UBS

British Computer Society
FiNSG and APSG

Public

Application of High Performance Computing in Investment Banks

Dr. Tony K. Chau
Lead Architect, IB CTO, UBS

January 8, 2014

Table of contents

Cautionary statement regarding forward-looking statements

This presentation contains statements that constitute "forward-looking statements", including but not limited to management's outlook for UBS's financial performance and statements relating to the anticipated effect of transactions and strategic initiatives on UBS's business and future development. While these forward-looking statements represent UBS's judgments and expectations concerning the matters described, a number of risks, uncertainties and other important factors could cause actual developments and results to differ materially from UBS's expectations. These factors include, but are not limited to: (i) the degree to which UBS is successful in executing its announced strategic plans, including its efficiency initiatives and its planned further reduction in its Basel III risk-weighted assets (RWA) and leverage ratio denominator (LRD); (ii) developments in the markets in which UBS operates or to which it is exposed, including movements in securities prices or liquidity, credit spreads, currency exchange rates and interest rates and the effect of economic conditions and market developments on the financial position or creditworthiness of UBS's clients and counterparties; (iii) changes in the availability of capital and funding, including any changes in UBS's credit spreads and ratings, or arising from requirements for bail-in debt or loss-absorbing capital; (iv) changes in or the implementation of financial legislation and regulation in Switzerland, the US, the UK and other financial centers that may impose more stringent capital (including leverage ratio), liquidity and funding requirements, incremental tax requirements, additional levies, limitations on permitted activities, constraints on remuneration or other measures; (v) uncertainty as to when and to what degree the Swiss Financial Market Supervisory Authority (FINMA) will approve reductions to the incremental RWA resulting from the supplemental operational risk-capital analysis mutually agreed to by UBS and FINMA, or will approve a limited reduction of capital requirements due to measures to reduce resolvability risk; (vi) the degree to which UBS is successful in executing the announced creation of a new Swiss banking subsidiary, a holding company for the UBS Group (including the announced offer to exchange shares of UBS AG for shares of such holding company), a US intermediate holding company, changes in the operating model of UBS Limited and other changes which UBS may make in its legal entity structure and operating model, including the possible consequences of such changes, and the potential need to make other changes to the legal structure or booking model of UBS Group in response to legal and regulatory requirements, including capital requirements, resolvability requirements and the pending Swiss parliamentary proposals and proposals in other countries for mandatory structural reform of banks; (vii) changes in UBS's competitive position, including whether differences in regulatory capital and other requirements among the major financial centers will adversely affect UBS's ability to compete in certain lines of business; (viii) the liability to which UBS may be exposed, or possible constraints or sanctions that regulatory authorities might impose on UBS, due to litigation, contractual claims and regulatory investigations; (ix) the effects on UBS's cross-border banking business of tax or regulatory developments and of possible changes in UBS's policies and practices relating to this business; (x) UBS's ability to retain and attract the employees necessary to generate revenues and to manage, support and control its businesses, which may be affected by competitive factors including differences in compensation practices; (xi) changes in accounting or tax standards or policies, and determinations or interpretations affecting the recognition of gain or loss, the valuation of goodwill, the recognition of deferred tax assets and other matters; (xii) limitations on the effectiveness of UBS's internal processes for risk management, risk control, measurement and modeling, and of financial models generally; (xiii) whether UBS will be successful in keeping pace with competitors in updating its technology, particularly in trading businesses; (xiv) the occurrence of operational failures, such as fraud, unauthorized trading and systems failures; and (xv) the effect that these or other factors or unanticipated events may have on our reputation and the additional consequences that this may have on our business and performance. The sequence in which the factors above are presented is not indicative of their likelihood of occurrence or the potential magnitude of their consequences. Our business and financial performance could be affected by other factors identified in our past and future filings and reports, including those filed with the SEC. More detailed information about those factors is set forth in documents furnished by UBS and filings made by UBS with the SEC, including UBS's Annual Report on Form 20-F for the year ended 31 December 2013. UBS is not under any obligation to (and expressly disclaims any obligation to) update or alter its forward-looking statements, whether as a result of new information, future events, or otherwise.

Important information related to numbers shown in this presentation

Use of adjusted numbers
Unless otherwise indicated, “adjusted” figures exclude the adjustment items listed on the previous slide, to the extent applicable, on a Group and business division level. Adjusted results are a non-GAAP financial measure as defined by SEC regulations. Refer to pages 13-14 of the 2Q14 financial report for an overview of adjusted numbers.

