CFRM 551: TRADING SYSTEMS

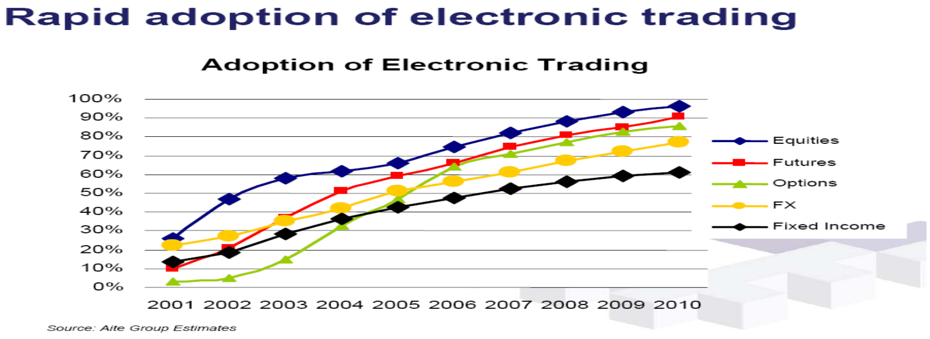
Slides to accompany High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems, 2nd edition, by Irene Aldridge

Emergence of algo trading

Evolution of trading

- Open outcry method of trading
- Virtual trading floor electronically
- Disintermediation
- Shorter settlement cycles
- Wall Street decimalization
- Standardization using FIX protocol
- Program trading - basket
- Digitalization of trading – Algorithmic trading

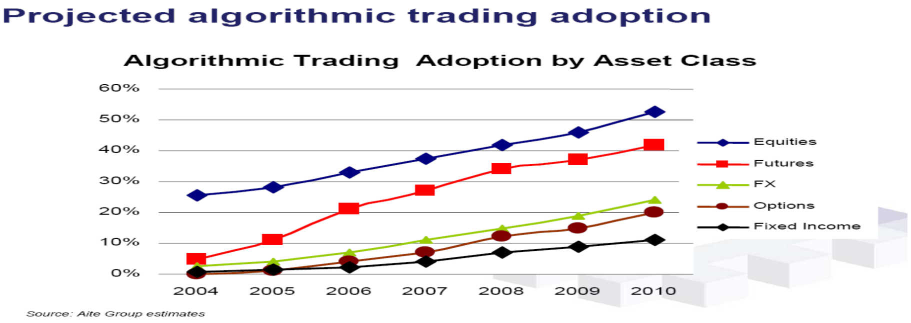
Adoption of electronic trading


Rapid adoption of electronic trading

Adoption of Electronic Trading

Source: Aite Group Estimates

Algorithmic trading adoption by asset class


Projected algorithmic trading adoption

Algorithmic Trading Adoption by Asset Class

Source: Aite Group estimates

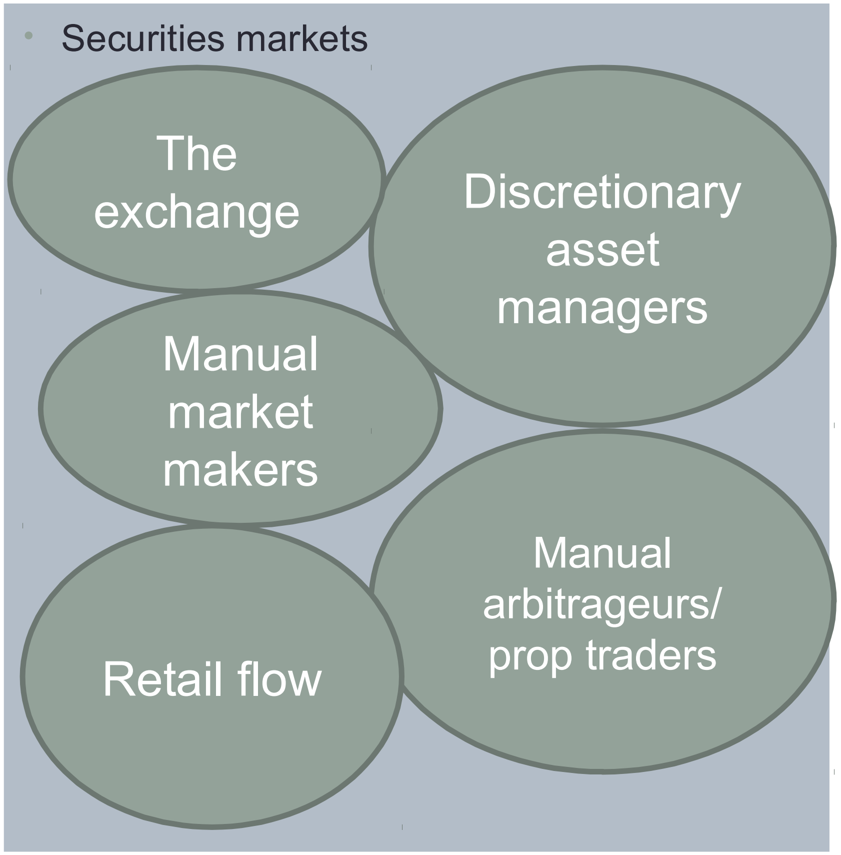
Algo trading is rapidly becoming the standard

Traditional markets

Traditional market participants

- Traditional players:
  - Discretionary asset managers
    - Pension funds, mutual funds and hedge funds
  - Manual market makers
    - Broker-dealers
  - Manual speculators
    - Broker-dealer proprietary trading
  - Retail flow
    - Mom-n-pop
  - Single non-profit exchange per asset class
- Key characteristics
  - High transaction costs
    - => low turnover of securities
  - Manual processing
    - => high degree of error
    - => high risk (traders judgment may fail)
  - High margin businesses

Before Electronization


Securities markets
The exchange
Discretionary asset managers
Manual market makers
Manual arbitrageurs/ prop traders
Retail flow

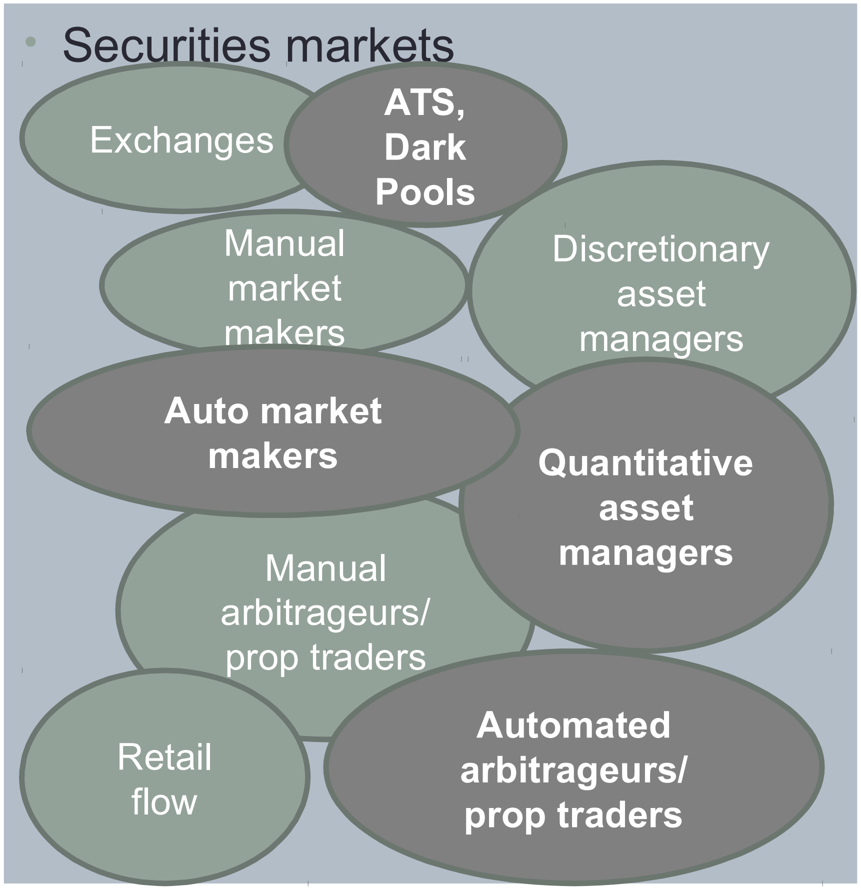
Steady-state labor-intensive processes generated high margins

