XVA Principles, Nested Monte Carlo Strategies, and GPU Optimizations

S. Crépey (joint work with Lokmane Abbas-Turki and Babacar Diallo)
LaMME, Univ Evry, CNRS, Université Paris-Saclay

https://math.maths.univ-evry.fr/crepey

RIO 2017 Conference

Outline

1. Introduction
2. XVA Fundamentals (Albanese and Crépey (2017))
3. XVA Equations
4. XVA Multi-layered NMC Tree
5. Nested XVA Toy Examples
6. Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

- Since the 2008 crisis, investment banks charge to their clients, in the form of rebates with respect to the counterparty-risk-free value of financial derivatives, various pricing add-ons meant to account for counterparty risk and its capital and funding implications.
- These add-ons are dubbed XVAs, where VA stands for valuation adjustment and X is a catch-all letter to be replaced by C for credit, D for debt, F for funding, M for margin, K for capital (!), and so on.

- Pricing XVA add-ons at trade level
    - funds transfer price (FTP)
- But also accounting XVA entries at the aggregate portfolio level
    - In June 2011 the Basel Committee reported that

      > *During the financial crisis, roughly two-thirds of losses attributed counterparty credit risk were due to CVA losses and only about o third were due to actual defaults*

    - In January 2014 JP Morgan has recorded a $1.5 billion FVA loss
- Individual FTP of a trade actually computed as portfolio incremental XVAs of the trade

XVAs represent a switch of paradigm in derivative management

- From hedging to balance-sheet optimization
- Derivative portfolio optimization for a market maker
  - Fixes prices, not quantities!

- XVAs deeply affect the derivative pricing task by making it global, nonlinear, and entity-dependent
- But first, before coming to these technical implications, the fundamental points are to
  - understand what deserves to be priced and what does not
    - $\supseteq$ double counting issues
  - establish not only the pricing, but also the corresponding collateralization, accounting, and dividend policy of the bank

Outline

1. Introduction

2. XVA Fundamentals (Albanese and Crépey (2017))

3. XVA Equations

4. XVA Multi-layered NMC Tree

5. Nested XVA Toy Examples

6. Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

Counterparty Risk,

with its funding and capital implications, is at the origin of all XVAs:

CVA Credit valuation adjustment

- The value you lose due to the defaultability of your counterparties

DVA Debit valuation adjustment

- The value your counterparties lose due to your own defaultability
- The symmetric companion of the CVA
- The value you gain due to your own defaultability?
- 2011 DVA debate

FVA Funding valuation adjustment

- Cost of funding variation and initial margin (MVA merged with FVA in this part to spare one “V/DA”)
- But what about the Modigliani-Miller theorem??
- 2013 FVA debate

FDA (aka Hull’s DVA2) Funding windfall benefit at own default

KVA Cost of capital

- The price for the bank of having to reserve capital at risk
- Unsettled KVA debate

CVA−DVA+FVA−FDA [+KVA]: The XVA debates

- FVA and FDA cash flows NPV-match each other

→ CVA-DVA yields the fair, symmetrical adjustment between two counterparties of equal bargaining power

- But "Contra-liabilities" DVA and FDA are only a benefit to the creditors of the bank, whereas only the interest of shareholders matters in bank managerial decisions

→ DVA and FDA should be ignored in entry prices

→ CVA+FVA

KVA and Funds Transfer Price

- Moreover, counterparty default losses cannot be replicated and a bank must reserve shareholder capital to cope with residual risk
- Shareholders that put capital at risk deserve a remuneration at a hurdle rate, which corresponds to the KVA

