

The Role of Latency in Financial Markets Technology and the Impact on State-of-the-Art Financial Exchange Data Centres

James Dow, Chief Technology Officer
CS Technology



DatacenterDynamics

SYDNEY 2010

Today's Premise

Exchanges will drive significant data centre activity:

- Exchange operators will drive significant amounts of data centre activity over the coming decade
- If data centre operators are to accommodate the needs of this industry, they first must understand it
- Latency management and market dislocation are the dominant characteristics to understand



Why Do Exchanges Care About Data Centres?

It's a sourcing question – What do we need and how do we get it?

- | | |
|--|--|
| 1) Deploy proprietary data centre capacity to meet core processing needs | <ul style="list-style-type: none">• \$40-\$60M capital spend• Does not provide any facility to accommodate HFT requirements <hr/> |
| 2) Deploy proprietary data centre capacity to meet core processing needs and co-location needs | <ul style="list-style-type: none">• \$240-\$300M capital spend• Accommodates HFT• Adds large burden on balance sheet <hr/> |
| 2) Partner with data centre provider for core and co-location needs | <ul style="list-style-type: none">• No or little capital requirement• Places core business value-add components in the hands of outsiders |

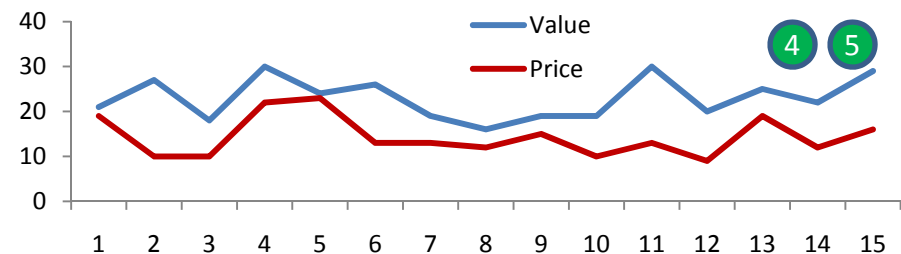
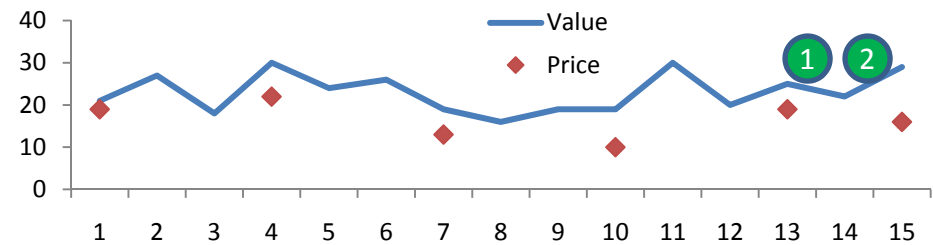
An exchange's strategic approach to acquiring data centre capacity is closely coupled to their strategic approach to the High Frequency Trading Business. This has tremendous impact on the structure, survivability, and economics of the core business.



Understanding the Trading Business

It begins with Econometrics 101:

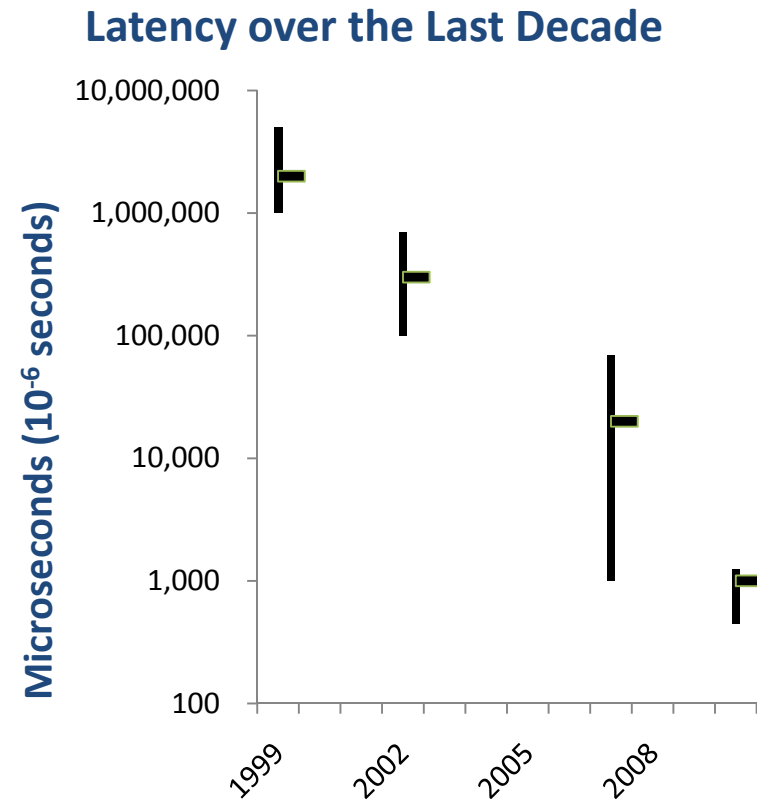
- 1) “Value” is a continuous
- 2) “Price” is discrete
- 3) More frequent pricing of an instrument creates a pricing curve that more closely approximates the value curve
- 4) By pricing frequently enough, we can drive the mathematically discrete pricing curve to exhibit characteristics of a continuous function
- 5) Continuous pricing creates a more efficient market – efficiency in this context is economic efficiency that a price is created that is an accurate representation of the value of the instrument



What Does “Frequently” Mean?

Frequently is a synonym for latency management

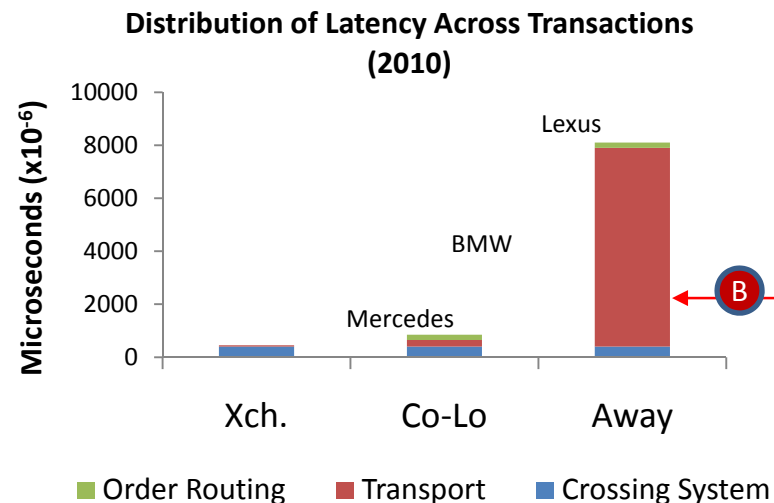
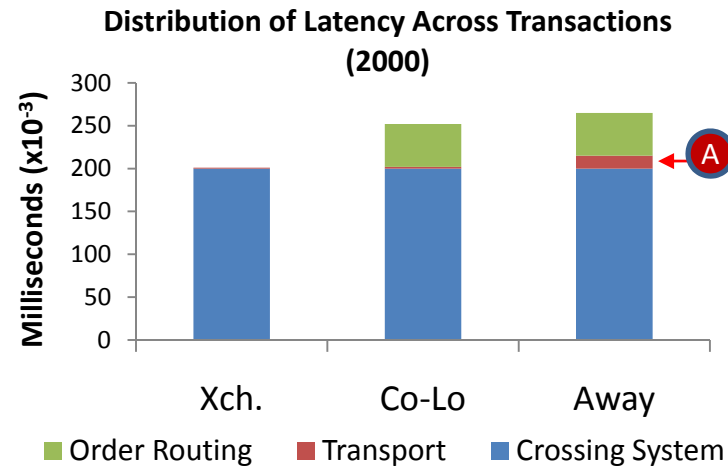
- Latency is the term used to describe the time lag between beginning and end of a transaction
- In 1999, latency was measured in 1's of seconds, *i.e.*, 10^1 seconds
- By 2002, latency within the equities and options worlds had fallen to milliseconds, *i.e.*, 10^{-3} seconds
- By 2007, latency had fallen into the range of single digit milliseconds, *i.e.*, 10^{-3} seconds
- Today, some exchange routinely operate at the range of hundreds of microseconds, *i.e.*, 100×10^{-6} seconds
- Business consumers make decisions based on fibre length, the impact of which is measured at 10^{-9} seconds in systems running at 10^{-5} seconds