Modern markets

Modern markets

- New entrants:
  - Quantitative money managers
    - Mutual funds and hedge funds
  - Automated market makers
    - Broker-dealers and hedge funds
  - Automated arbitrageurs
    - Stat arb hedge funds, prop traders
  - Multiple Exchanges
    - Alternative trading systems (ATS), dark pools
- Key characteristics
  - Democratic access to markets
  - Lower transaction costs
    - Examples: retail cost per trade in 1998: $70 with Merrill Lynch, retail cost per trade in $2008: $7.00 with Schwab, $0.70 with Interactive Brokers, a 100 times cost reduction over 10 years
    - Enables high turnover of securities
  - Automated trading, order routing and settlement
    - Low degree of error
  - Lower $$ margins for everyone

Now: new entrants


- Securities markets

Exchanges
ATS, Dark Pools
Manual market makers
Discretionary asset managers
Auto market makers
Quantitative asset managers
Manual arbitrageurs/ prop traders
Retail flow
Automated arbitrageurs/ prop traders

Automation opens up access, reduces margins for all

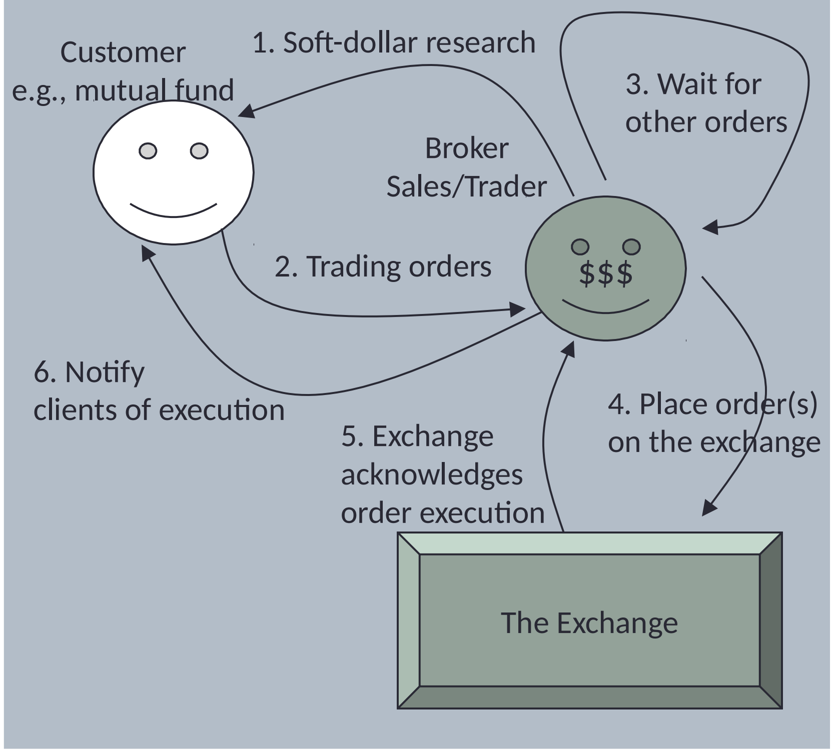
Traditional order routing

Traditional order sequence

1. Brokers deliver one-off trading ideas to customers (“soft-$$”)
   - Written or phoned-in research strategies
   - Based on analysis or market observation (e.g., what other customers are doing)
2. Customers decide to trade
   - Call the broker, place verbal order
3. The broker then waits until his orders fill up a “lot”
   - A lot is a round number of orders specified by the exchange
     - E.g. 10,000 shares
4. The broker routes the lot to the exchange
5. Exchange matches the order, acknowledges execution
6. Broker notifies client, makes $$

Before Electronization

- Securities markets



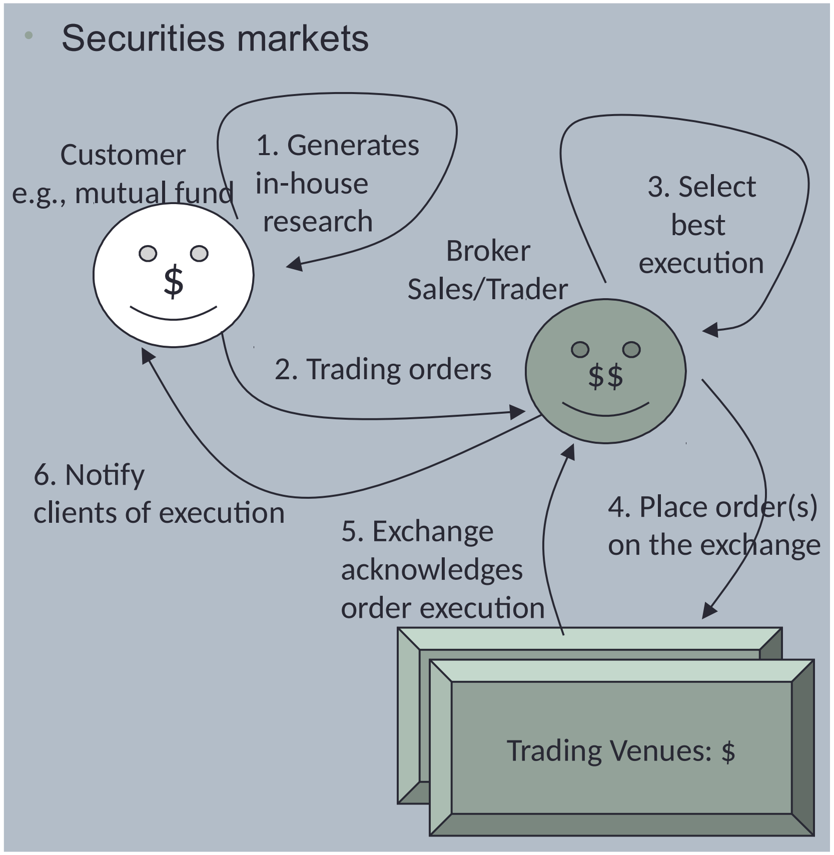
Traditional markets were broker-centric; time scale: few days

Modern indirect order routing

Modern order sequence: case 1

1. Customer generates research
   - Quant or HFT portfolio management
2. Customer decides to trade
   - Place an electronic order
3. The broker selects “best execution” algo for customer order
   - Depends on pre-determined customer needs
4. The algo electronically routes the order to exchange(s)
5. Trading venues match the order, acknowledge execution
6. Broker sends acknowledgement to customer, makes $$