Basel III RWA, Basel III capital and Basel III liquidity ratios
Basel III numbers are based on the BIS Basel III framework, as applicable for Swiss Systemically relevant banks (SRB). In the presentation are SRB Basel III numbers unless otherwise stated. Our fully applied and phase-in Swiss SRB Basel III and BIS Basel III capital components have the same basis of calculation, except for differences disclosed on page 85 of the 2Q14 financial report.

Basel III risk-weighted assets in the presentation are calculated on the basis of Basel III fully applied unless otherwise stated. Our RWA under BIS Basel III are the same as under Swiss SRB Basel III.

Leverage ratio and leverage ratio denominator in this presentation are calculated on the basis of fully applied Swiss SRB Basel III, unless otherwise stated.

From 1Q13 Basel III requirements apply. All Basel III numbers prior to 1Q13 are on a pro-forma basis. Some of the models applied when calculating pro-forma information required regulatory approval and included estimates (discussed with our primary regulator) of the effect of these new capital charges.

Refer to the “Capital Management” section in the 2Q14 financial report for more information.

Currency translation
Monthly income statement items of foreign operations with a functional currency other than Swiss francs are translated with month-end rates into Swiss francs. Refer to “Note 17 Currency translation rates” in the 2Q14 financial report for more information.

Rounding
Numbers presented throughout this presentation may not add up precisely to the totals provided in the tables and text. Percentages, percent changes and absolute variances are calculated based on rounded figures displayed in the tables and text and may not precisely reflect the percentages, percent changes and absolute variances that would be derived based on figures that are not rounded

Net profit attributable to preferred noteholders:
Purchase of UBS AG shares by UBS Group AG pursuant to the exchange offer to create a group holding company is expected to cause a triggering event which results in accruals for future distributions to preferred noteholders. Assuming the acceptance date for the exchange offer is in the 4Q14, we expect to attribute further net profit to preferred noteholders of up to approximately CHF 80 million in that period. If we have attributed net profit to preferred noteholders of CHF 80 million in 4Q14, we would expect to attribute net profit to preferred noteholders of approximately CHF 30 million in 2015 and CHF 80 million in 2016.