Role of Fixed Distance Latency

Transport latency is a physical fact:

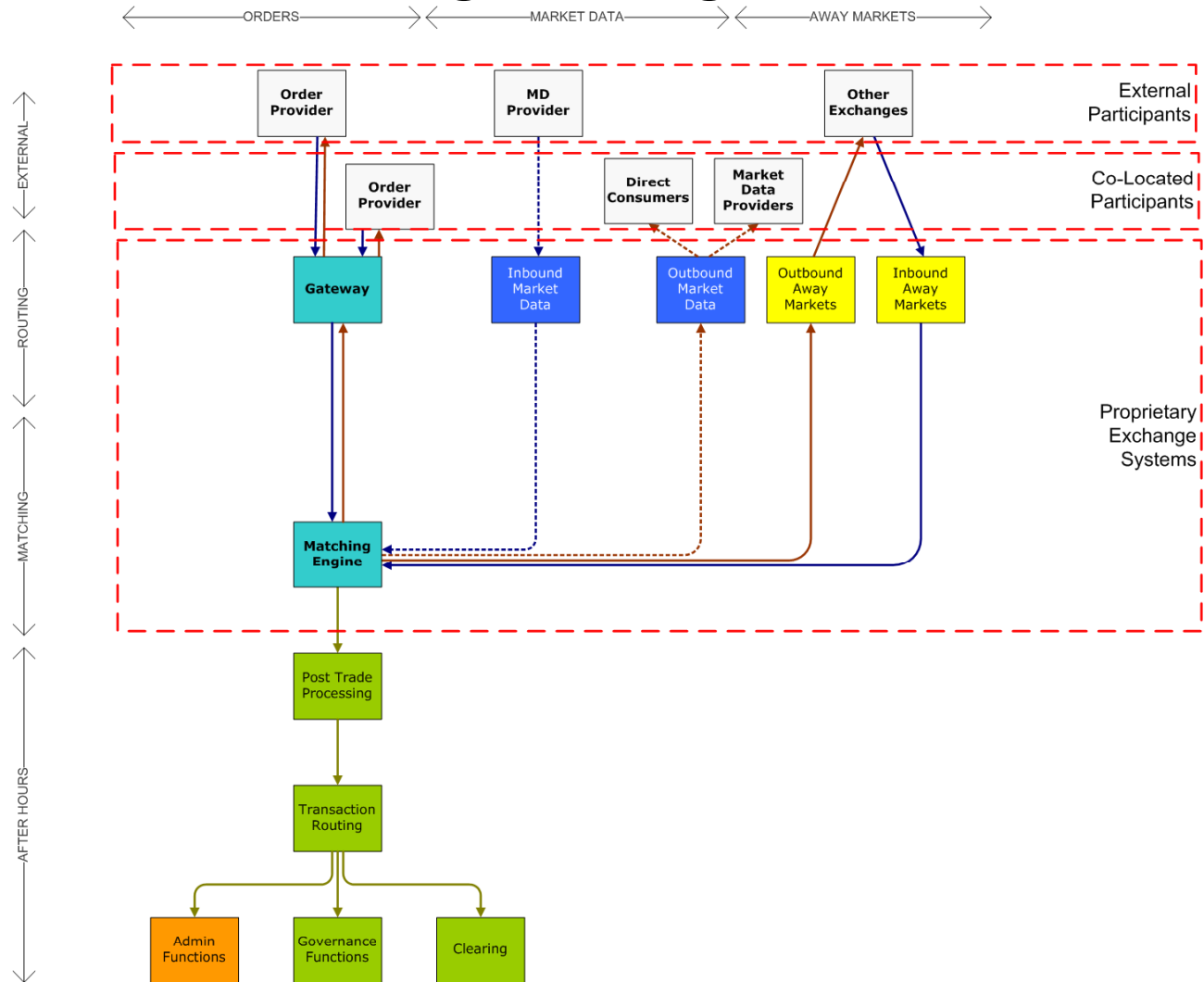
- Physical transport latency is governed by physical constants, *i.e.*, *speed of light*, *O-E-O translations*, *processing*
- Technological improvements have **decreased** transport latency since 2000, but has been outpaced by transactional speed improvements
- In 2000, transport latency was the smallest contributor to overall transactional latency – by 2010 it has become the largest, driving the industry to co-location with exchanges



In this example, transport latency at **B** is only 50% of **A** but seems much larger due to scale