Indirect order routing



- Securities markets

Customer
e.g., mutual fund

1. Generates
in-house
research

Broker
Sales/Trader

3. Select
best
execution

2. Trading orders

6. Notify
clients of execution

5. Exchange
acknowledges
order execution

4. Place order(s)
on the exchange

Trading Venues: $

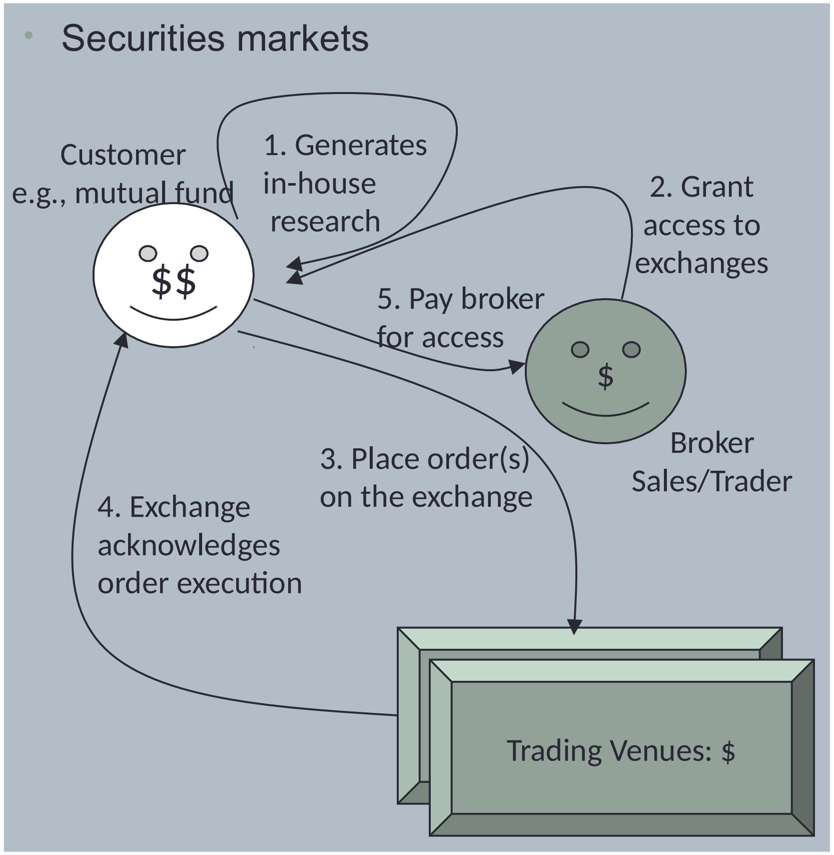
Customer-centric markets; time scale <= 1 day

Modern direct order routing

Modern order sequence: case 2

1. Customer systems generate research
   - Quant or HFT portfolio management
2. The broker grants customer direct access code to the exchange(s)
3. Customer systems decide to trade
   - Place electronic orders directly with the exchange, using customer's own order routing
4. Trading venues match the orders, acknowledge execution to client
5. Broker charges customer for use of the broker id

Direct order routing



Customer-centric markets, bypassing brokers; time scale <= 1 minute

Basic Terminology

- Market participant: anyone trading anything (broker-dealers, exchanges, HFT, low-frequency traders, retail mom-and-pop, …)
- Security = stock
- Financial instrument = stock, bond, futures contract, option, any other traded financial construct
- OTC = “over the counter,” not standardized contract
- Electronic = using computer systems, as opposed to paper and phone
- Algorithm = a process, a set of instructions for computer execution

Key outcomes

- Electronization => standardization => less OTC
  - More transparency
  - Less risk
- Electronization => higher execution speeds
  - Sub 20 millisecond execution is now standard
  - Manual brokers are unable to react to data fast enough, unable to place timely orders
  - Off-the-shelf computer technology is widely available to assist manual brokers

HFT

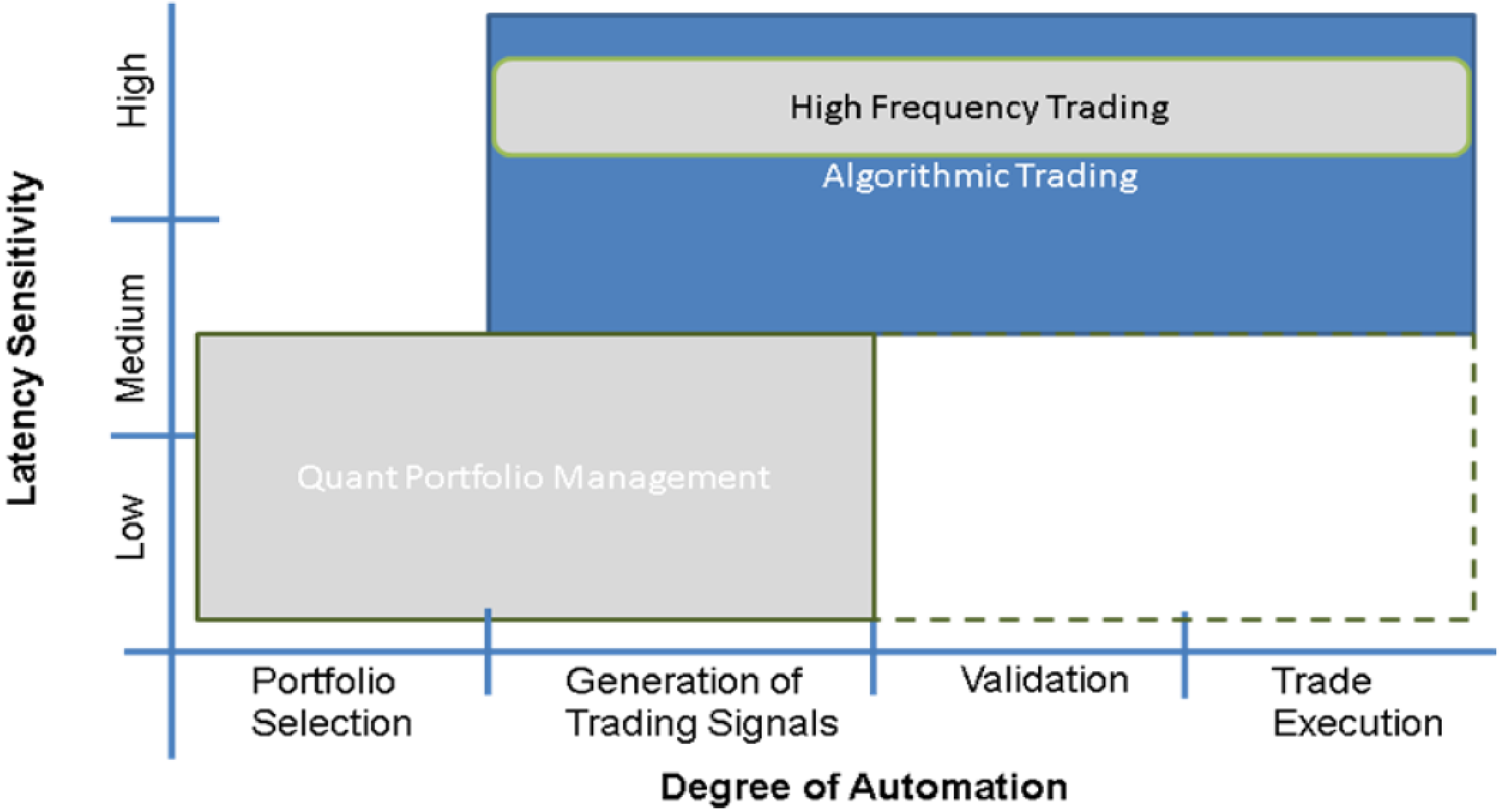
- A thousand definitions:
  - Everyone using electronic networks (CFTC working group)
  - Everyone using advanced electronic networks (academics)
  - Everyone holding intraday positions and quant reallocation (Hedge fund manager survey)
  - Specific market activity (CFTC/academics)
  - Activity impossible for humans (broker-dealers)
  - …

HFT Definition (CFTC subgroup)