Section 1

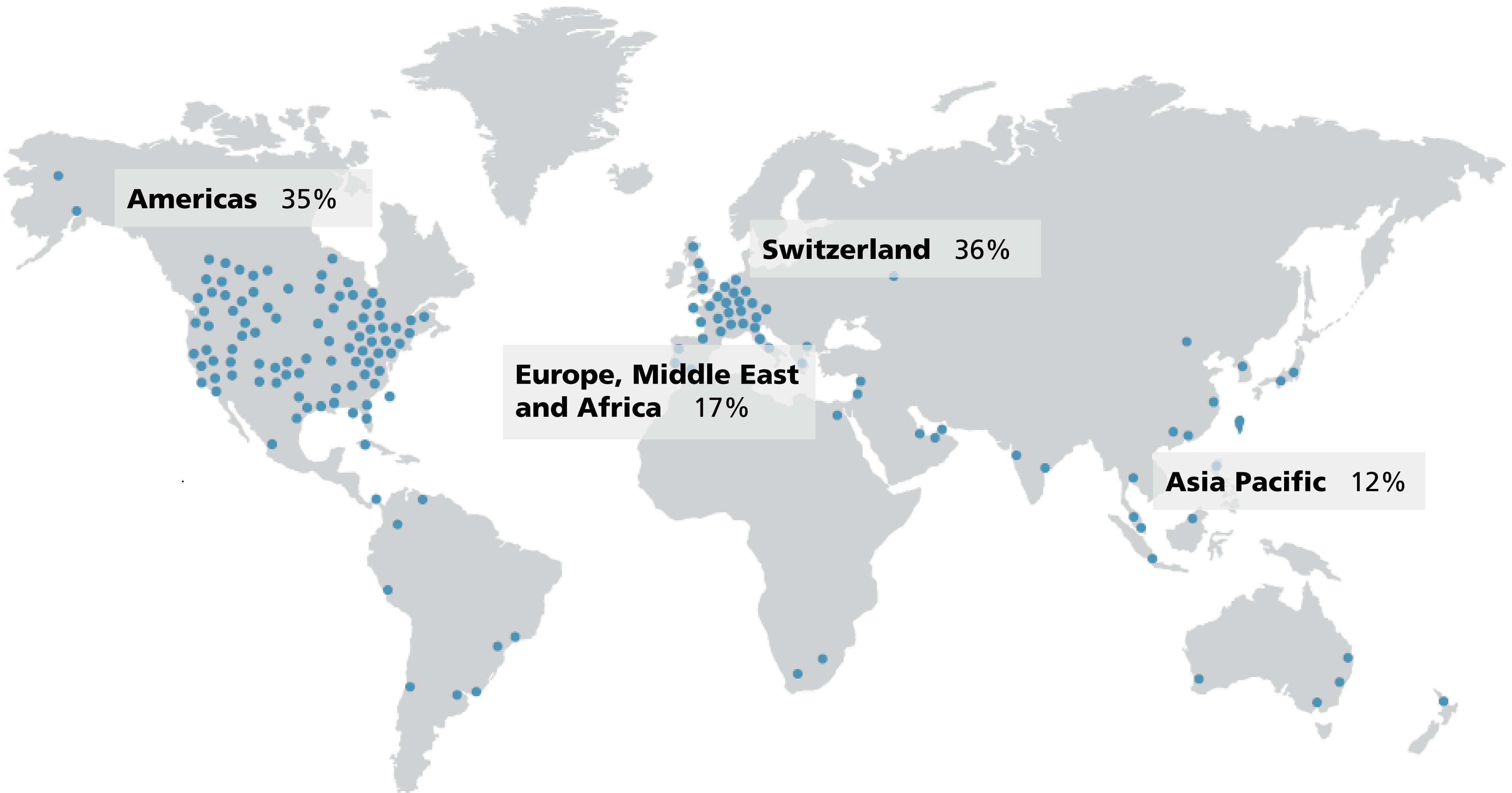
UBS General Information

UBS – one of the leading financial firms

- We draw on our 150-year heritage to serve private, institutional and corporate clients worldwide, as well as retail clients in Switzerland. Our business strategy is centered on our pre-eminent global wealth management businesses and our leading universal bank in Switzerland, complemented by our Global Asset Management business and our Investment Bank, with a focus on capital efficiency and businesses that offer a superior structural growth and profitability outlook.
- UBS is present in all major financial centers worldwide. It has offices in more than 50 countries, with about 35% of its employees working in the Americas, 36% in Switzerland, 17% in the rest of Europe, the Middle East and Africa and 12% in Asia Pacific. UBS employs about 60,000 people around the world.
- Its shares are listed on the SIX Swiss Exchange and the New York Stock Exchange (NYSE).