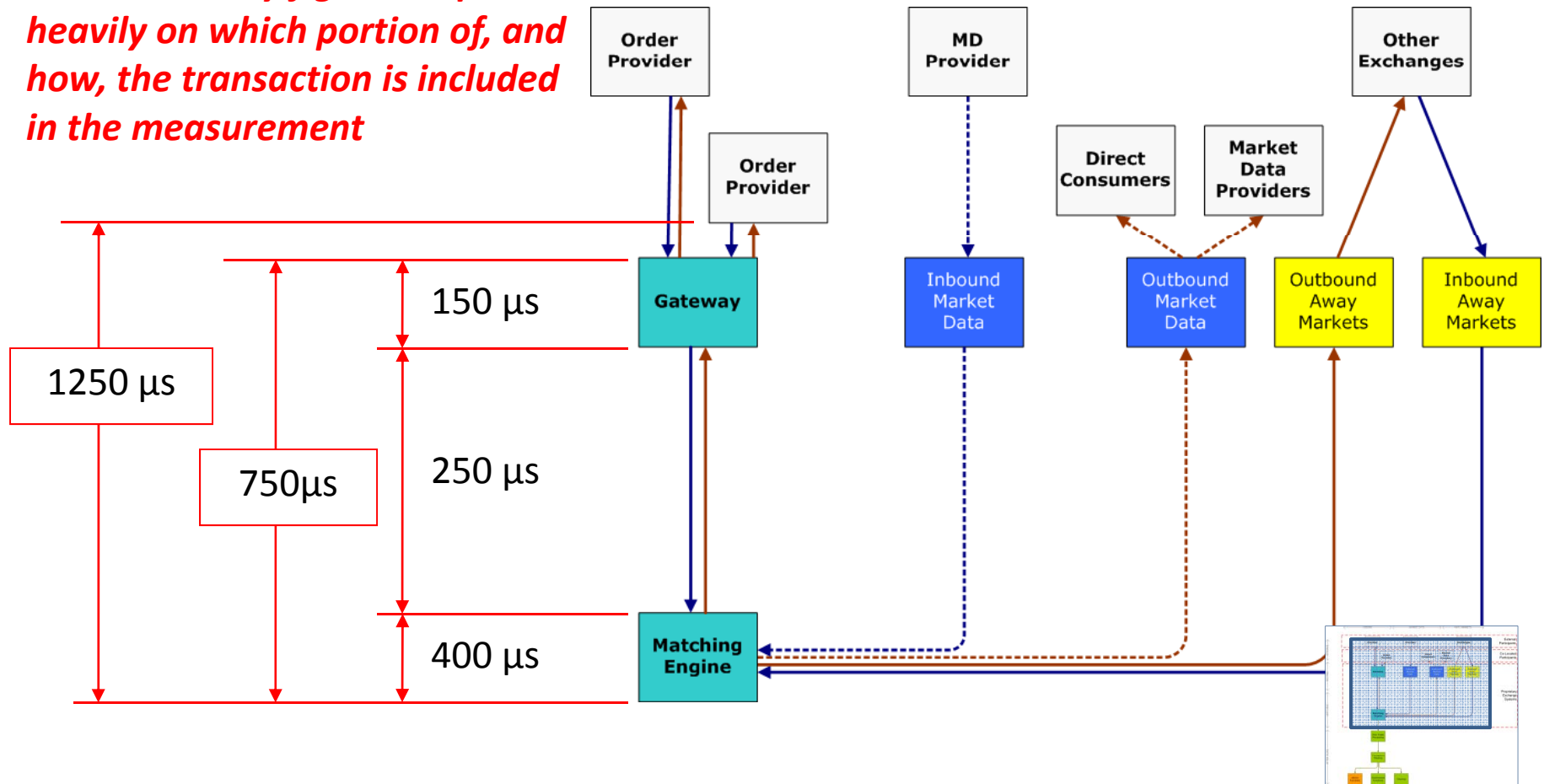


Understanding Exchange Execution



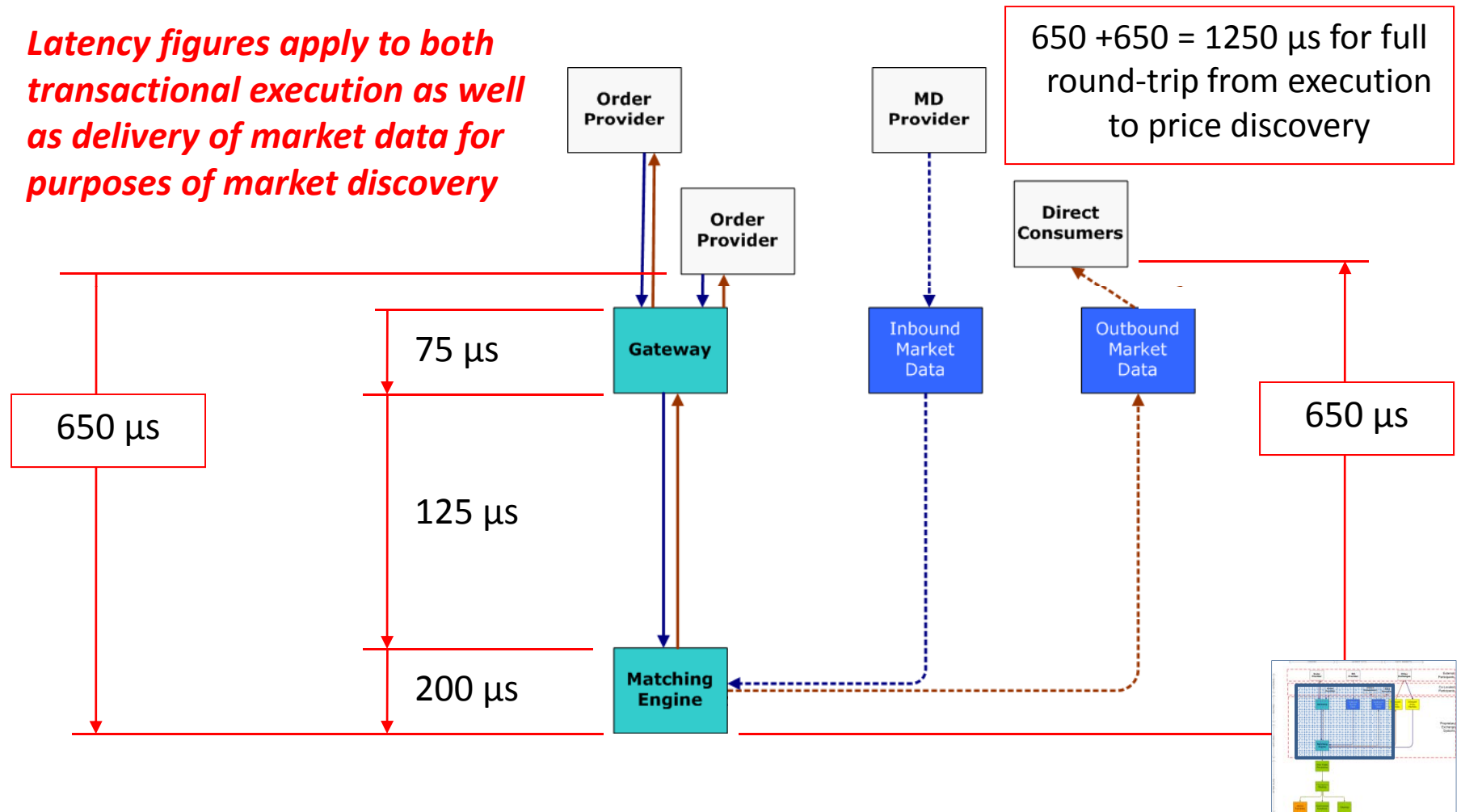
Today's Execution Latencies

Quoted latency figures depend heavily on which portion of, and how, the transaction is included in the measurement



Today's Execution Latencies

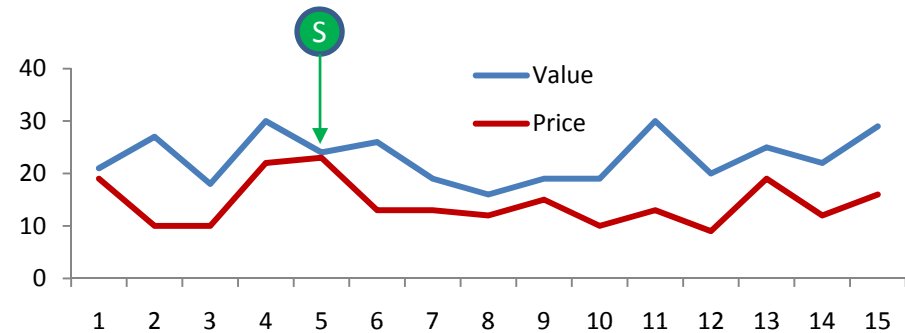
Latency figures apply to both transactional execution as well as delivery of market data for purposes of market discovery