- “High frequency trading is a form of automated trading that employs:
  - algorithms for decision making, order initiation, generation, routing, or execution, for each individual transaction without human direction;
  - low-latency technology that is designed to minimize response times, including proximity and co-location services;
  - high speed connections to markets for order entry; *and*
  - high message rates (orders, quotes or cancellations).”
- This definition applies to 95% of all market participants
  - HFT
  - Non-HFT
  - “Mom-and-pop” “point-and-click” traders
  - Anyone whose orders touch electronic networks

HFT Definition (academics)

- Gomber, Arndt, Lutat and Uhle (2011)
- HFT = quant trading on steroids



HFT definition (survey)

- High-frequency trading comprises of
  - Systematic,
  - Quant-based models
  - With holding periods from a fraction of a second to 1 day (no positions held overnight)

- Survey by FINalternatives (July 2009)
  - 300 responses from hedge fund managers who subscribe to FINalternatives (out of 10,000 questionnaires sent out)

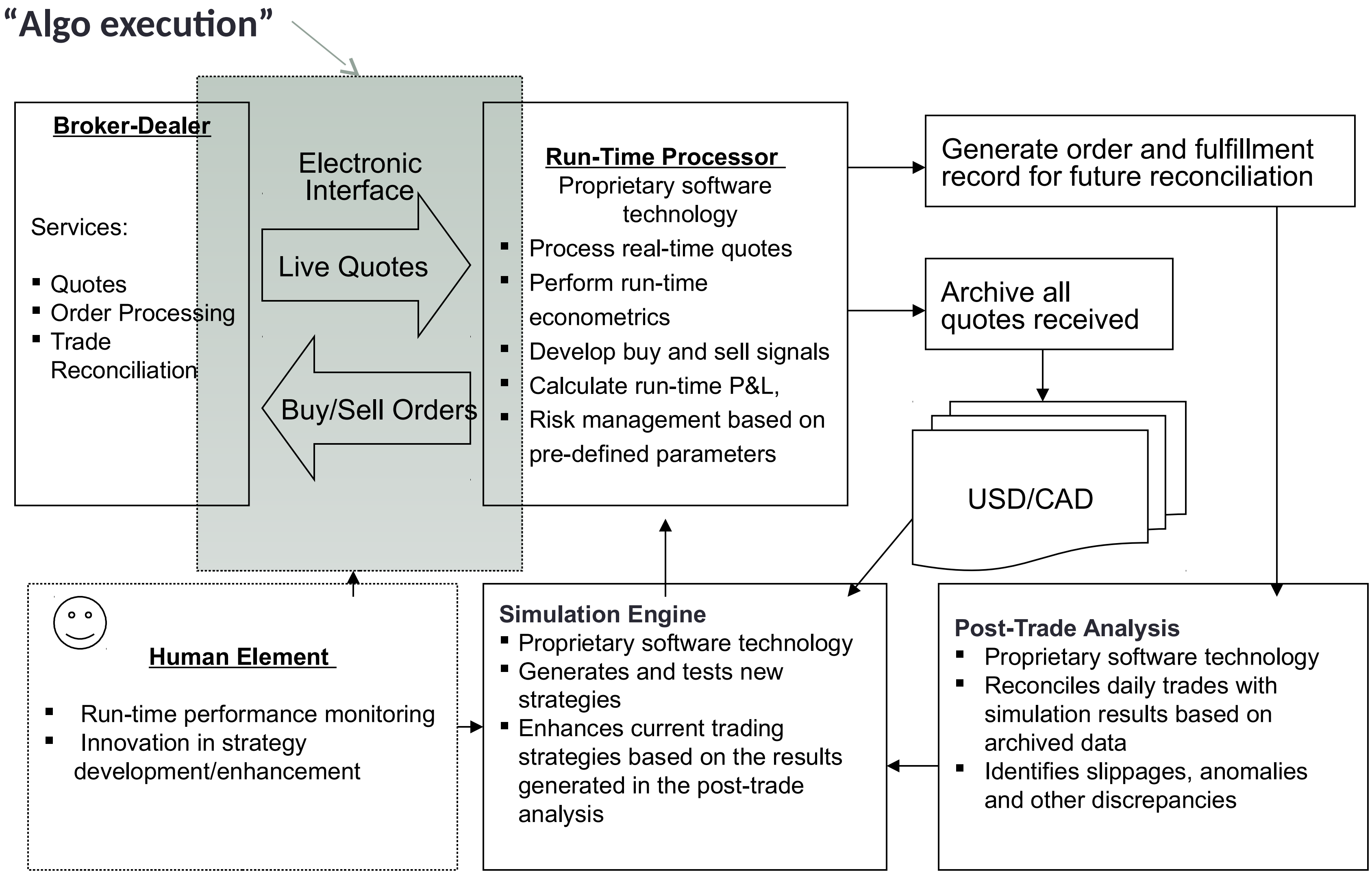
HFT Definition (academics 2)

- HFT = high market volume, low inventory.
- Kirilenko, Kyle, Samadi, Tuzun (2011)
- HFT is different from other market participants:
  - Intermediaries, characterized by low inventory, but not high trading volume
  - Fundamental buyers, who are consistent net buyers intraday
  - Fundamental sellers, who are consistent net sellers within a given day
  - Small traders, generating low volume
  - Opportunistic traders, loosely defined as all other traders, not fitting the definition of HFT or other categories above
- Somewhat arbitrary cut-off figures for low inventory and high volume

HFT Definition (broker-dealers)

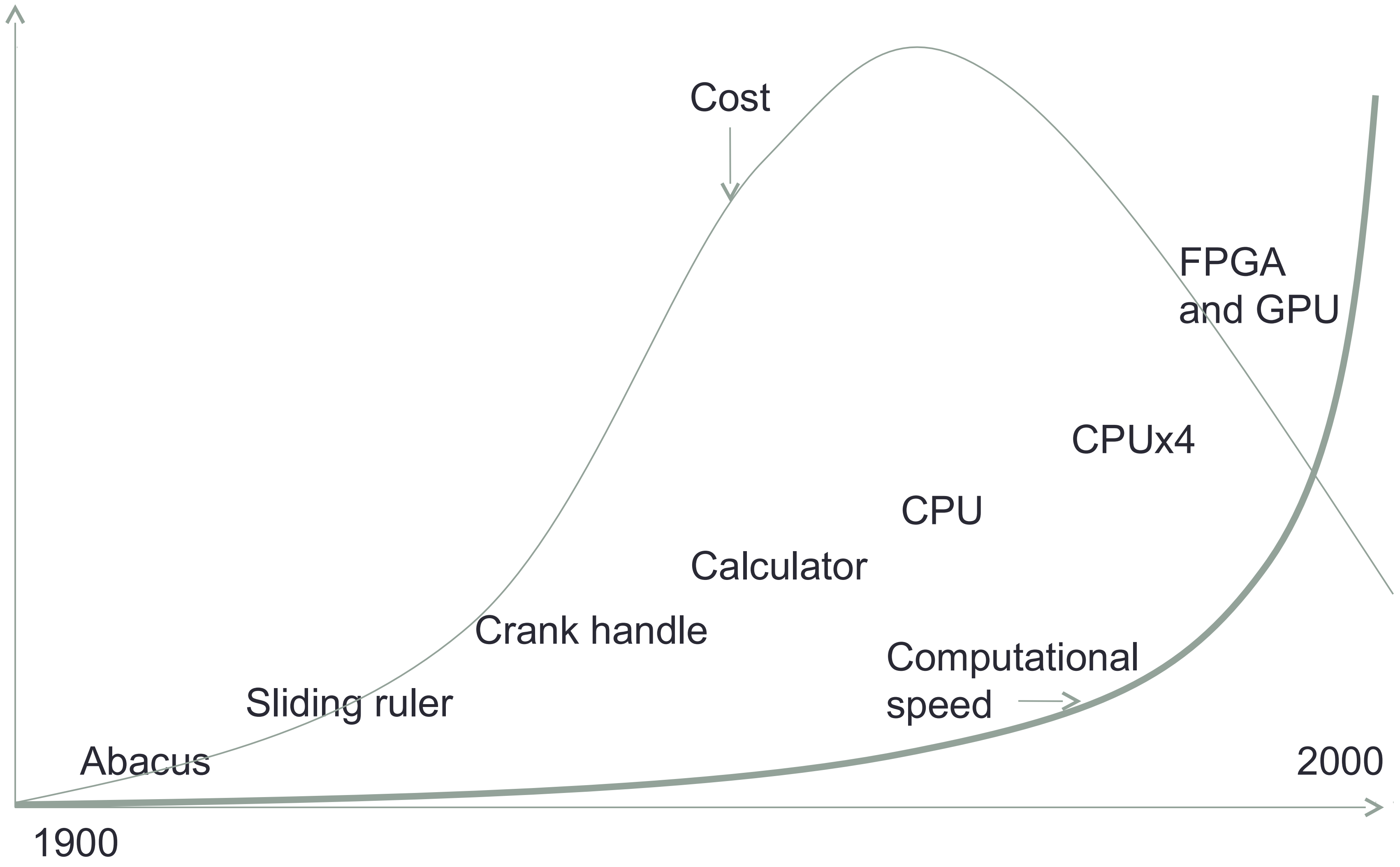
- HFT = trading behavior unattainable by human market participants
  - often used by brokers to segment their clients into HFT and non-HFT
  - attribution of trading activity of each specific account into human-feasible and human-infeasible
  - For example, an account generating 200 orders per second would be deemed HFT

