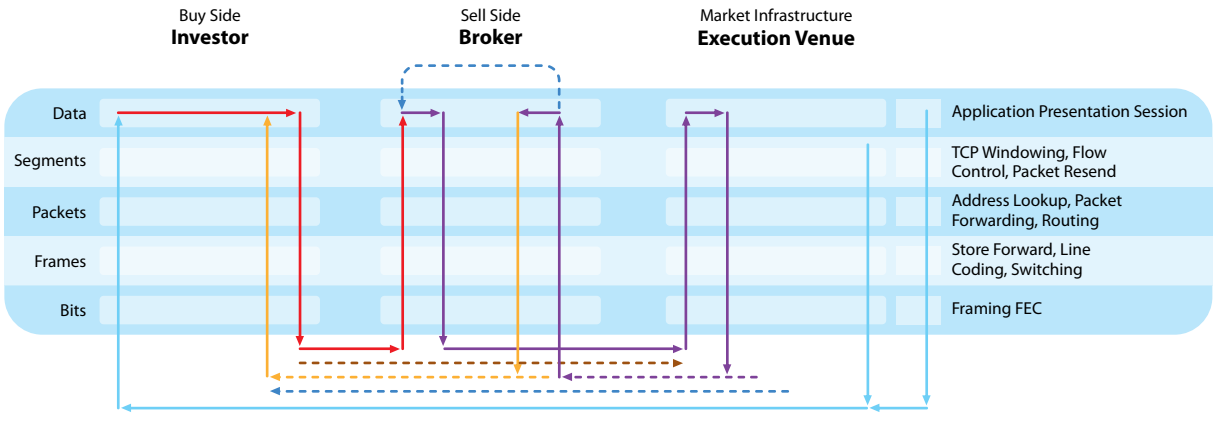
trade' and 'time gap between orders' on a trade book are examples of these imperatives. Another example is the focus on controls for the provision of 'direct market access' by brokerages to their trading clients. An interesting development, as a consequence of the search for liquidity, is 'Rebate Trading' which essentially exploits the incentivizing mechanism offered by exchanges for providing liquidity and spreads, similar to 'Market Making'.

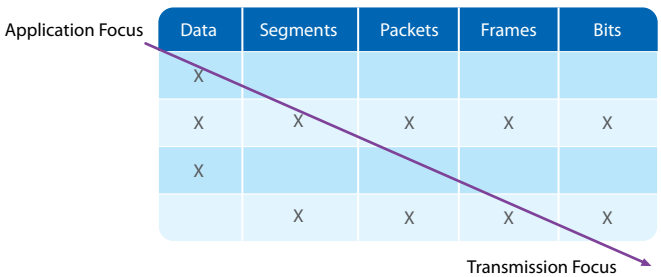
These latency imperatives are most visible in equity markets, specifically in the US, driven by enabling market structures around connectivity and cost. With increased fragmentation of liquidity venues geographically and diversification of investment strategies around instruments, these imperatives can be seen at play across geographies and instrument markets.

PART 2: Low Latency Sources and Manifestations

Manifestation of Latency:

In this section, we look at the sources of latencies and how they manifest themselves. To put the sources of latencies into perspective, let us look at a simplistic technology stack and examine the sources of latency within them:





The technology stack can be represented using a simplistic version of the OSI layers, moving from layer 1, which focuses on bits, to layers 5,6 and7 which deal with data and associated application processing. Communication between market participants (broadly categorized into Buy Side, Sell Side and Market Infrastructure in Fig 3 for simplicity) has to typically traverse from layers 7 to 1 and all the way back to layer 7 and so forth.

From a latency perspective, the layers themselves provide an intuitive assessment of the contribution to latency. If we consider the activities that need to occur at each of the layers, the contribution to latency steadily increases with the complexity of the activities themselves – higher layers contribute more to latency than the lower layers.