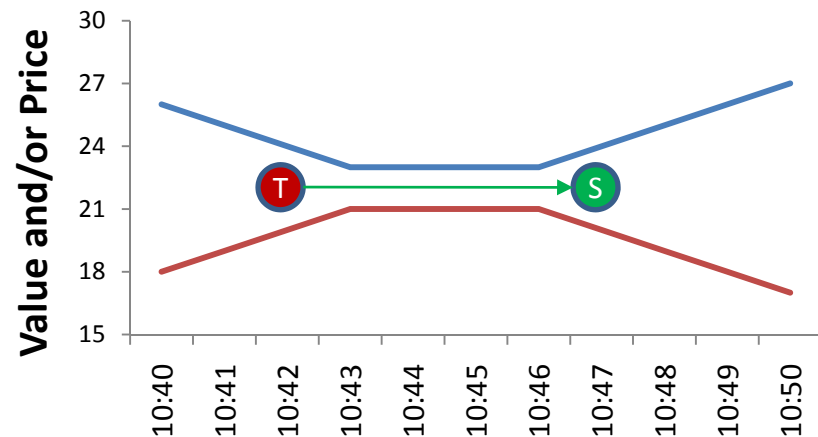
Latency Impact on Execution and “The Strike”

Going back to Value versus Price

- A “strike” occurs when the value and price coincide within narrow enough boundaries that both participants choose to execute a transaction, point **S**
- As the Price curve approaches the Value curve, a transaction “throat” is created **T**
- A strike may occur at any point within the transaction throat
- Duration of the transaction throat is dependent on the volatility and latency of the market



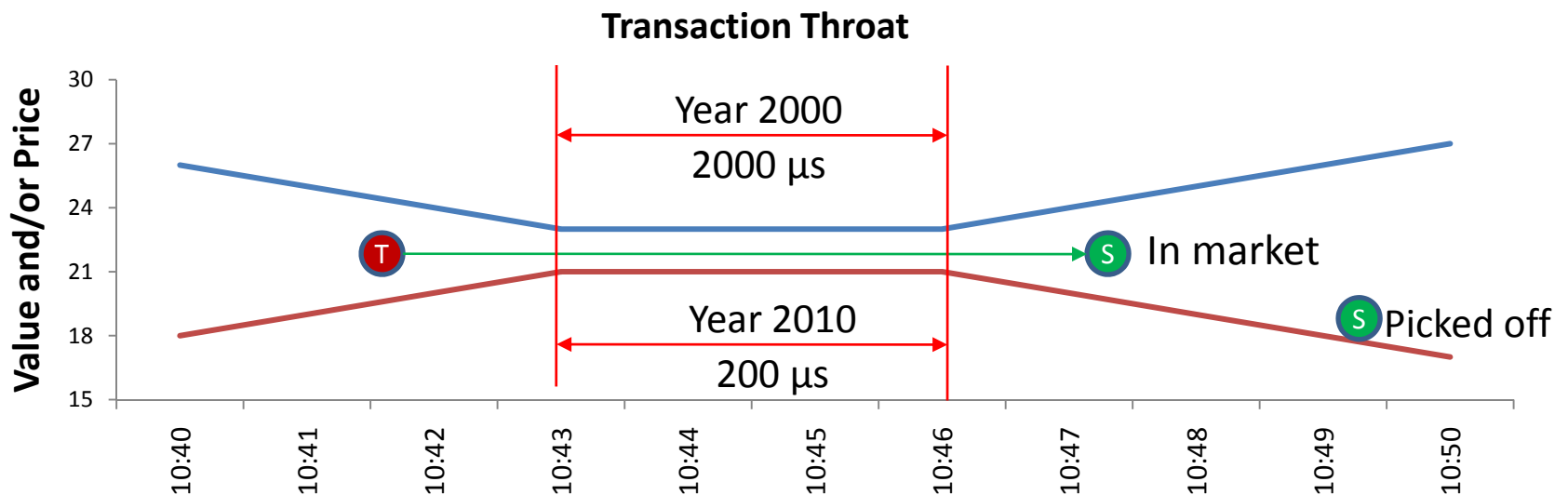
Transaction Throat



How Big Is the Transaction Throat?

Transaction throat shrinks with latency

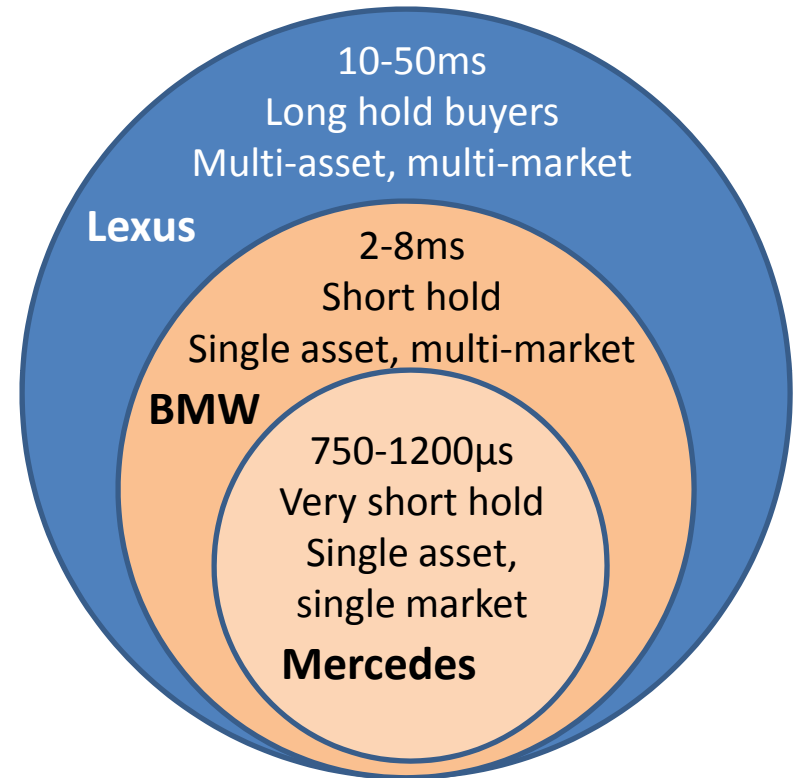
- Transaction throats in 2000 measured in milliseconds to seconds
- Transaction throats in 2010 measure in 10's to 100's of microseconds
- A strike out of the throat represents buying “out of the market”
- Both width, *i.e.*, *time duration*, and height, *i.e.*, *spread between price and value*, shrink as latency decreases



Differentiating Market Participants

Trade strategies influence latency sensitivity:

- Participants break into three classes
- Sensitivity to latency and width of transaction throat decrease from Mercedes to Lexus class
- “Not all participants exhibit the same sensitivities and buying patterns” – one size no longer fits all for exchange services
- Each class of participant, and member of the class, seeks to optimise the end-to-end trading stack for efficiency based on their proprietary trading models and strategies
- Hold times vary from seconds to years



Mercedes class participants must be within the latency thresholds, or transaction throat, for their strategies to be economically viable