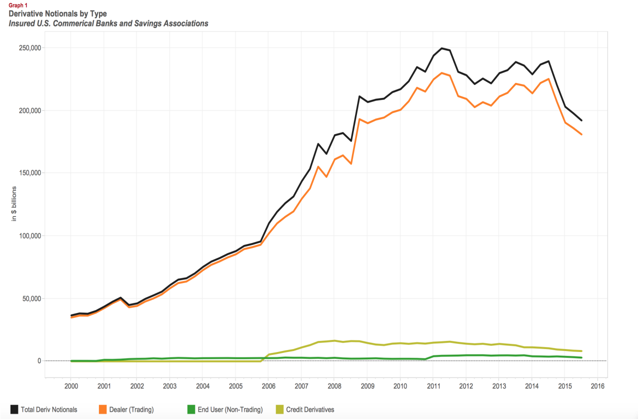
→ FTP=CVA+FVA+KVA

Incremental XVA Approach

- Meant incrementally at every new deal, the above XVA add-on is interpreted dynamically as the cost of the possibility for the bank to go into run-off,
    - i.e. lock its portfolio and let it amortize in the future,

  while staying in line with shareholder interest, from any point in time onward if wished.

→ Basel III Pillar 2 FTP as a “soft landing” or “anti-Ponzi” corrective pricing scheme accounting for counterparty risk incompleteness:



Graph 1
Derivative Notionals by Type
Insured U.S. Commerical Banks and Savings Associations

Connection with the Modigliani and Miller (1958) Theorem

- The Modigliani-Miller theorem includes two key assumptions.
  - One is that, as a consequence of trading, total wealth is conserved.
  - The second assumption is that markets are complete.
- In an XVA setup we keep the wealth conservation hypothesis but we lift the completeness.

- Hence the conclusion of the theorem, according to which the fair valuation of counterparty risk to the bank as a whole should not depend on its funding policy, is preserved.

- However, due to the incompleteness of counterparty risk, the interests of shareholders and creditors are not aligned to each other.

Outline

1 Introduction

2 XVA Fundamentals (Albanese and Crépey (2017))

3 XVA Equations

4 XVA Multi-layered NMC Tree

5 Nested XVA Toy Examples

6 Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

- Consider a bank engaged into bilateral trading with a single client, with promised cash flows process D and final maturity of the portfolio T.
- Let R_c denote the recovery rate of the client in case it defaults at time τ_c.

- Let VM denote the variation margin
  - collateral guarantee tracking the value of the client portfolio of the bank,
  - counted positively when received by the bank
- Let PIM and RIM denote the initial margins posted and received by the bank on its client portfolio
  - collateral guarantees set on top of VM against gap risk
- Let λ and $\bar{\lambda}$ denote the VM and IM funding spreads of the bank.
- Assume for simplicity an instantaneous liquidation of the bank portfolio in case the client defaults.

- Pricing stochastic basis $(\mathbb{F}, \mathbb{P})$ with risk-neutral discount factor β
- A no arbitrage risk-neutral martingale condition on the trading loss process L of the bank yields, for $0 \leq t \leq T$:

$$\begin{aligned}
\mathrm{CVA}_t &= \mathbb{E}_t\Big[\mathbb{1}_{\{t<\tau_c\leq T\}}\beta_t^{-1}\beta_{\tau_c} \times \\
&\qquad (1-R_c)\big(\mathrm{MtM}_{\tau_c} + D_{\tau_c} - D_{\tau_c-} - \mathrm{VM}_{\tau_c} - \mathrm{RIM}_{\tau_c}\big)^+\Big],\\
\mathrm{FVA}_t &= \mathbb{E}_t \int_t^T \beta_t^{-1}\beta_s \times \\
&\qquad \lambda_s\Big(\mathrm{MtM}_s - \mathrm{VM}_s - \mathrm{CVA}_s - \mathrm{FVA}_s - \mathrm{MVA}_s\Big)^+ ds\\
\mathrm{MVA}_t &= \mathbb{E}_t \int_t^T \beta_t^{-1}\beta_s\bar{\lambda}_s \mathrm{PIM}_s ds.
\end{aligned}$$

Economic Capital and Capital Valuation Adjustment

- On top of no arbitrage in the sense of risk-neutral “contra-assets” (actual liabilities)

$$\mathrm{CA} = \mathrm{CVA} + \mathrm{FVA} + \mathrm{MVA},$$

bank shareholders need be remunerated at some hurdle rate *h* for their capital at risk.