These latencies impact business contexts and we can examine the four different segments of the trade flow to identify and gain an understanding of the layer of the technology stack that plays an important role:

Investment Decision to Execution Need:

In this segment, the investment models, which generate decisions, are highly sensitive to latency. – Prevailing market prices and other market data, which change more frequently today, impact pricing models. The transaction cost analysis depends on the liquidity venues which vary dynamically owing to fragmentation and fast turnovers. In short, the effectiveness of investment decisions is increasingly dependent on faster executions. This is more so for asset classes such as equity and forex than for fixed income instruments. Accordingly, algorithms driving investment decisions as well as communication of execution need are under focus. Latencies further down in the technology stack are relatively insignificant for this segment.

Execution Options to Execution Feedback:

In this segment, the processes of searching for liquidity venues and getting into the queue for execution are most sensitive to latency. In the application part of the stack, the algorithms analyzing depth of the markets for smart order routing, ranging from order slicing for a particular market to distribution across several markets, is the key source of latency. Further down the technology stack, the need to get 'ahead of the queue' in a given liquidity venue depends on several factors. These include faster network access to the execution venues; , proximity to and co-location at the execution venues; and low-bandwidth message protocols. In addition, 'variability' or 'volatility' in the capacityand peak speed are of key importance. Jitters or clogs in the networks could create uncertainties which can have significant adverse impact on the execution quality.

Within the trading venues themselves, latency manifests at three levels-order receipt, matching and order communication. At the order receipt and matching level, the key sources of latency are the order receipt capacity and the matching algorithm. At the order communication level, the key source of latency is the network.

Execution Report to Investment Decision:

In this segment, the focus is back on the investment models on the buy side, wherein execution quality drives a more frequent fine-tuning of investment strategy. However, the key source of latency is the integration of the buy-side and sell-side order management systems (OMS).

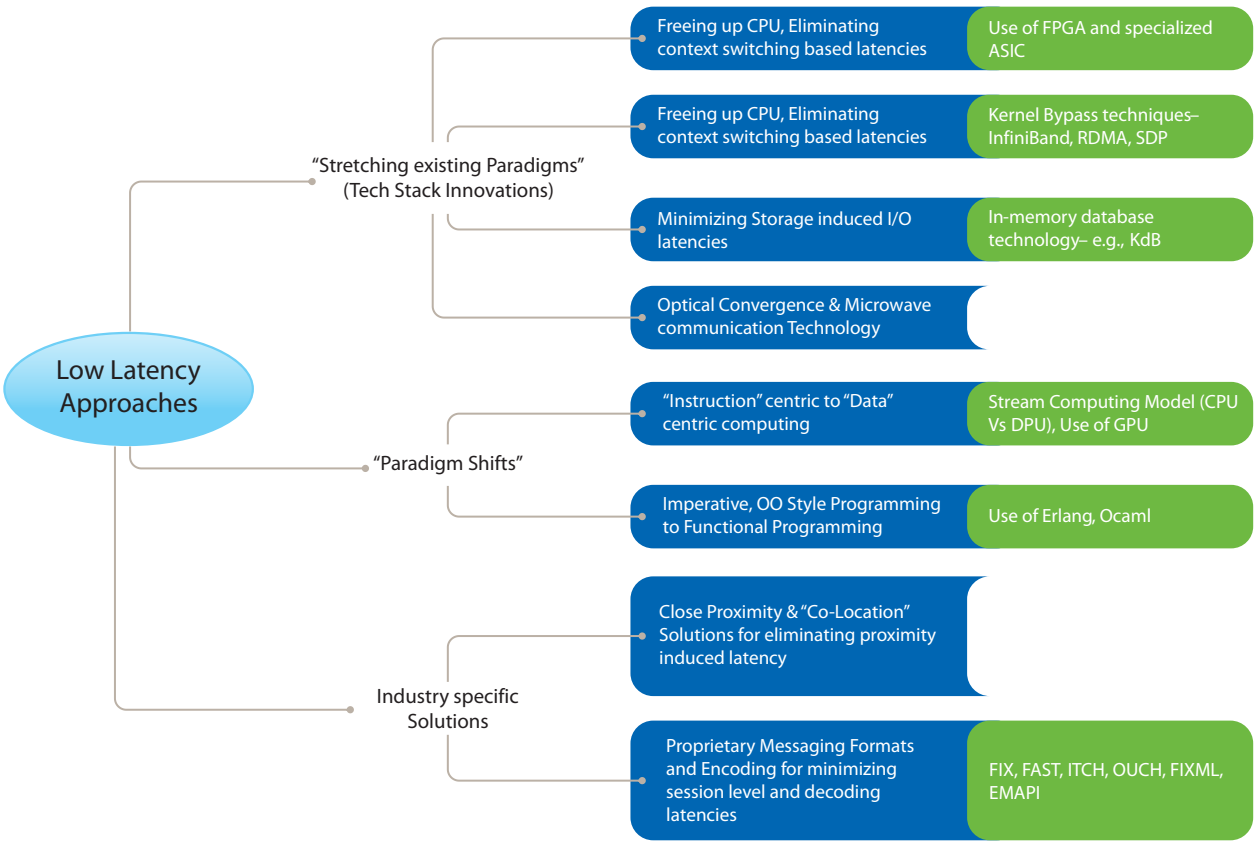
Market Data Feeds for Investment Decisions:

In this segment, the exclusive focus is on the message formats as well as the message distribution mechanisms. For trading venues, the capacity issue, which is at the core of latency, is driving approaches that look at the layers below the application layer. For market data vendors, collation, possible enrichment and distribution is driving approaches focused both on the application level (for innovative ways of parallel processing of market data) and the network level (for distribution of data customized to their customers' needs).

PART 3: Addressing Latency Challenges - Approaches

Low Latency Approaches:

Fig 4 depicts approaches addressing latency challenges. These approaches can be examined from three different perspectives:



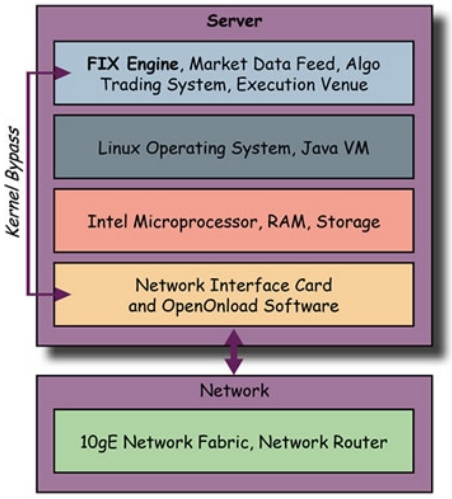
Stretching Existing Paradigms:

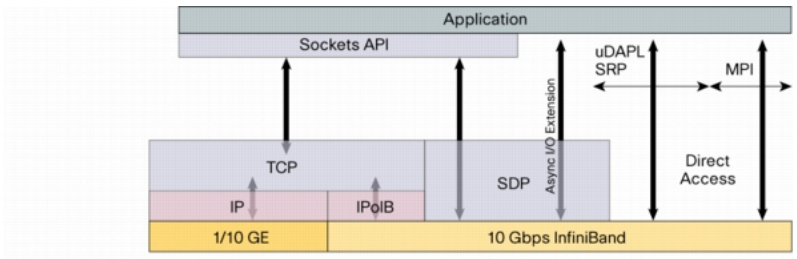
Approaches focused on stretching the existing computing and networking paradigms comprise innovations to the technology stack:

- CPU processing limitations and memory switching induced latency and 'jitter' are resolved through increased use of other specialized computing units such as FPGAs and specialist circuits (ASICs). These are mostly used for faster processing of market data feeds, with some basic level of checking/filtering logic
- Memory context switching induced by the operating system kernels use 'kernel' bypass techniques, which make frame level data accessible directly to user level programs. This can be achieved through the use of InfiniBand, RDMA and SDP in the data fabric for faster processing
- The latency induced by storage I/O in database processing has been addressed by increasing innovations in 'in-memory' database use. The particular success of products such as KdB illustrates this approach.
- On the network layer, propagation latencies have been addressed by adopting approaches around optical convergence and use of microwave for ultra-low latency transmissions