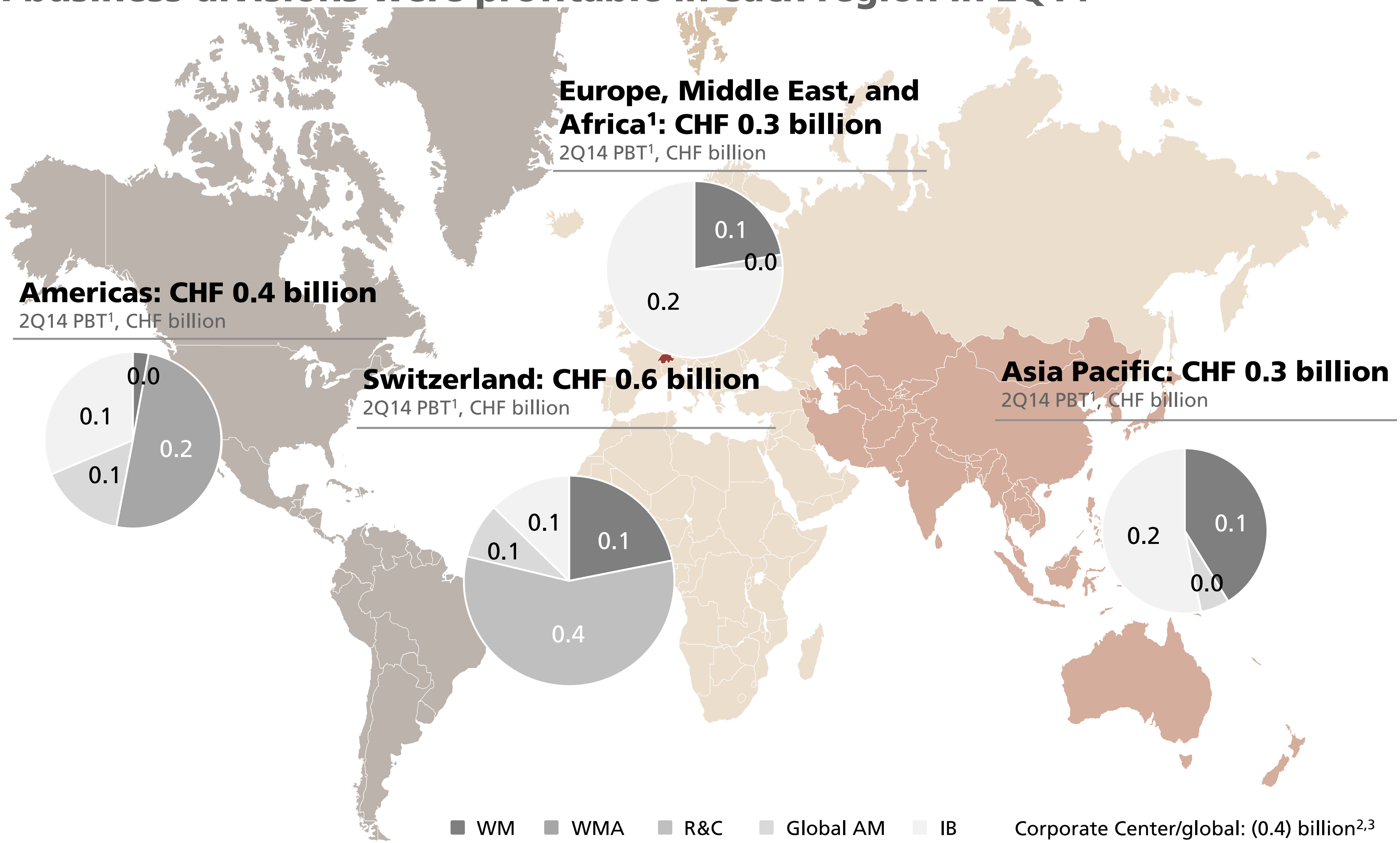
Worldwide presence

About 60,000 people in more than 50 countries



Broadly diversified profit sources

All business divisions were profitable in each region in 2Q14



1 Europe, Middle East, and Africa excl. Switzerland; 2 Numbers are not comparable to the disclosed financial statements of our main local subsidiaries; revenues are allocated in general following a client domicile view, which is supplemented by overlays to capture cross-country sales; this represents a more complete view of global and local sales for management purposes, as opposed to the split according to the legal entity where the transaction is recorded; 3 Includes Corporate Center and global operating income, expenses, and profit before tax that are not attributed to regions and are managed using a global view (CHF 36 million)