Algo Trading and HFT



"Algo execution"
Broker-Dealer
Services:
▪ Quotes
▪ Order Processing
▪ Trade Reconciliation
Electronic Interface
Live Quotes
Buy/Sell Orders
Run-Time Processor
Proprietary software technology
▪ Process real-time quotes
▪ Perform run-time econometrics
▪ Develop buy and sell signals
▪ Calculate run-time P&L,
▪ Risk management based on pre-defined parameters
Generate order and fulfillment record for future reconciliation
Archive all quotes received
USD/CAD
Human Element
▪ Run-time performance monitoring
▪ Innovation in strategy development/enhancement
Simulation Engine
▪ Proprietary software technology
▪ Generates and tests new strategies
▪ Enhances current trading strategies based on the results generated in the post-trade analysis
Post-Trade Analysis
▪ Proprietary software technology
▪ Reconciles daily trades with simulation results based on archived data
▪ Identifies slippages, anomalies and other discrepancies

A brief history of hardware



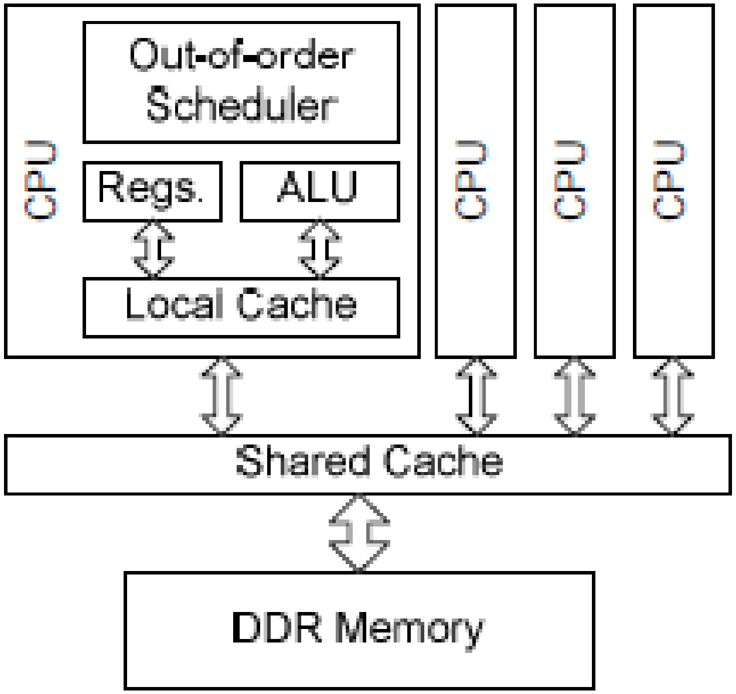
CPU

Key Characteristics

- “Central Processing Unit”
- “Brain” of a computer
- Decides how to store information in memory
- Scalable through multi-core approach
- CPUs use shared memory core-to-core cache coherency protocol for communication
- Arithmetic Logic Unit (ALU) manages current thread
- Other functions include management and scheduling tasks
- Core I7 8 cores
  - 169.79 GFLOPS (10^9)

Architecture

- Source: Thomas, Howes and Luk (2009)



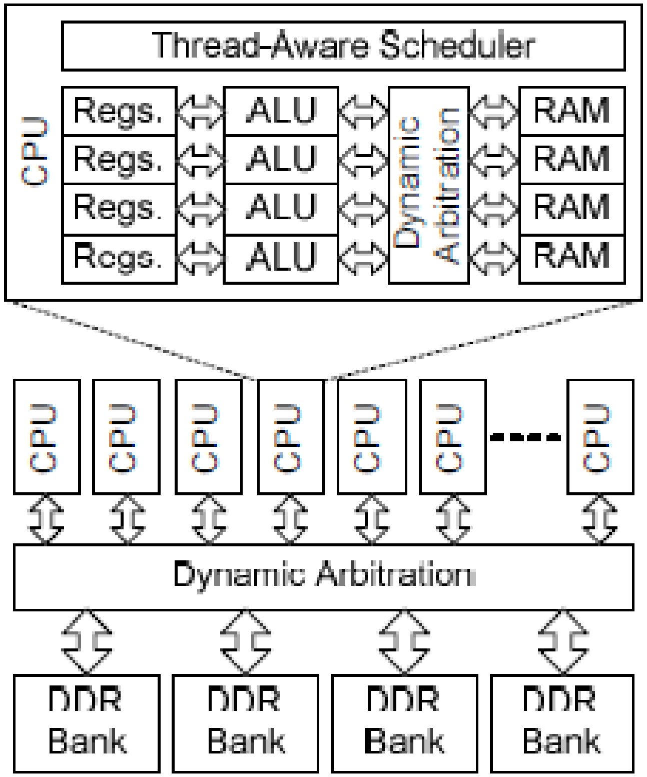
GPU

Key characteristics

- “Graphics Processing Unit”
- Most space on the chip is devoted to Arithmetic Logic Units (ALUs)
- Threads are executed in parallel batches of 32
  - Batches are called “warps”
  - To minimize latency, all threads in the same warp should be similar in terms of:
    - # of loops
    - conditional exits
- Nvidia Tesla K80, 2496 Cores
  - 2.91 TFLOPS (10^{12}) double precision performance
  - 8.74 TFLOPS single precision

GPU Architecture

- Source: Thomas, Howes and Luk (2009)