A few of these approaches are depicted below:

(a) Kernel Bypass Techniques (Source: Aviat Networks, Cisco)





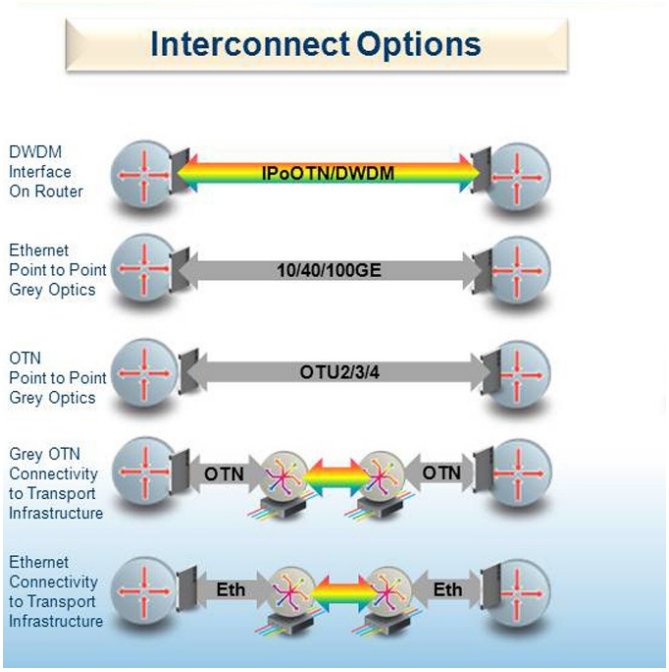
Throughput	1 Gbps	3.9 Gbps	4.1 Gbps	4.5 Gbps	7.9 Gbps	8 Gbps	8 Gbs
Latency	40 usec	40 usec	20 usec	15 usec	15 usec	8 usec	4.5 usec
CPU Utilization	30%	80%	85%	25%	4%	1%	>1%

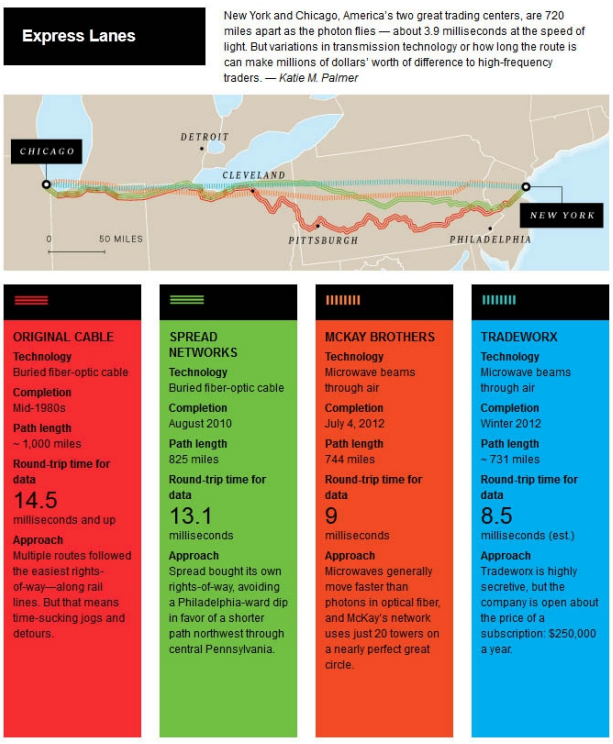
MRI: Message Passing Interface
SRP: SCSI Remote Protocol
uDAPL: User-Level Direct Access Programming Language

(b) Use of specialized circuits for 'lean'' logic processing (Source: HFTreview.com)

	Standard 10GE Network Card	Low Latency 10GE Network Card	FPGA	ASIC
Latency	20 micros + application processing	5 micros + application processing	3-5 micros	Sub-micro
Ease of Deployment	Trival	Kernel driver installation	Retraining of programmers	Specialist
Man Years Effort to Develop	Week	Weeks	2-3 man years	2-3 man years
Elapsed Time	Week	Weeks	6 months - year	Year +
Costs	$50 - $200	$500+	$100 - $20,000	$1million +