UBS

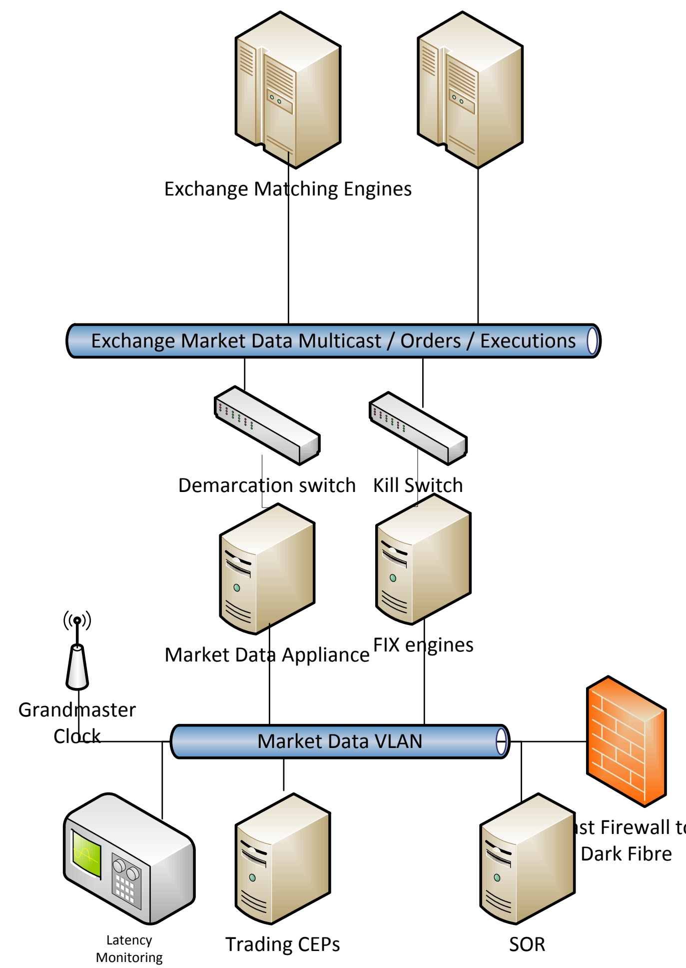
Section 2

Investment Bank Trading Infrastructure

Pre-Trade

Investment Banks provide a number of services to support different types of client requirements. They range from voice, Web based, through a consolidator or direct execution clients.

- Fault Tolerant Redundant Infrastructure
  - Fire Wall / Load Balancer (ADC)
  - Network Switches (for low latency, top of the rack)
  - Monitoring
  - Network Interface Card
  - High performance Local / Wide Area Network
  - Kill Switches
- Applications
  - FIX engine (Financial Information eXchange is a open protocol for order/execution)
  - Credit and Suitability Check
  - Global Smart Order Routing
  - Algorithmic execution

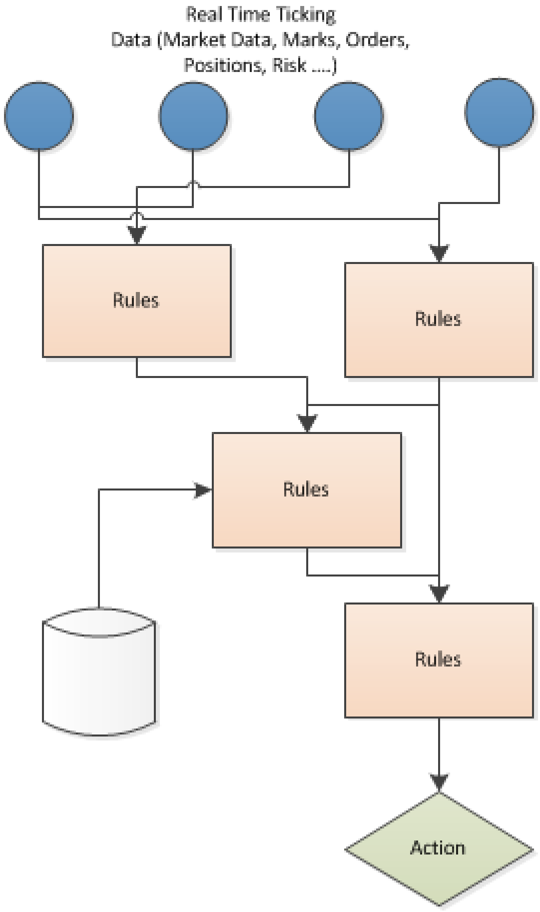


Market Data

Each Exchange or Electronic Crossing Networks provides real time market data for quotes / orders / execution. The market data is push out in across number of multicast addresses on 10G Ethernet. The update rate can peak at 16 million / second (e.g. US Option market – OPRA feed*)

- Market Data Feed Handlers
  - Software base feed handlers with kernel bypass
    - In process
    - Server based
  - Field Programmable Gate Array (FPGA) Appliance
  - Custom FPGA network interface card
- In-house market data distribution
  - Custom switches and putting compute into the data plane
  - Efficient local market data distribution (within a Data Centre)
  - Wide area market data distribution (back to the trading floor)

Reactive DAG



UBS

Performance enhancement techniques

In order to be efficient, one has to be very close to the hardware and have very good instrumentation. Techniques to optimize the memory access path with very little abstractions are common practise.

- R&D into network interface and ensure Non-uniform memory access (NUMA) compliant (e.g. use CPU cores associated with the NIC for processing)
- How the trading stack co-exist with the Operating System (kernel bypass – e.g. OpenOnload™)
- In process computation – avoid context switches at all cost + very careful thread management
- If one is using a language that requires Garbage Collection (GC), ensure that GC doesn't fire in critical path
- Efficient Ring Buffers that leverage the on-chip L1/L2 cache
- Very careful about memory access (and the management of the TLB – Translation Lookaside Buffer)
- Use of on chip share memory to communicate between different cores that analyse the data in parallel

Custom Hardware

In order to be more efficient, one can offload lower level functions into hardware (ASICs) as well as using configurable hardware (like Field Programmable Gate Array – FPGA) for doing some functions.