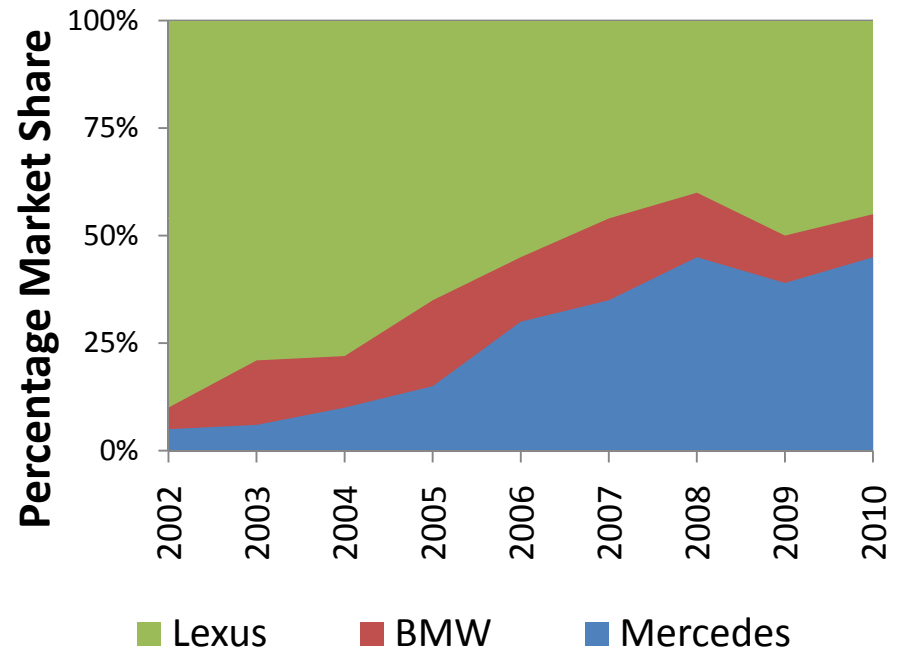


Influence of HFT Participants

HFT participants drive market liquidity:

- High Frequency Trading (HFT) now contribute the largest percentage transactional share by trading class
- Once growing, remote HFT is now a shrinking asset class
- Percentage of market share to HFT is inversely proportional to venue latency; as latency falls, HFT participation increases
- HFT participation drives basic market liquidity, depth of liquidity, and more efficient pricing
- ***High Frequency Trading has emerged as a critical and necessary component to exchanges seeking to compete in the global capital markets environment***

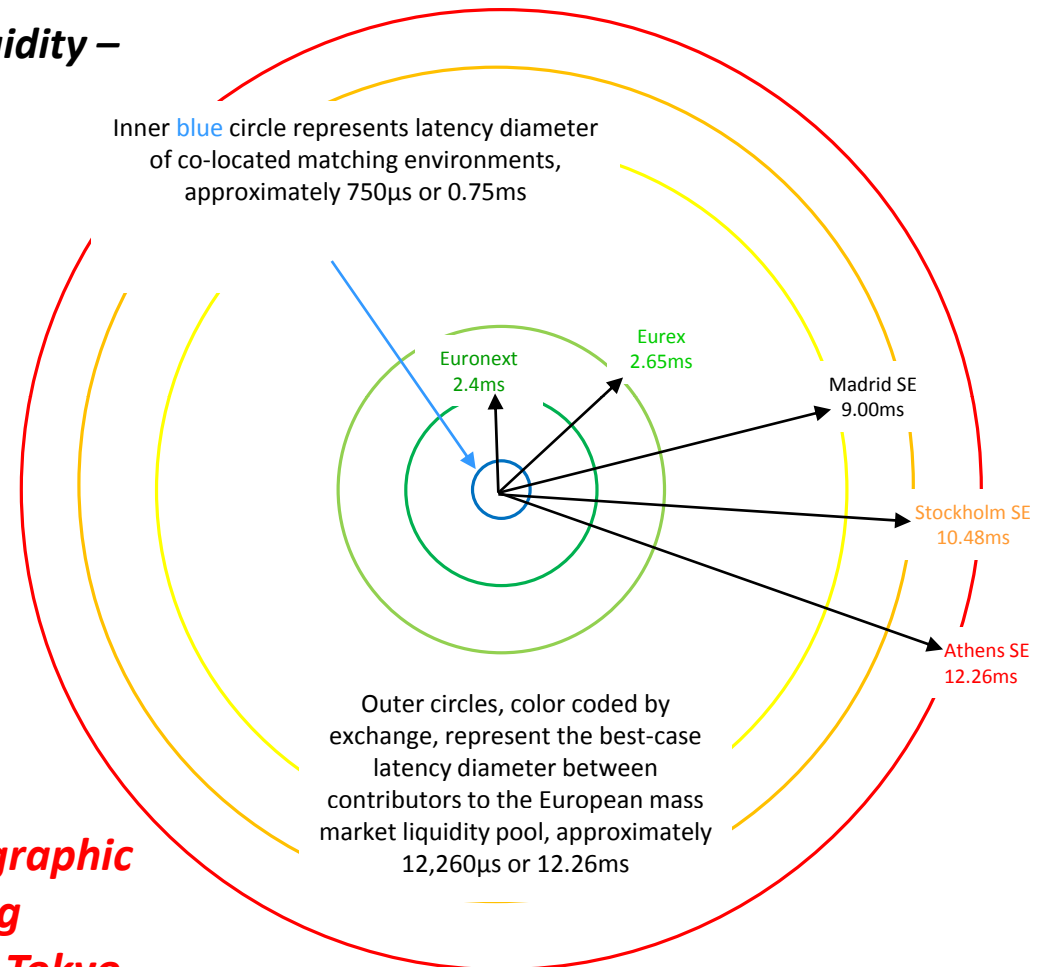
HFT Market Participation



Impact of Latency on Markets and Liquidity

Reduced latency creates islands of liquidity – Geographic fragmentation:

- Market expectation of “0 latency at 0 cost” creates a shrinking geographic diameter over which liquidity is relevant
- Fixed location markets are becoming increasingly isolated as latencies decrease
- Market and liquidity consolidation are emerging factors with which exchange operators must address
- ***A small number of specific geographic locales will emerge representing global availability of liquidity – Tokyo, Singapore, HK, Frankfurt, London, etc.***

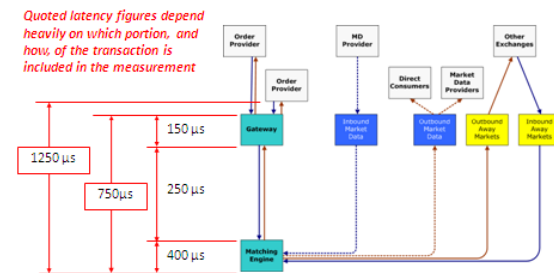


Exchanges in the Present Day

Key factors to remember:

- Exchanges are notoriously protective of their latency figures, often obscuring either measurement points or the actual round-trip times
- No latency figures from this presentation apply to any specific exchange, but rather the broad behaviours across the industry
- Matching engine latencies generally measure in the 300 μ s range
- Full end-to-end latencies are in the 900-2000 μ s range for “fast” exchanges; others continue to operate in the 300 millisecond range
- Disparity between exchanges in execution speeds has increased
- “Clock walking” and negative time

Latency In Today's Execution Landscape



Please note that we are approaching “Heisenberg” limits of our technology – In mass-produced electronic systems, we cannot reliably measure a single events with durations shorter than $\sim 8\mu$ s due to manufacturing disparities. We can measure effects through aggregation of 1000's of events, but this introduces the impact of statistical distributions and long tails of anomalous effects



Impact of Regulation on Markets and Liquidity

Competition creates structural fragmentation:

- Markets may or may not have exclusive rights to their instruments – in the US there are over 33 separate venues for Over-The-Counter (OTC) equities
- The question arises as to what right national governments have to their industrial equity, *viz., Irish equities trade on a German exchange operator in Frankfurt*
- Some markets may trade not for their instruments, but for their liquidity, particularly in futures and other derivatives markets
- Fungibility (*or ability to move between venues*) of either instruments or liquidity reduces the security or hegemony a single operator can claim on a market
- Increased fungibility, regulatory permissiveness, competition, discrepancies in the global regulatory environment all contribute to create structure fragmentation, which increases opportunities for arbitrage
- All market operators are struggling to define their business models in today's environment

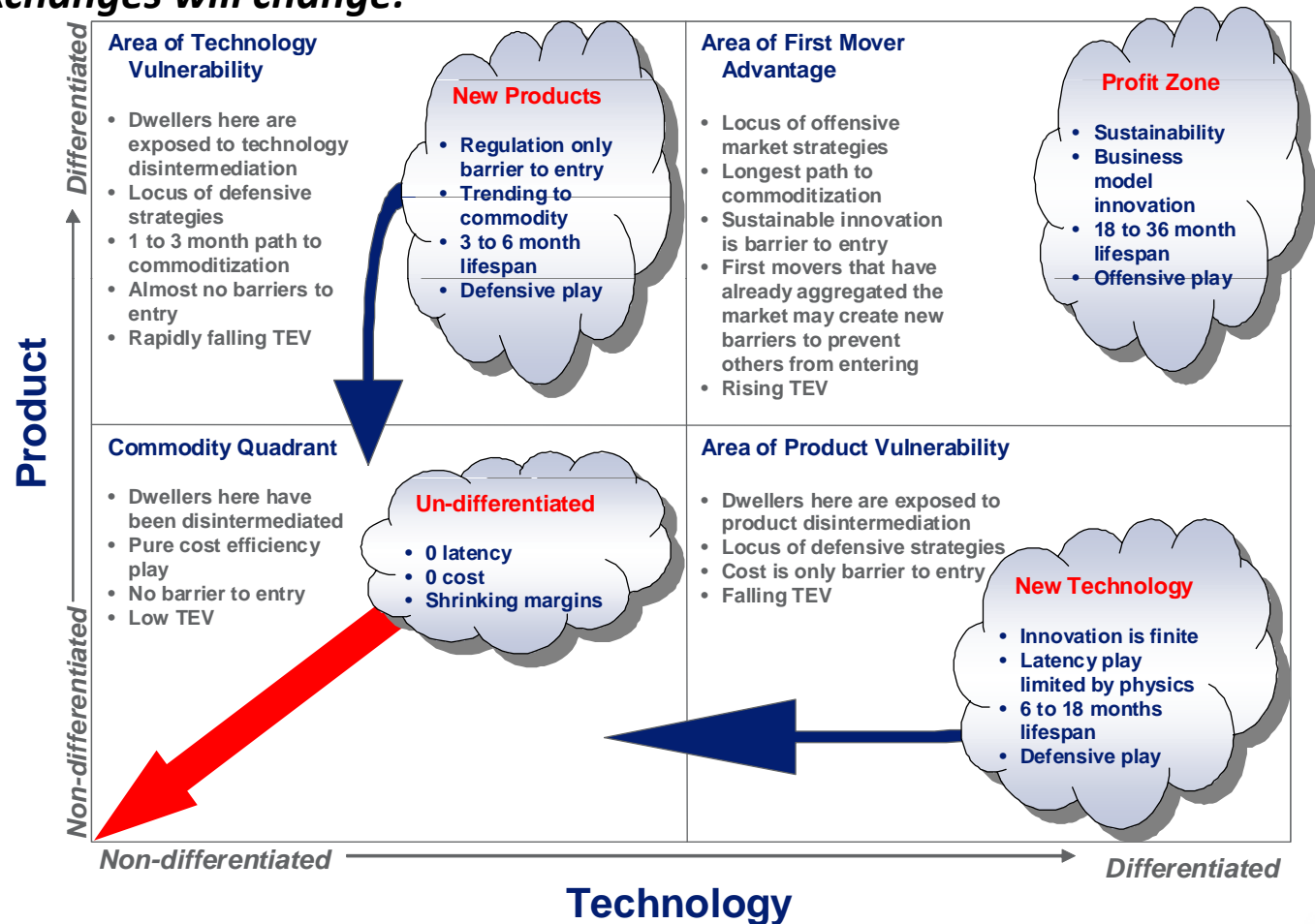


Is Exchange Commoditisation Inevitable?

Business models for exchanges will change:

Exchanges must cope with:

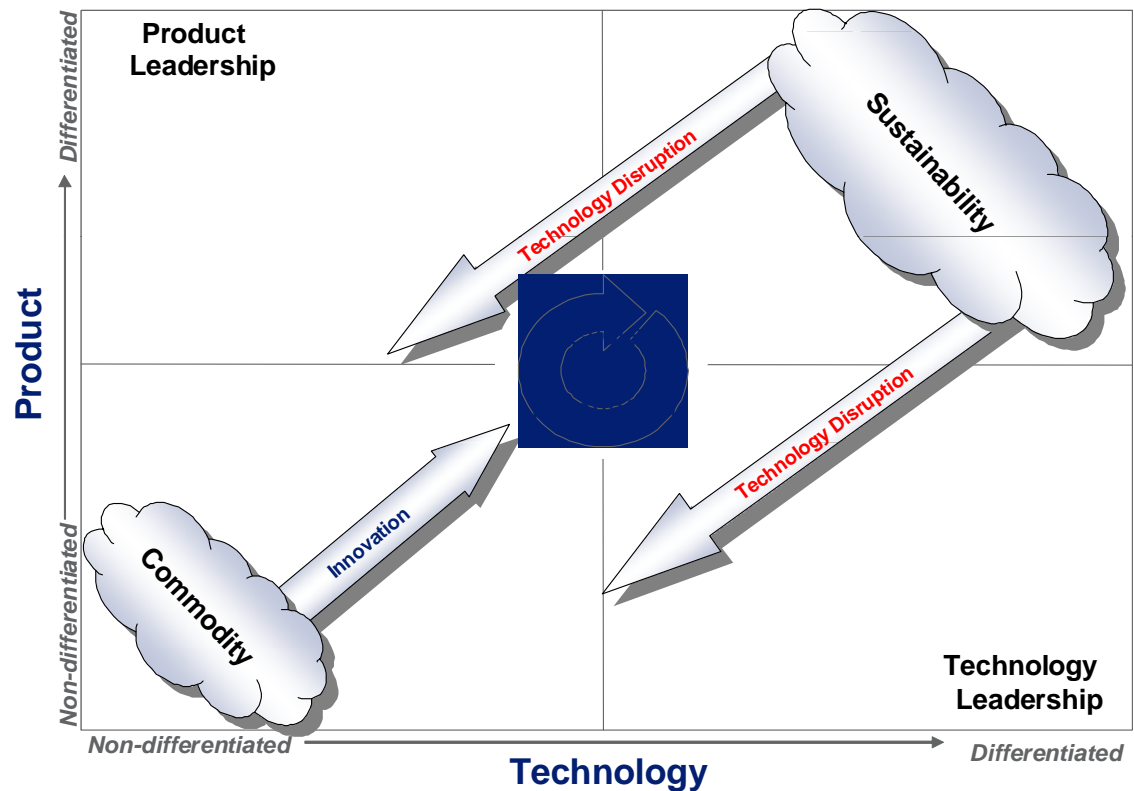
- Fragmentation, both structural and geographic
- Rising technology costs to drive down latency
- Shrinking margins
- Increasing market volatility
- Uncertain regulatory climate