FPGA

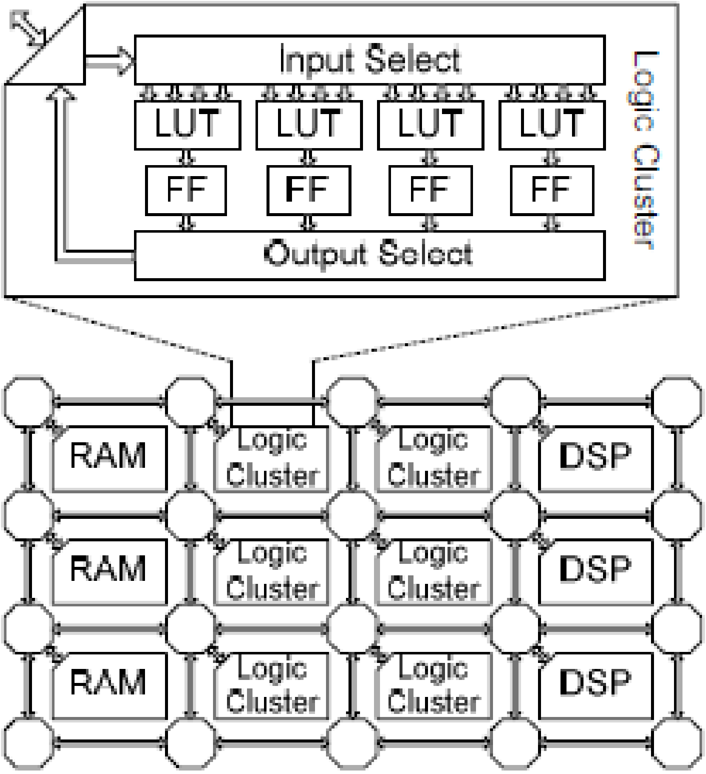
Key Characteristics

- Do not have any fixed instruction-set architecture
- Provide a fine-grain grid of bit-wise functional units
- Can be composed to create any desired circuit or processor
- Much of the FPGA area is dedicated to the routing infrastructure, which allows functional units to be connected together at run-time.
- Modern FPGAs also contain a number of dedicated functional units, such as DSP blocks containing multipliers, and RAM blocks.

- Popular model: Xilinx Virtex-5

Architecture

- Source: Thomas, Howes and Luk (2009)



FPGAs – “souped up” microprocessors

- Source: EETimes, March 3, 2004

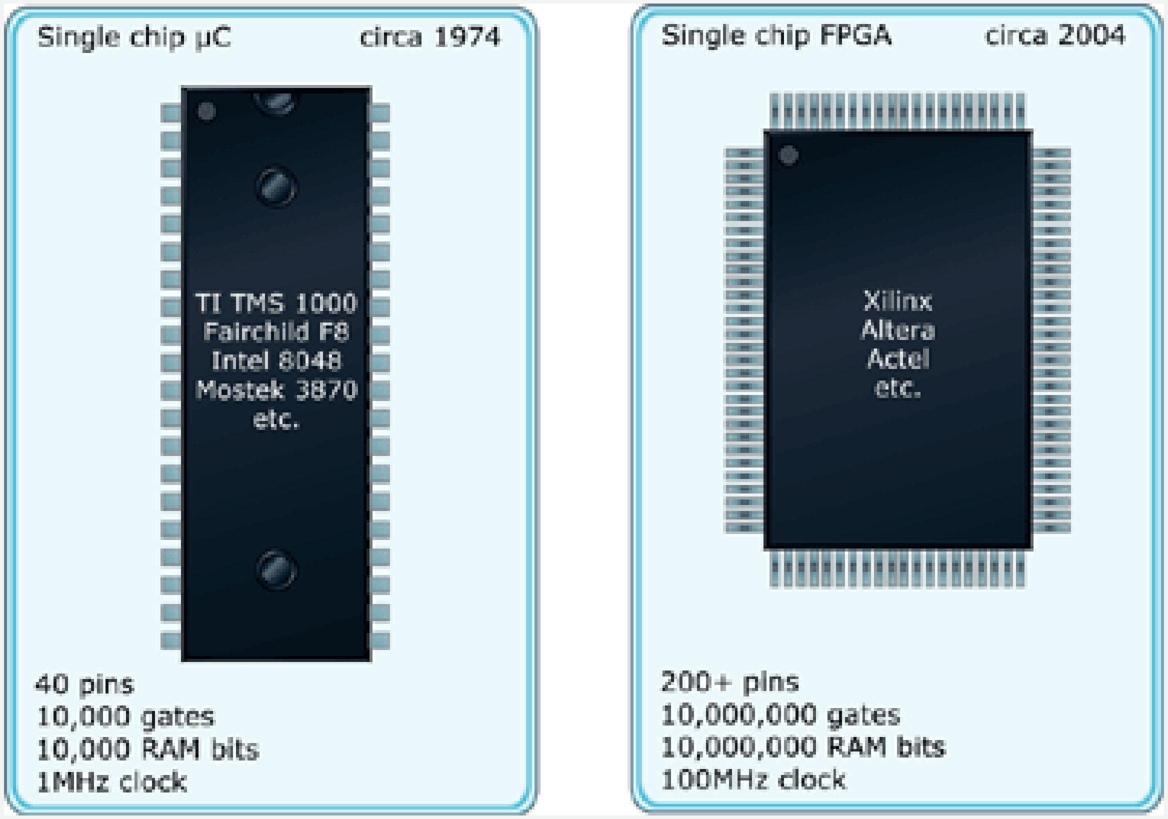
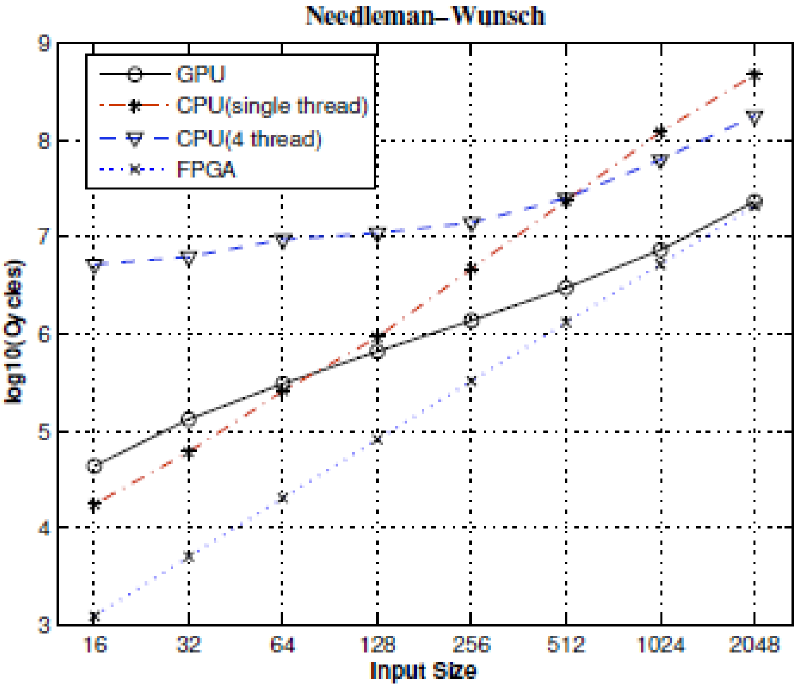


Figure 1: Comparison of first microprocessors to current FPGAs

Accelerator speed depends on clock

Comparative speeds in cycles



How fast is one cycle?