- The economic capital (EC) of the bank is dynamically modeled as the conditional Expected Shortfall (ES) at some quantile level a of the one-year-ahead loss of the bank, i.e., also accounting for discounting:

$$\mathrm{EC}_t = \mathbb{ES}_t^a\left(\int_t^{t+1} \beta_t^{-1}\beta_s dL_s\right).$$

- As seen in (Albanese and Crépey 2017), assuming a constant h, the amount needed by the bank to remunerate its shareholders for their capital at risk in the future is

$$\mathsf{KVA}_t = h\mathbb{E}_t \int_t^T e^{-\int_t^s (r_u+h)du} \mathrm{EC}_s ds \,, \quad t \in [0, T].$$

XVA Metrics for Bilateral Trade Portfolios

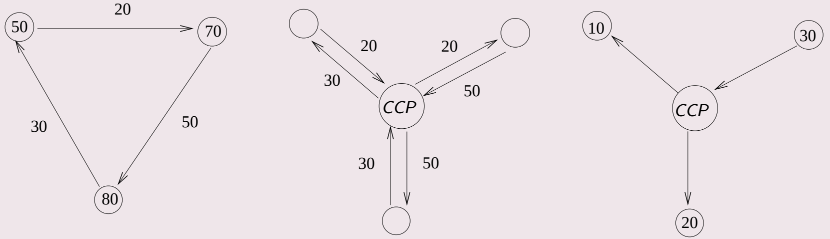
Assuming n netting sets (and one funding set):

- nonlinear CVA terminal payoffs, hence the CVA can only be computed at the level of each netting set
- semilinear FVA equation, hence the FVA can only be computed at the level of the overall portfolio of the bank
- The KVA can only be computed at the level of the overall portfolio and relies on future conditional risk measures of the trading loss process of the bank, which itself involves future fluctuations of other XVAs, as these are part of the bank liabilities

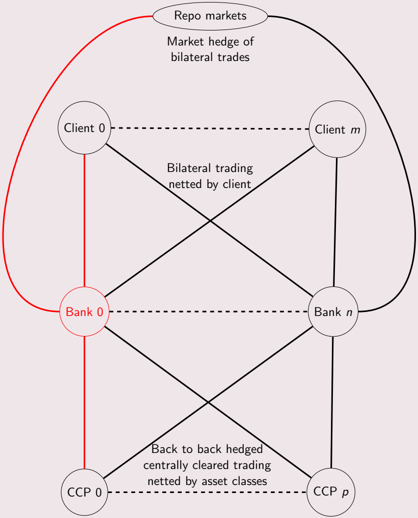
XVA Metrics for Centrally Cleared Portfolios

- Central clearing is becoming mandatory for vanilla products on the markets
- The alternative being bilateral transactions under SIMM

In a centrally cleared setup, the clearinghouse (or CCP, for "central counterparty") interposes itself in all transactions, becoming "the buyer to every seller and the seller to every buyer"

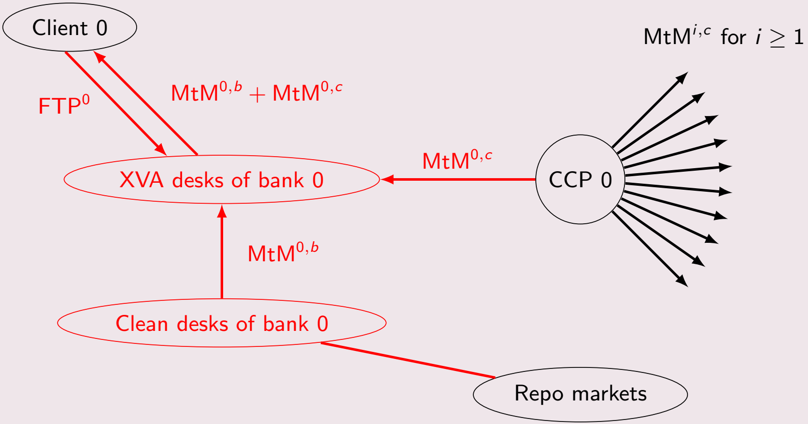


Financial network of clients, banks, and CCPs. Solid edges represent cash flows between the related entities. Bilateral trades correspond to the upper part of the picture (banks and above) and centrally cleared trades to the lower part (banks and below).