Technology and Product Innovation Cycle is Accelerating

Business models for exchanges will change:

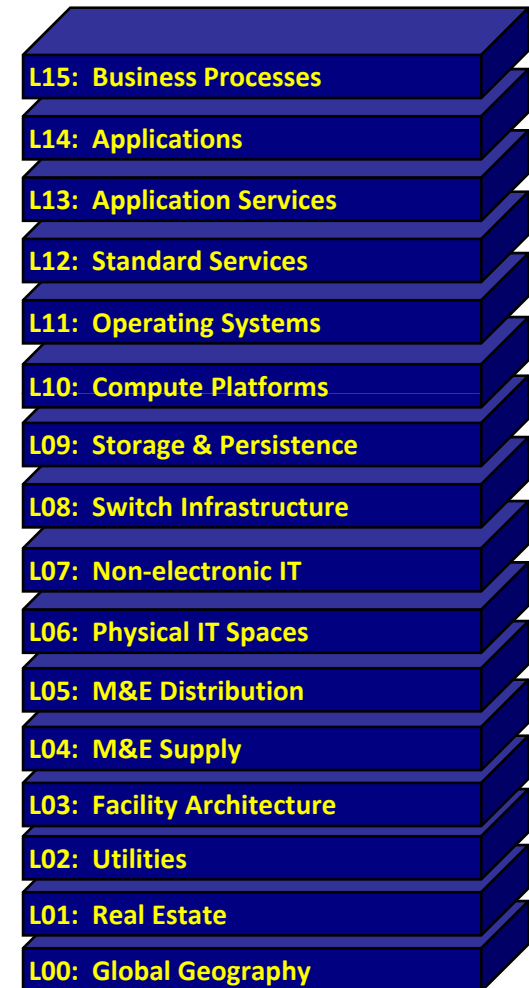
- Product innovations no longer enjoy regulatory protection and are rapidly copied
- Technology innovations have short lifespan
- Rigid business models provide short-term runway
- Vertical or horizontal integration of products and technology are required to survive



The Business of Markets

Providing integration of the entire stack:

- Margins on transactions are shrinking; other products and services must be brought to market to drive revenue and provide an integrated, optimised stack
- Proprietary portions of the stack provided internally, *i.e., license to trade, crossing engine technology*, others through partners
- Uniqueness must be preserved to protect business – primary uniqueness comes from existence as a licensed exchange; secondarily from liquidity created by community
- Trading licenses, competition, and asset fungibility can be eliminated by the stroke of a regulatory pen, so generation of community and liquidity become paramount concerns



Economic Considerations

Differentiation within the Data Centre Co-location market

- The co-location market segments into three discrete types of capacity, differentiated by quality, use, and pricing structure
- Tranches (or types) of co-location service available on the external market:
 - Commodity Co-Location Services
 - Financial Services Co-Location Services
 - Electronic Trading Co-Location
- Key metrics for co-location services, *e.g., average w/m² density, average kW/cabinet density, average contract size and length*, vary significantly by tranche
- Power based pricing, *i.e., pricing based on kW provisioned and powered versus square footage allocation*, varies significantly by tranche
- Note: these comparisons effective only for loads of 50 kW or more



Economic Considerations

Commodity Co-location characteristics:

- Generic co-location capacity without a significant differentiator (baseline case)
- Mostly constructed as Tier II or Tier III facilities
- Available in multiple primary, secondary, and tertiary real estate markets
- Customer base is predominantly non-financial enterprises, small businesses, and some financial services
- Contracts are generally cabinet based, *viz., not power-based pricing, in units of 5 to 50 cabinets*
- Contracts are held by IT and average 18-24 months with renewal clause price protection
- Average power density: 500-750 W/m²; 2-4 kW/cabinet
- [US] Market price point (per kW/month): \$500-\$650 (primary markets); \$350-\$475 (secondary markets)
- Limited and declining availability; investment mostly speculative in secondary markets by 2nd tier providers



Economic Considerations

Financial Services Co-location characteristics:

- Co-location differentiated by catering to predominately financial services clients
- Mostly constructed as Tier III or Tier IV facilities
- Mostly available in primary real estate markets
- Customer base is predominantly large scale financial services enterprises
- Contracts are generally power-based, *viz., not cabinet-based pricing*, in units of 500kW to 2MW
- Contracts are held by IT and average 48-60 months with limited renewal clause price protection
- Average power density: 1000-1500 W/m²; 4-10 kW/cabinet
- [US] Market price point (per kW/month): \$650-\$800
- Market premium over 3 years has exhibited 10% pa increase over commodity co-location
- Limited availability; investment mostly in primary markets with little speculative development



Economic Considerations

Electronic Trading Co-location characteristics:

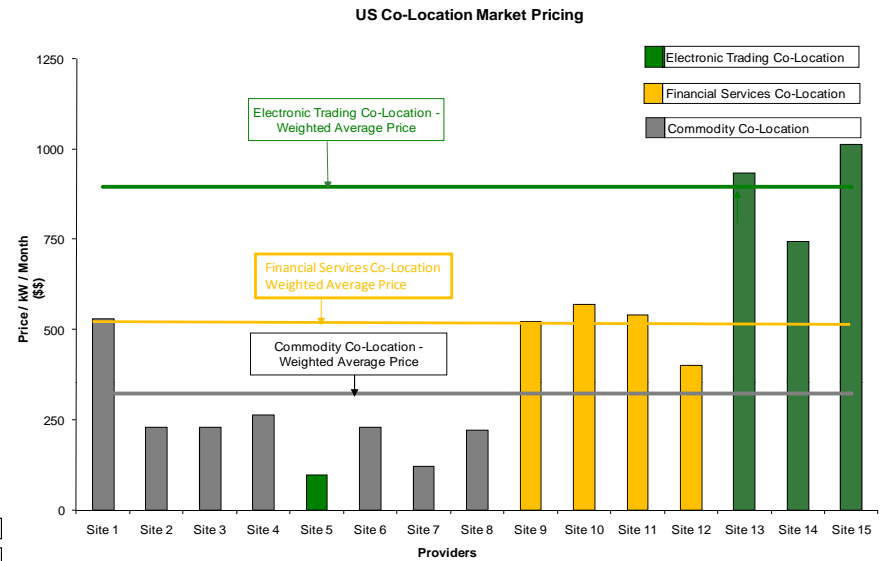
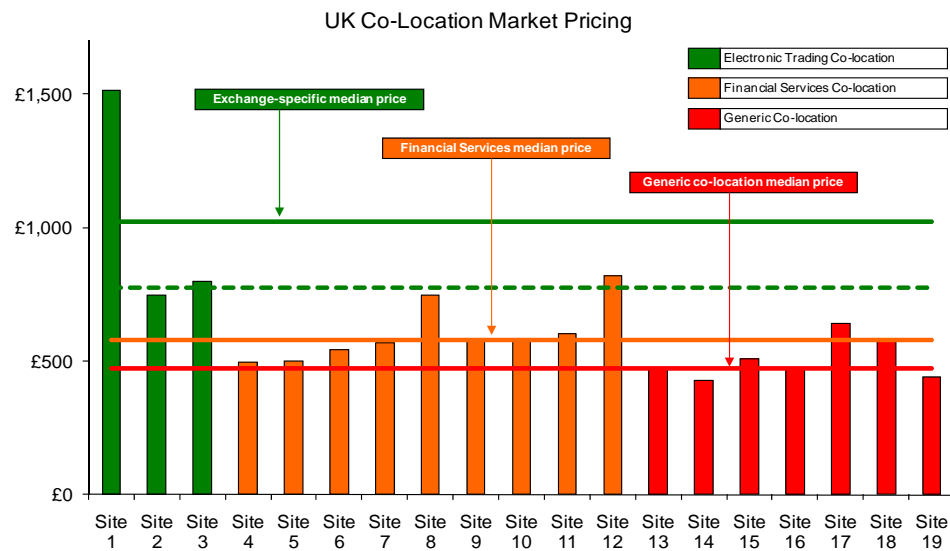
- Co-location defined by crossing services, *e.g.*, *exchanges, dark pools*
- Mostly constructed as Tier III or Tier IV facilities in primary trading markets globally
- Contracts are generally cabinet based in units of 5-10 cabinets;
- Contracts (held by business desk) average 3 months with limited price protection
 - Pricing and renewals heavily dependent on, or subsidised by, order flow to contributed controlling crossing service
- Average power density: 1000-1500 W/m²; 4-8 kW/cabinet
 - Emerging trend to increasing density; customer demand for 1500-2500 w/m² with 8-12 kW/cabinet densities requested
- [US] Market price point (per kW/month): \$950 - \$1250
- Market premium over 3 years has exhibited a significant rise from 125% (2004) to 250% (2009) with continued escalation even in the adverse economic climate of 2009