- TCP / IP handling in hardware
- PTP in hardware
- Line A / B arbitrage in FPGA
- Line handler (Binary protocol normalisation /decoding)
- Order Book building in hardware
- Higher level function (e.g. conflation)….
- Regulatory "kill" switches
  - Real time price limit checks
  - Real time market impact checks
  - Real time "Kill and disconnect"

Post Trade

Big Data Streaming Analytics support for various activities (regulatory reporting, fraud detection, alerts and perishable insights).

- Time Series database
  - Collection of tick data for regression testing
  - Forensic evident
  - Implementation shortfall calculation (Transaction Cost Analysis)
- Risk management
  - Portfolio risk (vs greeks)
  - Liquidity risk
  - Tracking risk
  - Volatility
- For instruments that have optionality
  - Spread management
  - Automatic delta hedges

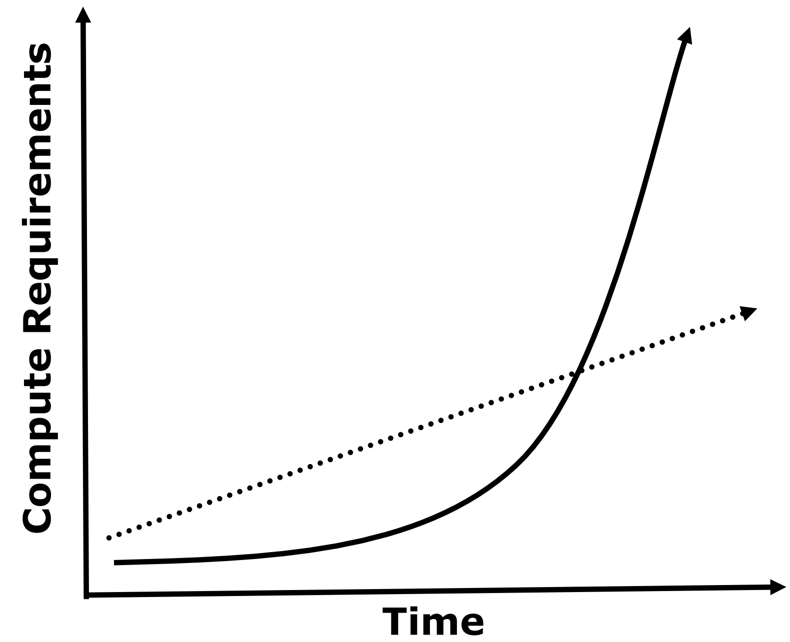
Section 3

High Performance Computing

Financial Services Compute Requirements

The growth in compute requirements for Financial Services firms has been phenomenal in recent years

- There are three primary factors that have contributed to this:
  - Calculation Method
  - Volumes
  - Complexity
- There are thousands of products that make up the Financial Services spectrum, and they vary widely in terms of compute requirements