Repo markets
Market hedge of bilateral trades
Client 0
Client m
Bilateral trading netted by client
Bank 0
Bank n
Back to back hedged centrally cleared trading netted by asset classes
CCP 0
CCP p

Zoom on a reference bank, labeled by 0, focusing on its transactions with client 0 and CCP 0 (corresponding to the red part in the previous figure).



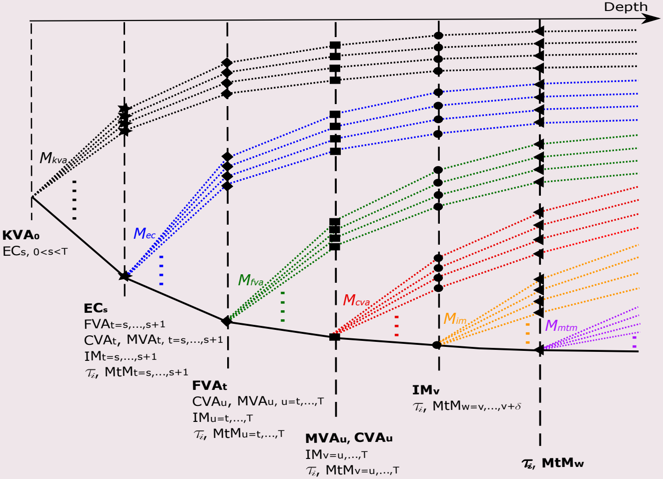
Outline

1. Introduction
2. XVA Fundamentals (Albanese and Crépey (2017))
3. XVA Equations
4. XVA Multi-layered NMC Tree
5. Nested XVA Toy Examples
6. Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

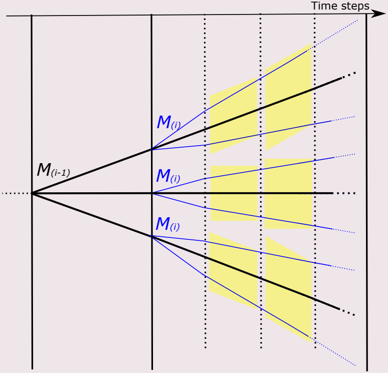
Algorithmic and computational challenges

- Heavy computations at the portfolio level
- Yet needs accuracy so that incremental XVA computations are not in the numerical noise of the machinery

XVA NMC simulation tree, from the most outer layer to the most inner one. The sub-tree rooted at the lowest node on each inner layer should be duplicated starting from each node above on the same layer.


Depth
M_{kva}
$\mathbf{KVA_0}$
EC_s, $0<s<T$
M_{ec}
$\mathbf{EC_s}$
$FVA_{t=s,\dots,s+1}$
CVA_t, MVA_t, $t=s,\dots,s+1$
$IM_{t=s,\dots,s+1}$
$\mathcal{T}_i$, $MtM_{t=s,\dots,s+1}$
M_{fva}
$\mathbf{FVA_t}$
CVA_u, MVA_u, $u=t,\dots,T$
$IM_{u=t,\dots,T}$
$\mathcal{T}_i$, $MtM_{u=t,\dots,T}$
M_{cva}
$\mathbf{MVA_u}$, $\mathbf{CVA_u}$
$IM_{v=u,\dots,T}$
$\mathcal{T}_i$, $MtM_{v=u,\dots,T}$
M_{im}
$\mathbf{IM_v}$
$\mathcal{T}_i$, $MtM_{w=v,\dots,v+\delta}$
M_{mtm}
$\boldsymbol{\mathcal{T}_i}$, $\mathbf{MtM_w}$

FVA (inner) backwardations



- If the user is only interested in some of the XVA components, then only the sub-XVA tree corresponding to the most outer XVA of interest in the figure needs be processed computationally;
- If one or several layers can be computed by exact or approximate formulas instead of Monte Carlo simulation, then the corresponding layers drop from the picture.