©Networking Propagation Technologies (Source: Businesswire.com, zerohedge.com)



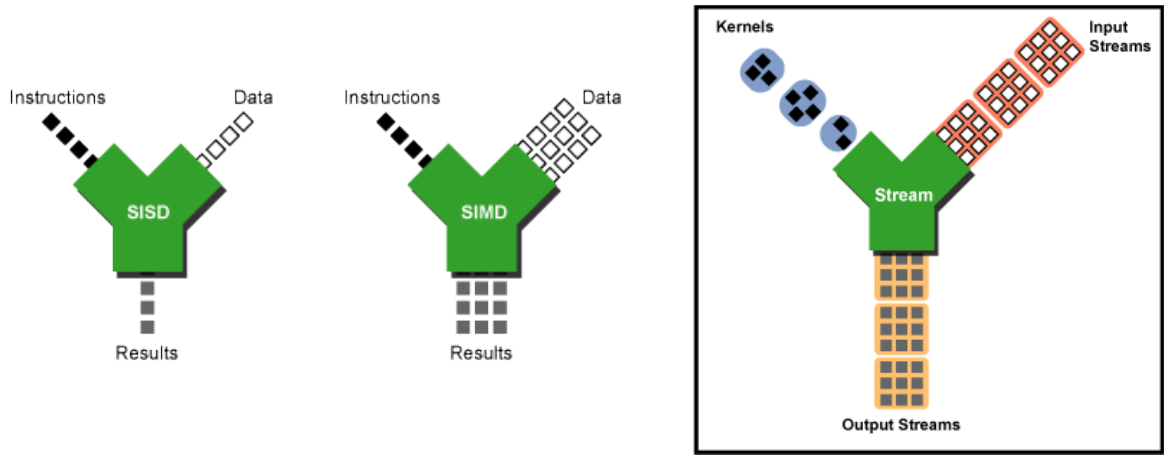


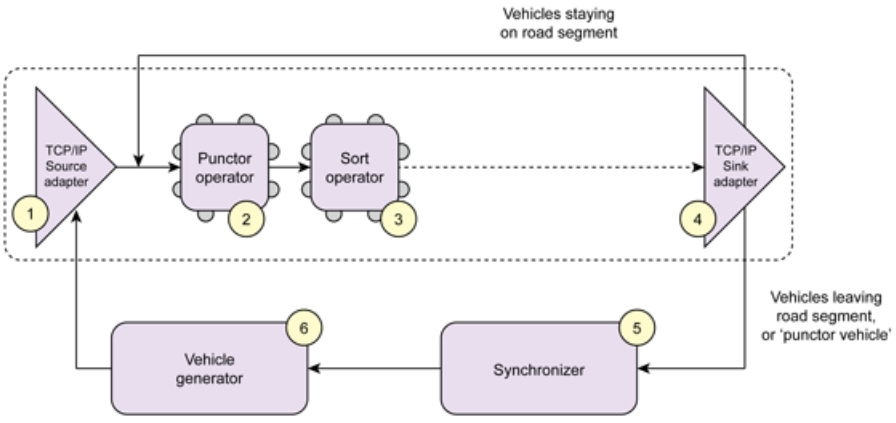
Paradigm Shifts

The two key developments responsible for paradigm shifts in addressing low latency challenges are:

- Computing paradigms are seeing a shift from instruction centric processing to a data centric processing. In essence, this paradigm involves a shift from a computing engine focused on instruction stream sequencing to data stream sequencing with an emphasis on 'parallel computing'. This shift is exemplified by 'Stream Computing', which in essence, turns conventional data processing on its head, replacing the 'store-then-process' approach with a 'process-on-the-go (data-in-motion)' approach. The IBM Infosphere streams product is a good example of the stream computing approach.
- Programming language adoption is seeing a shift from imperative and object oriented (OO) languages to functional languages. The key drivers of this shift are the 'side effects' and limited concurrency in the imperative and OO models, which the functional languages handle better through their inherent immutability. The use of Erlang and OCaml in the algorithmic programming domain exemplify this shift of preference.