- Speed is determined by the "oscillator crystal"
  - shipped with CPU
  - Pentium 4 (2002) has clock rate of 3 billion cycles/second = 10^9 cycles/second
- Performance comparison:

Hardware	Data Size	Clock Cycles	Seconds
CPU	128	10^6	0.001
CPUx4	128	10^7	0.01
GPU	128	10^6	0.001
FPGA	128	10^5	0.0001

Hardware	Data Size	Clock Cycles	Seconds
CPU	2048	10^9	1
CPUx4	2048	10^8	0.1
GPU	2048	10^7	0.01
FPGA	2048	10^7	0.01

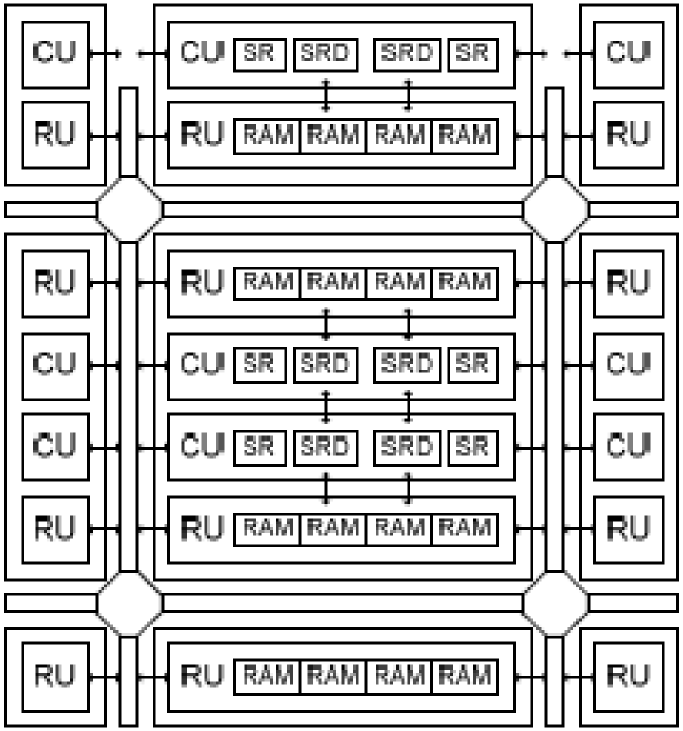
MPPA

Key characteristics

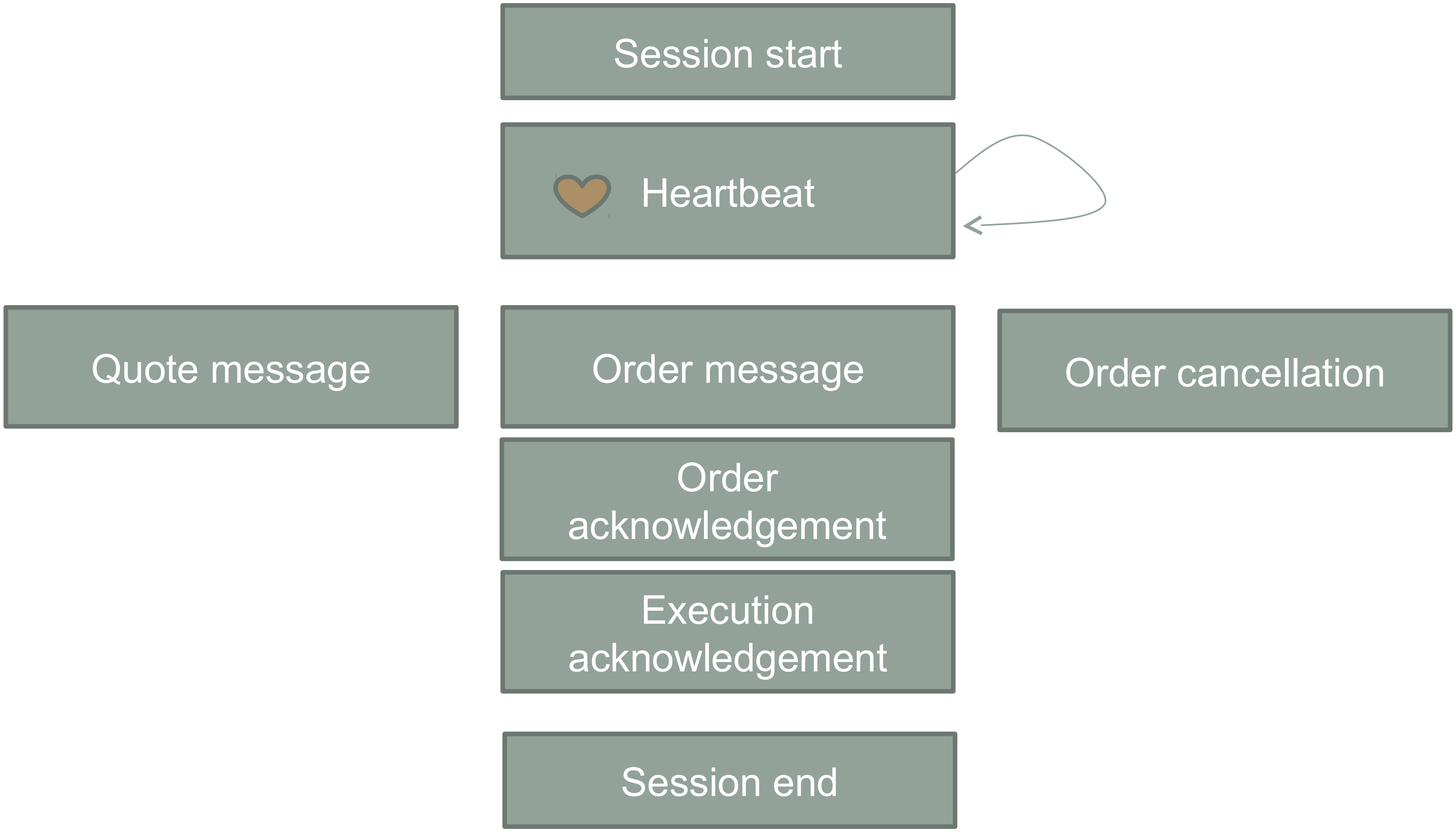
- Massively Parallel Processor Array
- Newest type of architecture:
  - The CPUs are instantiated in a regular grid
  - 2D communication channels between CPUs
  - Small local memories.
- Such small CPUs provide excellent efficiency in terms of peak performance per mm2 or power efficiency
- May present problems when partitioning applications
- Popular model: Ambric AM2000

Architecture

- Source: Thomas, Howes and Luk (2009)



Core message architecture



Session start

Heartbeat

Quote message

Order message

Order cancellation

Order acknowledgement

Execution acknowledgement

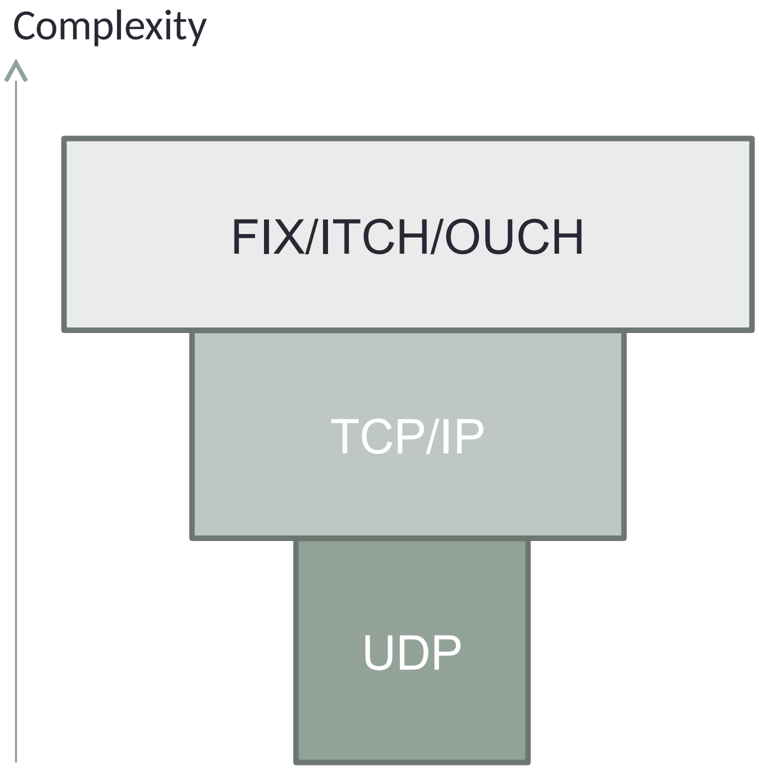
Session end

Messaging Protocols in Trading

Key Characteristics

- FIX/ITCH/OUCH are the “high-level” communication protocols
  - **F**inancial **I**nformation e**X**change protocol
    - ITCH and OUCH are NASDAQ’s alternatives
  - Requires messaging delivery protocol like TCP/IP
- TCP/IP is a standard Internet communication protocol
  - **T**ransmission **C**ontrol **P**rotocol/**I**nternet **P**rotocol
  - Used for e-mail, web browsing
  - All packets are numbered, the total number of bytes within the packet is counted, undelivered data is resent
- UDP is a fast low-level protocol
  - **U**ser **D**atagram **P**rotocol
  - Used for media streaming
  - Does not carry acknowledgements