- Using GPU parallel computing for portfolio-wide nested Monte Carlo and conditional risk measure computations
- Using machine learning techniques for
  - dealing with certain nonlinearities that can be accounted for by stochastic approximation "quasi-regression" schemes
  - solving high-dimensional non-convex XVA compression problems

XVA NMC Design Parameterization

- Assuming the same variance created through the different layers of the tree, the order of accuracy of an $M_{(0)} \otimes M_{(1)} \otimes \ldots \otimes M_{(i)} = M_{(0)} \otimes M_{(0)} \otimes \ldots \otimes M_{(0)}$ NMC is the same as the one of an $M_{(0)} \otimes \sqrt{M_{(0)}} \otimes \ldots \otimes \sqrt{M_{(0)}}$ NMC
    - $O(M_{(0)}^{-\frac{1}{2}})$ order of accuracy
- Moreover, the variances (at least, the corresponding constants) are not homogeneous with respect to the stages.

Accordingly, the design of our XVA NMC algorithm reads as follows:

- Select layers of choice in a XVA NMC sub-tree of choice, with corresponding tentative number of simulations denoted by $M_{(0)}, \ldots, M_{(i)}$, for some $1 \leq i \leq 5$ (we assume at least one level of nested simulation).
- By dichotomy on $M_{(0)}$, reach a target relative error (in the sense of the outer confidence interval) for $M_{(0)} \times M_{(1)} \ldots \times M_{(i)}$ NMCs with $M_{(1)} = \ldots = M_{(i)} = \sqrt{M_{(0)}}$.
- For each j decreasing from i to 1, reach by dichotomy on $M_{(j)}$ a target bias (in the sense of the impact on the outer confidence interval) for $M_{(0)} \times M_{(1)} \times \ldots \times M_{(j)} \times \ldots \times M_{(i)}$ NMCs.

- For instance, considering the overall 5-layered XVA NMC, in order to ensure a 5% relative error (in the sense of the corresponding confidence interval) at a 95% confidence level, which we can take as a benchmark order of accuracy for XVA computations in banks, the above approach may lead to M_{mtm}, M_{im}, and M_{cva} somewhere between $1e2$ and $1e3$.

- As the FVA is obtained from the resolution of a BSDE that involves preconditioning, M_{fva} can be even smaller than $1e2$ without compromising the accuracy.

- Due to the approximation of the conditional expected shortfall risk measure involved in economic capital computations, M_{ec} has to be bigger than $1e3$ but usually can be smaller than $1e4$.
- As this conditional expected shortfall is an average on XVA trading loss tail values and because such tail events are mostly driven by default events rather than by market volatility swings (at least with intensity models of default times), it has then a very small variance and M_{kva} can vary between $1e2$ and $1e3$.

Using Economic Capital as Variation Margin

- Accounting for the possibility for a bank to post economic capital (EC) as variation margin (VM), the VM funding needs are reduced from $(\mathrm{MtM} - \mathrm{VM} - \mathrm{CA})^+$ to

$$(\mathrm{MtM} - \mathrm{VM} - \mathrm{CA} - \mathrm{EC}(L))^+.$$

- FVA anticipated BSDE (ABSDE)

Outline

1. Introduction
2. XVA Fundamentals (Albanese and Crépey (2017))
3. XVA Equations
4. XVA Multi-layered NMC Tree
5. Nested XVA Toy Examples
6. Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

In this part all our simulations are run on a laptop that has an Intel i7-7700HQ CPU and a GeForce GTX 1060 GPU programmed with the CUDA/C API.

Nested CVA Toy Example

CVA: $M_{cva} = 1024 * 100$, Counterparty spread $=$ 100bp.

M_{mtm}	CVA value	CI 95%	Rel. err.	Computation time (second)
128	2358.92	±76.47	3.24%	5.96
256	2358.46	±76.45	3.24%	11.54
512	2367.90	±76.49	3.23%	23.18
1024	2373.83	±76.51	3.22%	46.12

For $M_{cva} = 1024 * 100$ paths, taking $M_{mtm} = 128$ is already enough

- The gain in bias that results from taking a M_{mtm} greater than 128 is negligible with respect to the uncertainty of the simulation (size of the confidence interval).

Nested FVA Toy Example

FVA: $M_{fva} = 128 * 100$, $\lambda = 50bps$, keeping 90% of the information in the regressions for the conditional expectations.