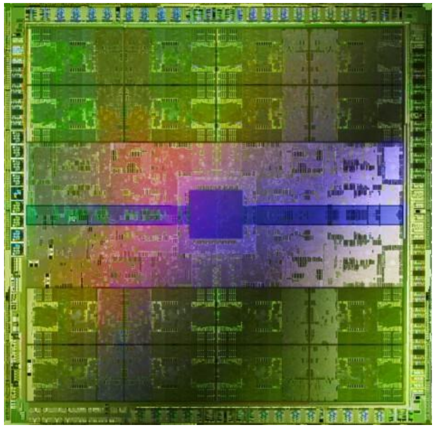


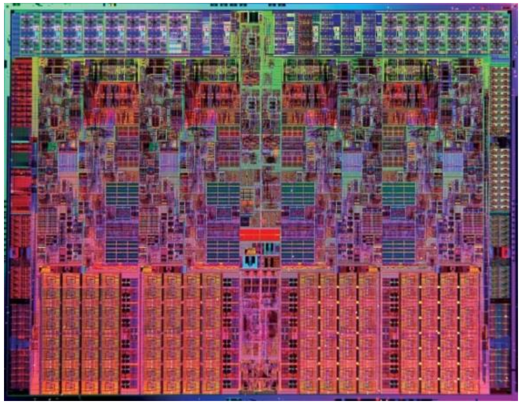
Vanilla ⟷ **Exotic**

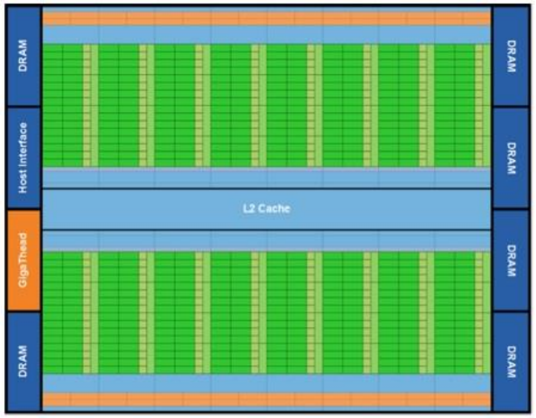
Low Complexity	High Complexity
High Volume	Low Volume
Low Latency	High Latency

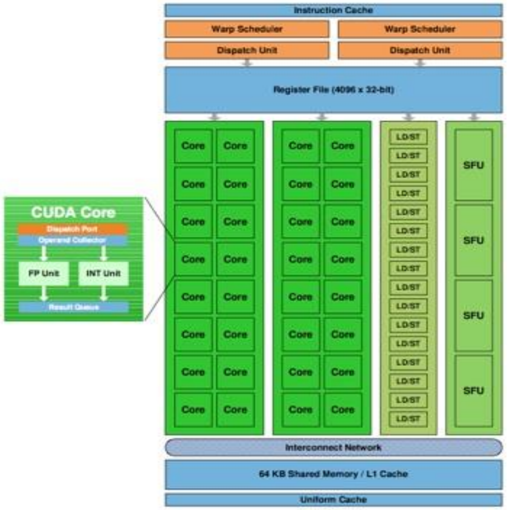
Way to solve this…

CPU clock speed is not getting any higher so the only way to go faster is to do things in parallel. This basically involves using more compute cores (either in the vector processor on the CPU – Intel® AVX2™ instruction) or acceleration device (like GPU) or custom hardware in FPGA