Graphic Representation



Messaging protocols

Speed of protocols

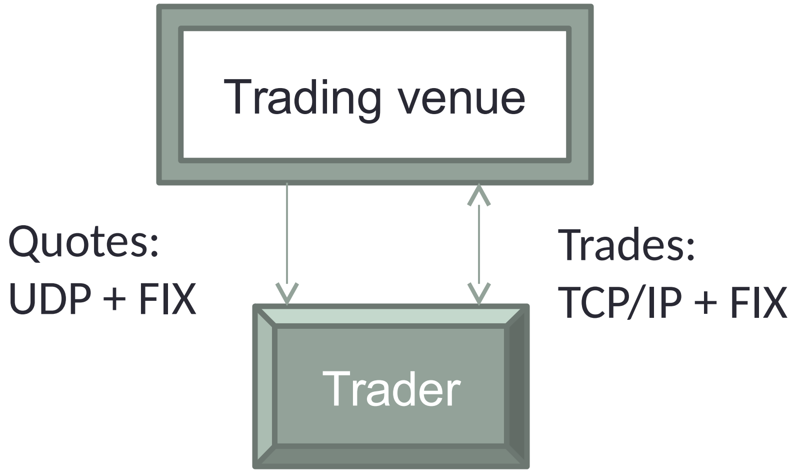
- Transmission speed depends on:
  - Size of message
  - Connection bandwidth
  - TCP/IP and UDP window sizes (how many bytes the clients are willing to receive at once)

Protocol	Header size
UDP	8 bytes
TCP/IP	20 bytes
FIX	100+ bytes

- As a rule:
  - TCP/IP is 3 times slower than UDP

Protocol security and usage

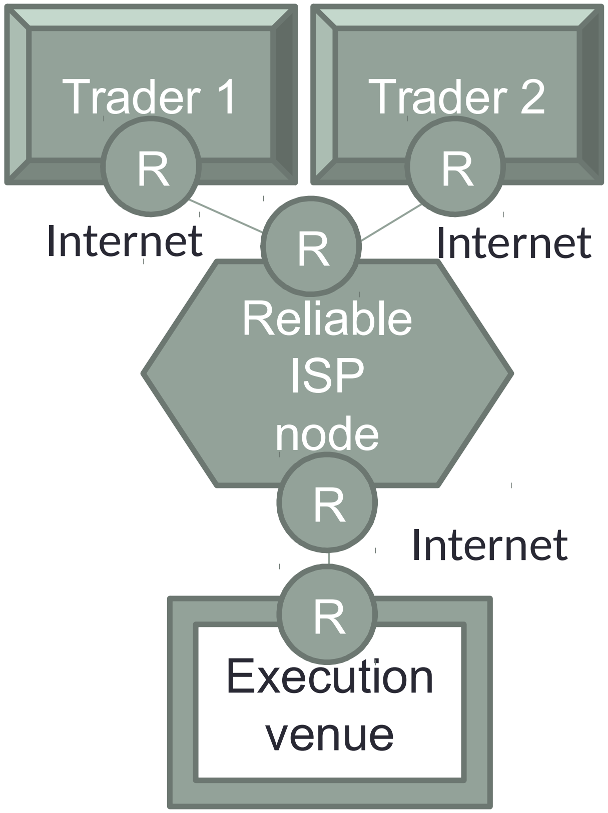
- UDP: None
- TCP/IP: some
  - Basic destination and packet count checks
- FIX: optional encryption



- Huge vulnerabilities persist

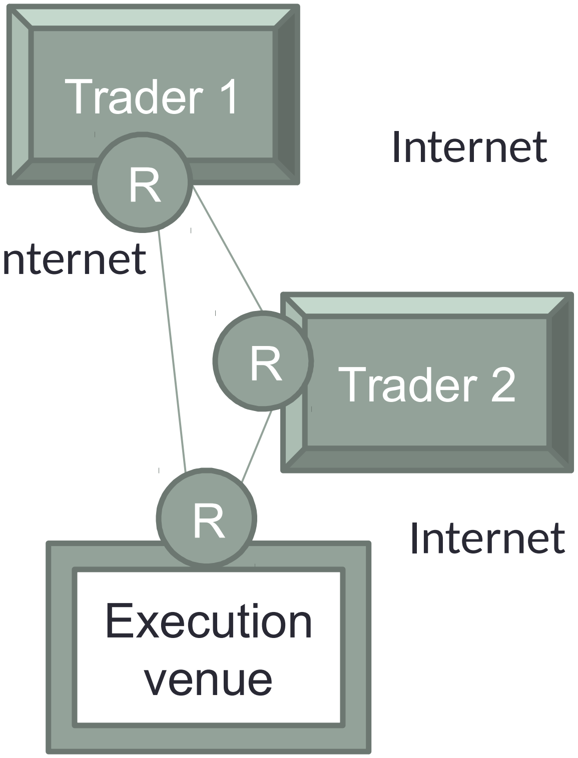
Messaging Architecture

Client-server model



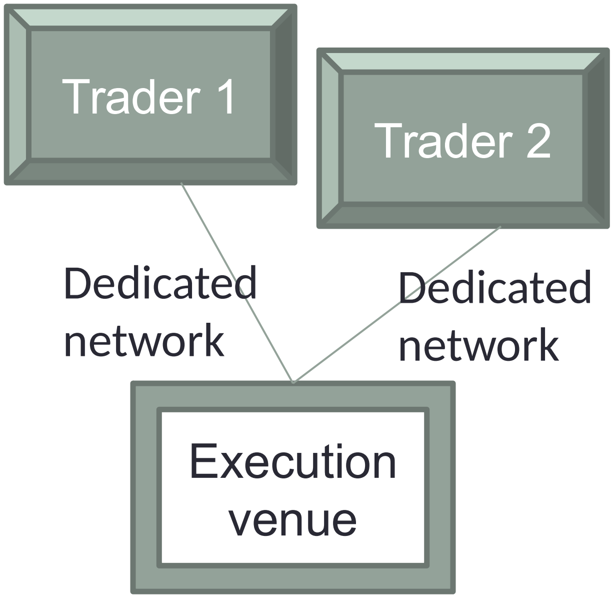
Moderately secure, but slow

Peer-to-peer model



Faster, but security can be
Compromised by Trader 2

Co-location model



Fast and secure

Co-location

Key Characteristics

- Definition:
  - Trading servers housed in the same facility as execution servers
  - Trading servers often have dedicated network access
- Alternatives
  - System warehousing in close proximity to the exchange
- Benchmark costs:
  - $3000/rack of 20 servers

Savings from Colocation

- Time saving
  - Round-trip latency between New York and Chicago: 17-22 milliseconds
  - Savings can be high in volatile market conditions
- Added security
  - Dedicated networks prevent packet snooping, message hacking
  - Unlimited savings from cyber attack prevention