M_{mtm}	FVA value	CI 95%	Rel. err.	Computation time (second)
128	1861.66	±52.03	2.79%	81.21
256	1861.96	±52.03	2.79%	162.12
512	1861.83	±52.03	2.79%	324.67

- The second column shows that the FVA values are already stabilized after a low number M_{mtm} (such as $M_{mtm} = 256$) of inner simulations
- The third and fourth column show that $M_{fva} = 128 * 100 \approx 10K$ outer simulations is enough to ensure a reasonable accuracy.

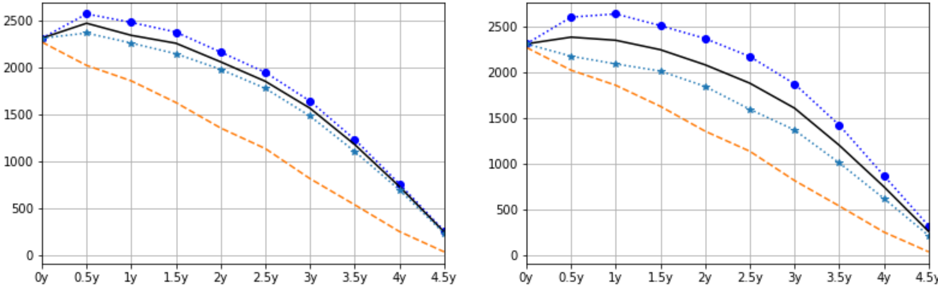
FVA: $M_{fva} = 128 * 100$, $\lambda = 50bps$, keeping 95% of the information in the regressions for the conditional expectations.

M_{mtm}	FVA value	CI 95%	Rel. err.	Computation time (second)
128	1886.60	±53.75	2.87%	81.41
256	1886.73	±53.75	2.87%	162.52
512	1886.91	±53.75	2.87%	324.97

Analogous results when 95% of the information is kept in the regressions for the conditional expectations (instead of 90% before).

Nested KVA (CCP) Toy Example

ES. Left: $M_{kva} = 128$ and $M_{ec} = 1024 * 100$. Right: $M_{kva} = 128 * 100$ and $M_{ec} = 1024 * 100$ '- -' unconditional, '. . . $*$' quantile 1%, '—' mean, '. . . ●' quantile 99%.



KVA_0 at a quantile level 99%: $M_{ec} = 1024 * 100$.

M_{kva}	KVA_0 value	CI 95%	Rel. err.	Computation time (second)
128	800.77	± 0.90	0.11%	1.42
256	802.47	± 0.60	0.00%	2.74
512	808.41	± 0.42	0.00%	5.41
1024	807.25	± 0.23	0.00%	9.34

The variance is acceptable already for a low number, such as $M_{kva} = 512$, of outer simulations

- which relates to the low volatility of the $\mathbb{ES}_t^{a_{df}}$ process

KVA_0 without nested simulation at a quantile level 99%.

M_{kva}	KVA_0 value
$128 * 100$	562.86
$256 * 100$	567.86
$512 * 100$	564.99
$1024 * 100$	560.33

The impact on the time 0 KVA of working with the unconditional expected shortfall is important, even bearing in mind the model risk intrinsic to such computations

- KVA underestimated by a factor almost two when compared with the output of the NMC computation

Toy Examples Computation Times and Speedups

CVA $M_{cva} = 1024 * 100$, $M_{mtm} = 128$; FVA $M_{fva} = 128 * 100$, $M_{mtm} = 128$; KVA $M_{kva} = 128$, $M_{ec} = 1024 * 100$ (times in seconds).

	CVA		FVA		KVA	
	Time	Speedup	Time	Speedup	Time	Speedup
Nested simulation	5.80	3.5	80.51	3.5	0.96	3.5
Listing default	0.06	1.4				
Sorting default					0.22	1.2
Outer regression			0.05	2		
Inner regression	0.06	2	2.54	2		
Risk measure					0.12	20
Other	0.1		0.64		0.11	

- NMC XVA computations (with less paths on the inner layers without loss of accuracy) are within reach provided GPU and the related optimizations are used, even for computations involving portfolios of one hundred credit names, time backwardation of nonlinearities, or conditional risk measure computations.
- Optimizations are important for effectively benefiting from the expected $\sim$ 100 speedup factor when compared with an optimized CPU implementation.