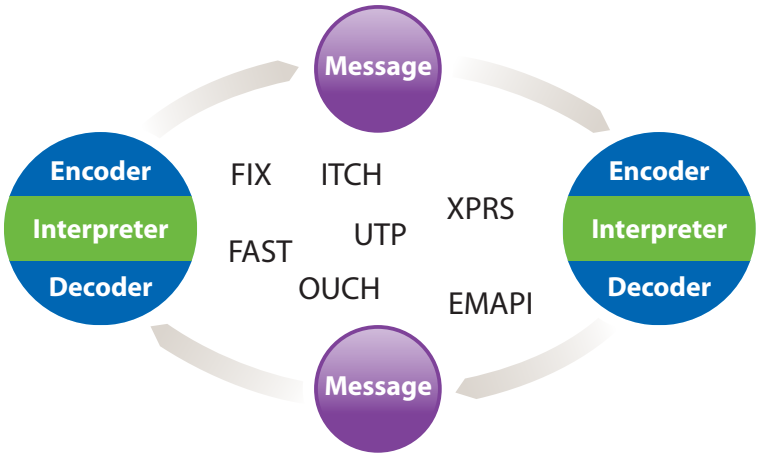
Fig depicts the stream computing paradigm shift(Source: sciencedirect.com, ibm.com):





Industry specific approaches:

Industry specific solutions in capital markets primarily focus on improving proximity through co-location and inter-communication between participants through adoption of better messaging formats. Co-location services offered by trading venues are eliminating proximity induced latencies. The proliferation of proprietary formats and protocols is centered on the debate around speed vs. flexibility. While FIX enjoys widespread adoption, many market infrastructures support it as a secondary connectivity option, preferring proprietary messaging formats and encoding to enable faster interconnectivity with market participants. The use of ITCH and OUCH by NASDAQ and UTP by the New York Stock Exchange (NYSE) exemplify this approach.



PART 4: Summary and Outlook

Summary

This section summarizes the key imperatives and approaches discussed in the earlier sections.

The previous sections have progressively discussed low latency in Capital Markets starting with the context of low latency, its manifestation (primarily in the trading chain) and finally the approaches for dealing with the various challenges imposed by low-latency. Fig depicts a consolidated view of latency challenges facing capital markets:

Latency Challenges - Market Participants in Trading Chain

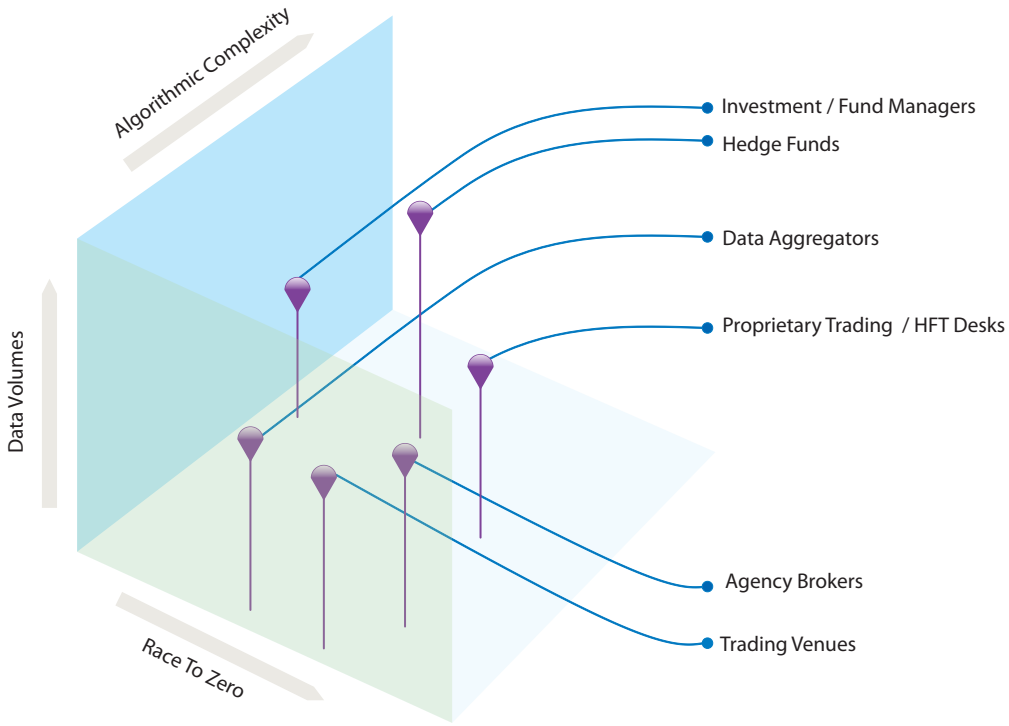
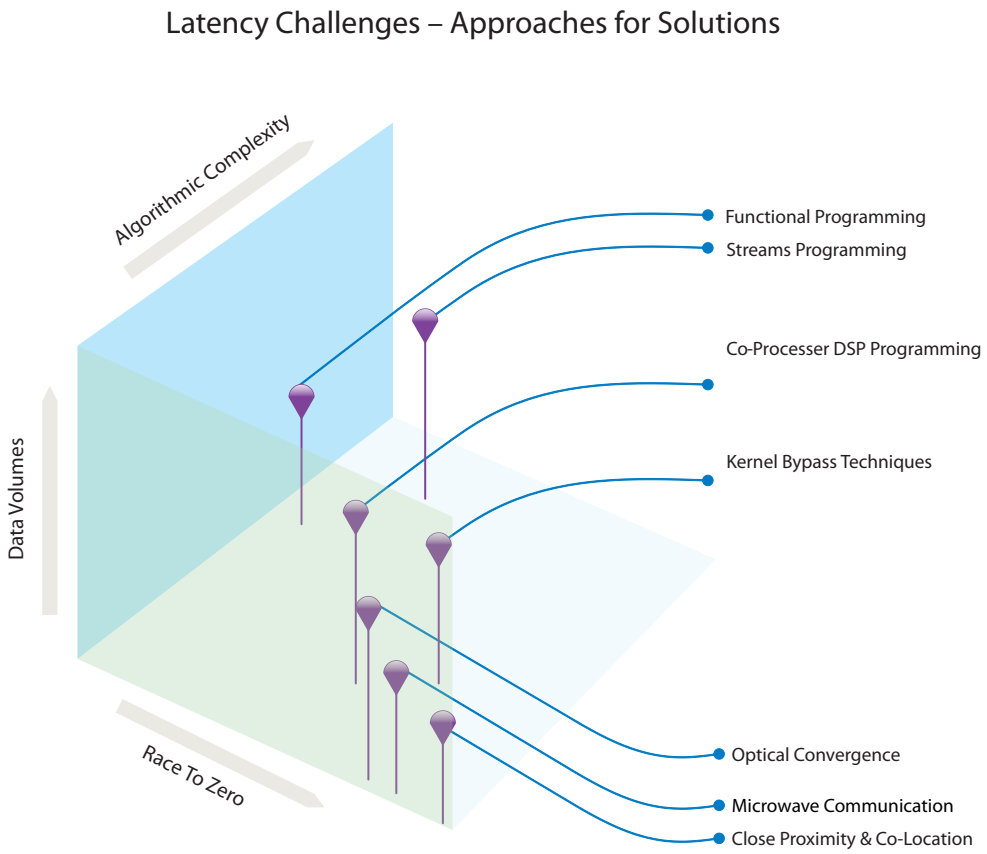


Fig depicts the approaches for dealing with latency challenges:



Outlook:

The existing latency imperatives and the approaches for addressing the latency challenges would result in the following developments:

Globalization and Diversification

The key impediments to globalization of computerized trading strategies are (a) lack of convenient connectivity solutions with acceptable latency and (b) trading costs at execution venues. On both these aspects, significant improvements are underway. On the connectivity front, Fig illustrates the latencies between the top 10 financial centers and the improvements in connection speeds (source: capacitymagazine.com)

Latency benchmarks between the top 10 financial centres

Internet / Lightspeed

	Shanghai		Hong Kong		Tokyo		Singapore		London		Zurich		Frankfurt		Paris		Chicago		New York	
Shanghai	–	–	30	12	55	17	65	38	230	91	260	89	265	87	260	92	195	112	215	117
Hong Kong	30	12	–	–	55	29	35	26	265	95	325	92	330	91	320	95	215	124	235	128
Tokyo	55	17	55	29	–	–	75	53	255	95	270	95	275	92	260	96	170	100	185	107
Singapore	65	38	35	26	75	53	–	–	205	107	275	102	270	102	225	106	230	149	260	152
London	230	91	265	95	255	95	205	107	–	–	20	8	15	6	10	3	90	63	75	55
Zurich	260	89	325	92	270	95	275	102	20	8	–	–	10	3	15	5	115	71	90	63
Frankfurt	265	87	330	91	275	92	270	102	15	6	10	3	–	–	15	5	105	69	95	61
Paris	260	92	320	95	260	96	225	106	10	3	15	5	15	5	–	–	105	66	85	58
Chicago	195	112	215	124	170	100	230	149	90	63	115	71	105	69	105	66	–	–	20	11
New York	215	117	235	128	185	107	260	152	75	55	90	63	95	61	85	58	20	11	–	–

On the costs front, increasing competition by alternative liquidity venues- Chi-X-in Europe, Canada and Australia have significantly lowered trading costs by up to 50%. With both these impediments melting, a significant focus on globalization and diversification of computerized trading strategies can be expected in the near future.

New Technology Focused Offerings

With increasing focus on latency, new technology offerings in the value chain can be expected–especially in pre-trade analytics around 'intelligent' processing of market signals and in renewed focus on latency 'measurement' (extending the niche created by the likes of TS-Associates, Correlix and Corvil).

Increasing Emphasis of Smart Order Routing

With algorithmic trading dis-intermediating the traditional execution centric business model of the sell side, smart order routing capabilities driven by fragmentation, globalization and diversification would emerge as a focus area.

Technology Services Offerings Shift to Functional Paradigm

With increasing emphasis on low latency and parallelism, technology services are likely to be increasingly based on functional programming models, moving away from object oriented and imperative styles.

Reference Acknowledgements

1. Washington Blog (www.washingtonsblog.com) - April 2012
2. Sydney Morning Herald (www.smh.com.au) - October 2012
3. BIS Report - High Frequency Trading in Forex Markets - September 2011
4. Eric Schulte - Non-Von Neumann Computation - University of New Mexico - May 2010
5. The Bureau of Investigative Journalism (www.thebureauinvestigates.com) - Infographic: Trading at the speed of light - September 2012
6. IOSCO - Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency (www.iosco.org) - October 2011
7. Altera Corporation - Accelerating High Performance Computing with FPGAs - October 2007
8. ADVA - Ultra Low Latency Financial Networking - May 2010
9. Latency Stats (www.latencystats.com) - Latency Transparency for Market Data - September 2012
10. Robert Litzenberger et al - The Impacts of Automation and High Frequency Trading on Market Quality
11. Gideon Saar et al - Low Latency Trading
12. David H. Jones et al - GPU Versus FPGA for high productivity computing
13. Christian Leber et al - High Frequency Trading Acceleration using FPGAs
14. Srinivasan S. et al - Understanding Global FInancial Crisis Through Systems Dynamics

TATA CONSULTANCY SERVICES

Experience certainty.

Contact

For more information, contact **bfs.marketing@tcs.com**

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w

Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at **www.tcs.com**

IT Services
Business Solutions
Outsourcing

TCS Design Services | M | 03 | 13