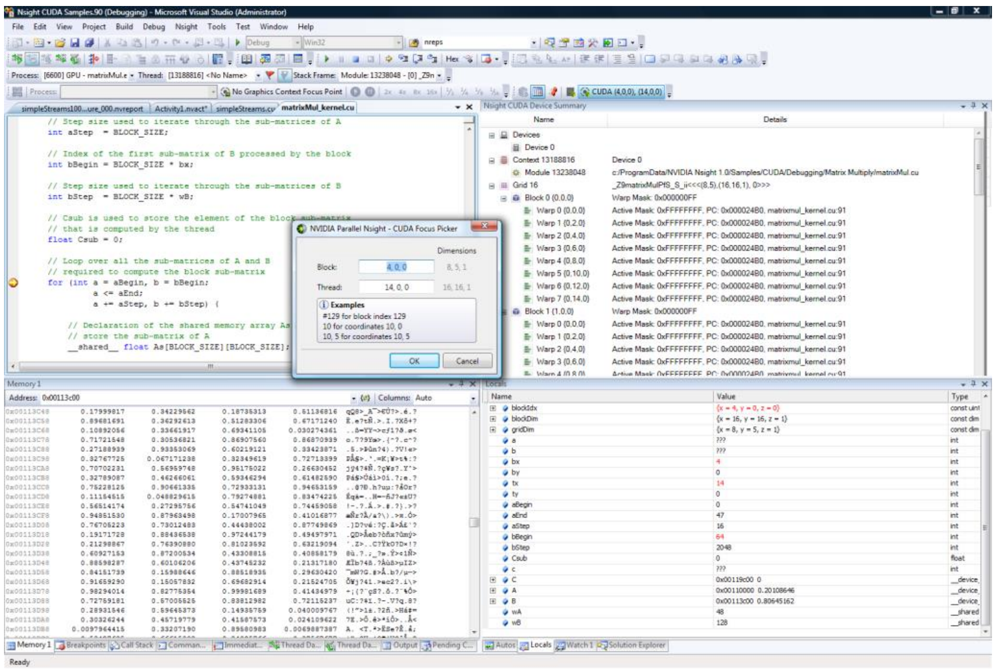






GPGPU Programming Model

Thousands of Threads in CUDA™ – very easy to do in C with production quality support on Windows and Linux + good free “Eco system” and lots of developers




```
// Device code
__global__ void VecAdd(float* A, float* B, float* C, int N)
{
int i = blockDim.x * blockIdx.x + threadIdx.x;
if (i < N)
C[i] = A[i] + B[i];
}
// Host code
int main()
{
int N = ...;
size_t size = N * sizeof(float);
// Allocate input vectors h_A and h_B in host memory
float* h_A = (float*)malloc(size);
float* h_B = (float*)malloc(size);
// Initialize input vectors
...
// Invoke kernel
int threadsPerBlock = 256;
int blocksPerGrid = (N + threadsPerBlock – 1) / threadsPerBlock;
VecAdd<<<blocksPerGrid, threadsPerBlock>>>(d_A, d_B, d_C, N);
// Copy result from device memory to host memory
// h_C contains the result in host memory
cudaMemcpy(h_C, d_C, size, cudaMemcpyDeviceToHost);
```

Intel® TBB or OpenMP Programming Model

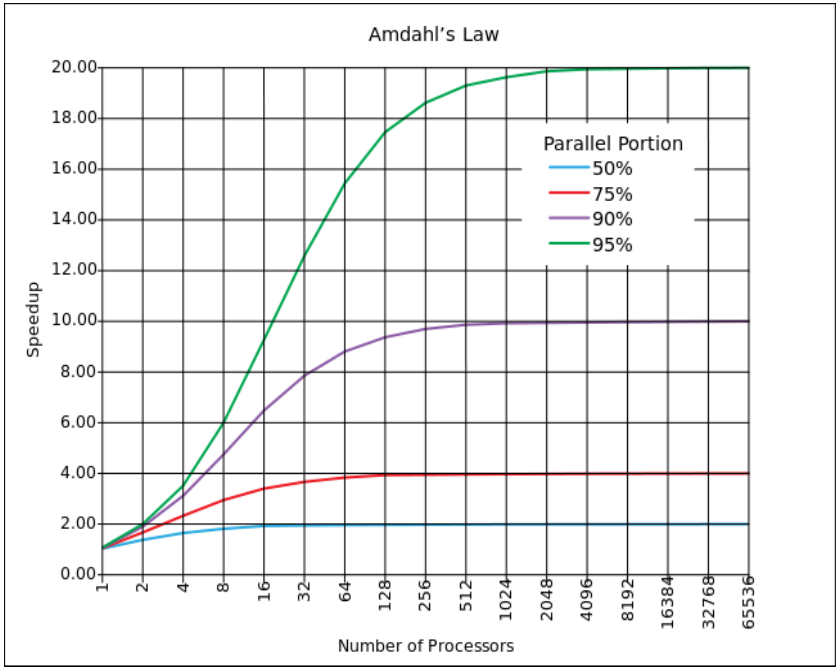
Other ways of parallelizing loops (like Intel® TBB or OpenMP) as well as vectorising code to take advantage of the vector processor are available for accelerating computation

- Ways to optimize
  - Understand where time is being spent
  - Look at ways to optimise by parallel execution
    - Vectorising using SIMD
    - Threads using TBB
- Similar considering to Low latency compute (memory layout)
- Use "more silicon" and try not to "move data" and keep the compute pipeline as full as possible
- Other techniques like Dataflow architecture (e.g. in FPGA design) is useful in dealing with streaming analytics

Why parallelization is hard – Amdahl's Law

"If it is so easy, why ain't people doing it?"

- This argument is named after Gene Amdahl
- "The speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program"
- Example: Program take 100s to execute end to end; 95% of the program can be made to run on a parallel processors (either FPGA or GPU). The part that runs on the parallel processors is 200x quicker so it will take only 0.4s to execute the parallel portion (95s/200). However, the sequential part is still 5% (i.e. 5s) so the over end to end time is 5.4s i.e. 100/5.4 or only 18.2x quicker.
- If the parallel part is speed up by 100x (e.g. using a GPU), we are talking about 0.9s or a end to end total speedup of 17x
- Hence, the debate of choosing FPGA or GPU is rather in-significant since the limitation of using either technology is not constraint by the parallel portion
- If GPU is easier to program, cheaper to buy, integration to grid schedulers, validate, maintain – the perceived 5x improvement of speed of FPGA is hard to justify for PV/Risk calculations
- In general, we find 95% is possible with Monte Carlo type code. Around 80% for tree type code.



UBS

Section 4

Q & A