Outline

1 Introduction

2 XVA Fundamentals (Albanese and Crépey (2017))

3 XVA Equations

4 XVA Multi-layered NMC Tree

5 Nested XVA Toy Examples

6 Real Portfolio Study (Albanese, Caenazzo, and Crépey (2017))

Client Portfolio

- Representative banking portfolio with about 2,000 counterparties, 100,000 fixed income trades including swaps, swaptions, FX options, inflation swaps and CDS trades.

NMC XVA Approach

- Nested Monte Carlo simulations for approximating the loss process L required as input data in the KVA computations.
    - Contra-assets (and contra-liabilities if wished) are computed at the same time.
- Accounting for the impact on the FVA of the funding sources provided by reserve capital and economic capital
- IM $= 0$.

Modeling Choices

- Market and credit portfolio models of Albanese, Bellaj, Gimonet, and Pietronero (2011) calibrated to the relevant market data.
- Risk factors are simulated forward, whereas the backward pricing task is performed by fast matrix exponentiation in floating arithmetics.
- 20,000 primary scenarios up to 50 years in the future run on 100 underlying time points, with 1,000 secondary scenarios starting from each primary simulation node, which amounts to a total of two billion scenarios.

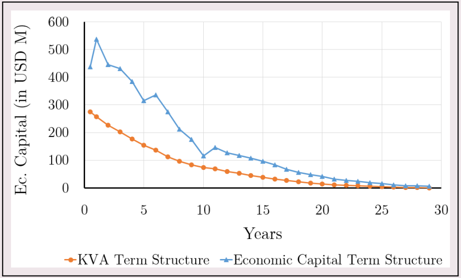
Hardware/Software Choices

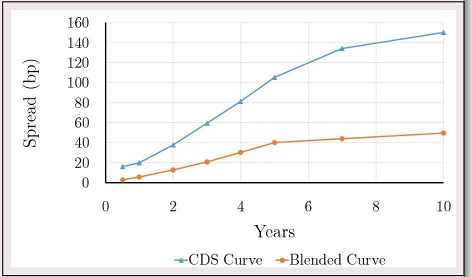
- All the computations are run using a 4-socket server for Monte Carlo simulations, Nvidia GPUs for algebraic calculations and Global Valuation Esther as simulation software.
- Using this super-computer and GPU technology the whole calculation takes a few minutes for building the models, followed by a nested simulation time in the order of about an hour for processing a billion scenarios on the bank portfolio.

XVA values for the large portfolio.

XVA	$Value
CVA_0	242 M
$\mathrm{FVA}_0^{(0)}$	126 M
FVA_0	62 M
KVA_0	275 M
FTDCVA	194 M
FTDDVA	166 M

Left: Term structure of economic capital compared with the term structure of KVA.
Right:FVA blended funding curve computed from the ground up based on capital projections.





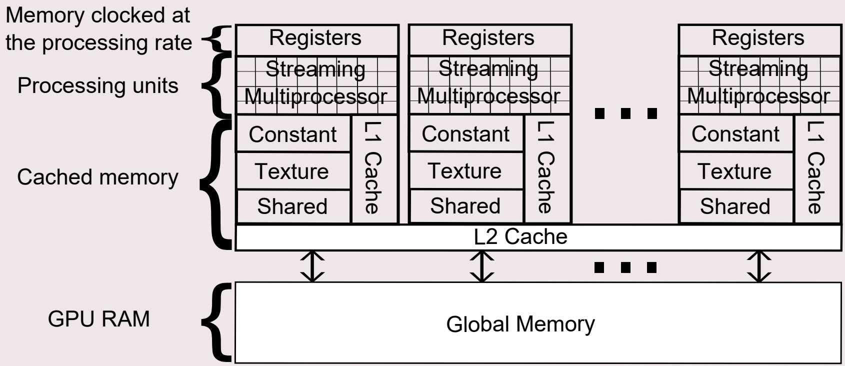
Thanks for your attention!