Economic Considerations

Co-location market pricing:

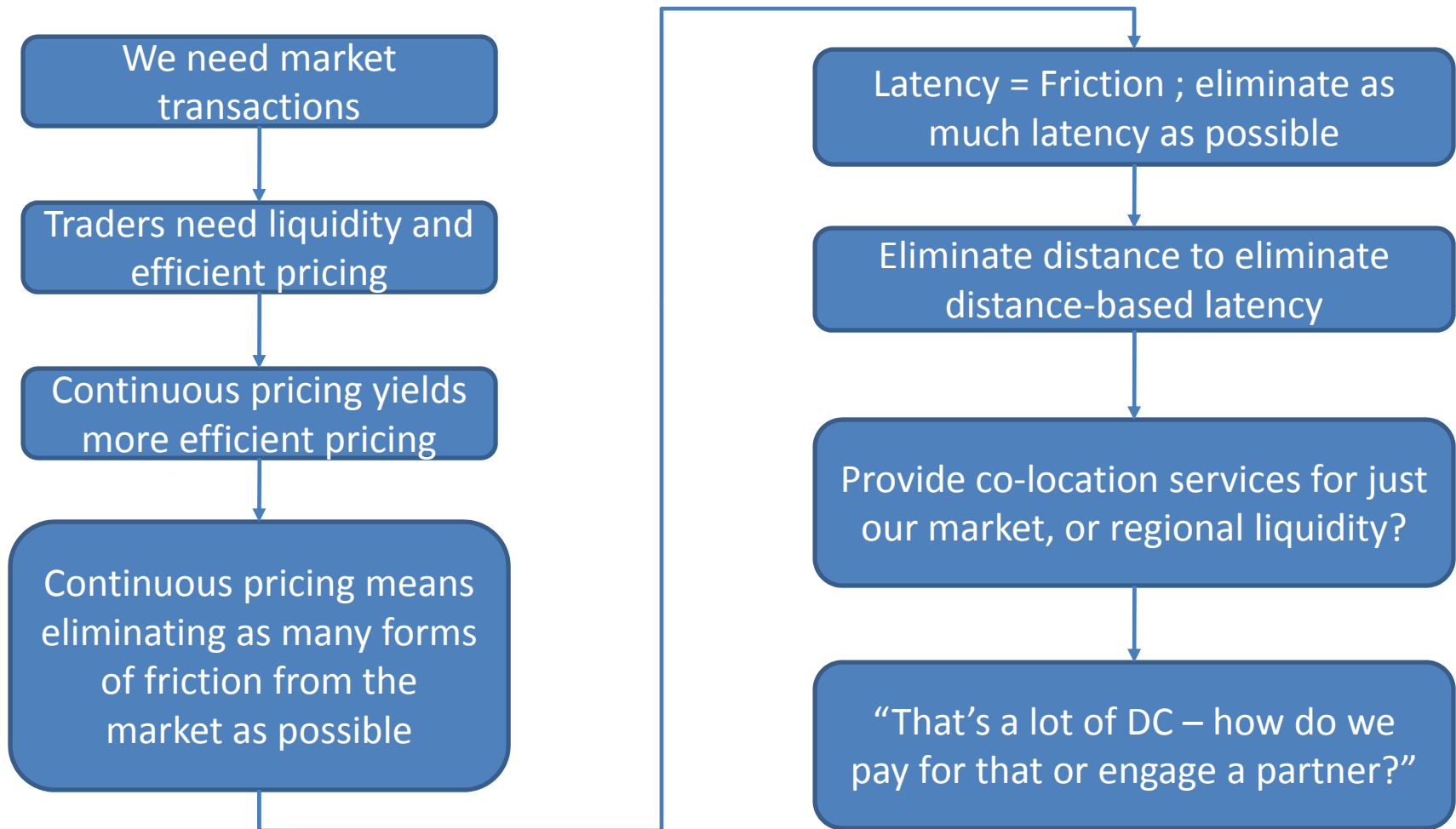
- Varies significantly by class of service
- Essential differentiation by class of service holds across markets



- Differentiation by presence of financial transaction services clearly creates pricing flexibility
- Business models offer opportunities to create offers for “carry-on” financial loads



Following the Logic Chain



Markets are Systems – with Systemic Behaviours

There is a flaw inherent within the view of current global of markets – that flaw is believing that value is continuous:

- Lottery tickets are an example of discontinuous value, with probabilities Instantaneously coalescing at 0% or 100%
- Market structures attempting to artificially make pricing behave as a continuous function create a highly optimised and efficient system
- This functions well when values remain continuous, but highly efficient systems of any kind rapidly trend to disequilibrium when subjected to extreme perturbation
- Discontinuities in valuation create exactly the type of extreme perturbation that disrupts highly efficient systems, creating rapid feedback cycles, system induced oscillation, and control (regulatory) challenges
- Measurement of inertia (volatility) becomes critical
- ***This will lead to market disruption and will impact data centre operators***



So What?

Exchanges will drive a significant amount of data centre activity:

- Technology acceleration reduces latency which allows for increased pricing accuracy, which in turn drives the increased opportunities for arbitrage which will define the business problems of exchange operators for the next decade
- Exchanges will require large scale capacity, which will become a driver for the data centre industry for the next several years
- Where geopolitical boundaries inhibited the aggregation of compute loads within ANZ to the multi-megawatt size of US and EMEA loads, exchange aggregation of HFT trading will drive large scale data centres
- Exchange operators will seek capital, construction, or operational partners to provide this capacity
- The uniqueness of the transactional nature of exchanges creates value – exchanges and their data centre sourcing partners must decide the equitable split of that value, should the exchange not hold it entirely internal