Outline

7. XVA NMC GPU Computing Resources Optimization Techniques

- By current standards, a GPU implementation means a potential speedup of roughly one hundred with respect to a CPU implementation
    - assuming roughly 4000 processors operating at 2GHz each on a GPU, versus 20 physical cores operating at 4GHz on a CPU

A simple presentation of the Nvidia GPU architecture



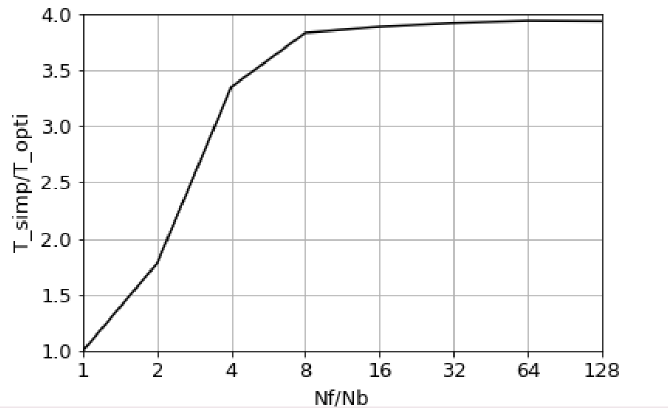
- However, the bottleneck with GPU is memory and, more precisely, memory bandwidth
  - Volume of data retrievable per second from the memory

→ GPU locality programming principle

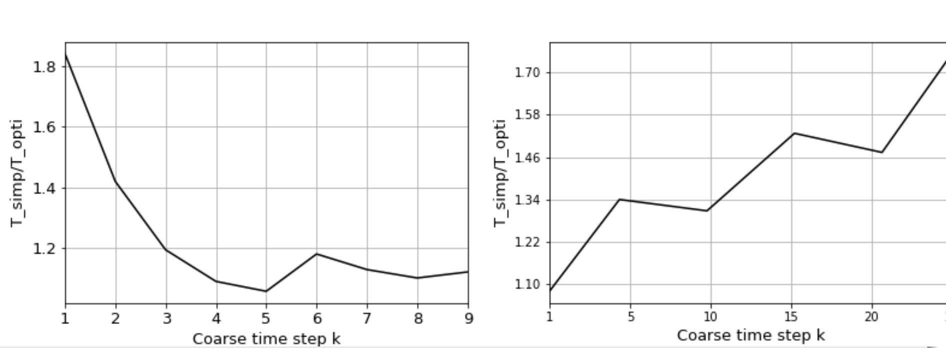
Coarse and fine parallelization strategies:

- if only one GPU is available, then the paths should be allocated between the $\sim$ 4000 streaming processors from the most inner nested layer of simulation to the most outer one, in a fine grain stratification approach;
- if several GPUs on one single node are available, then one should allocate between them the most outer paths of the simulation, using in turn a fine grain stratification approach for the allocation of the computational task between streaming processors on each on them;
- if even several computing nodes (each equipped with several GPUs) are available, then one should allocate between them the most outer paths of the simulation and, for each of them, allocate between the corresponding GPUs intermediary levels of the simulation, using in turn a fine grain stratification approach on each on them.

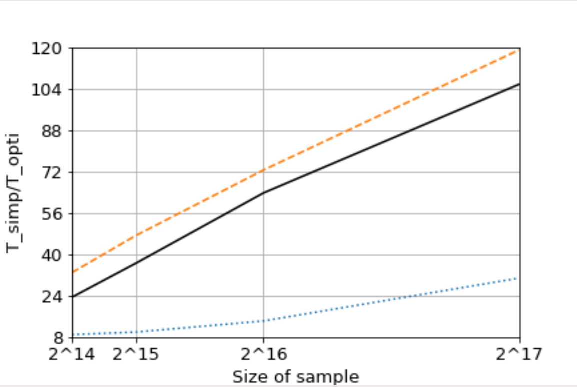
The speedup obtained from using registers and shared memory during the nested simulation of the underlying process as a function of N_b, for N_f fixed to 128.



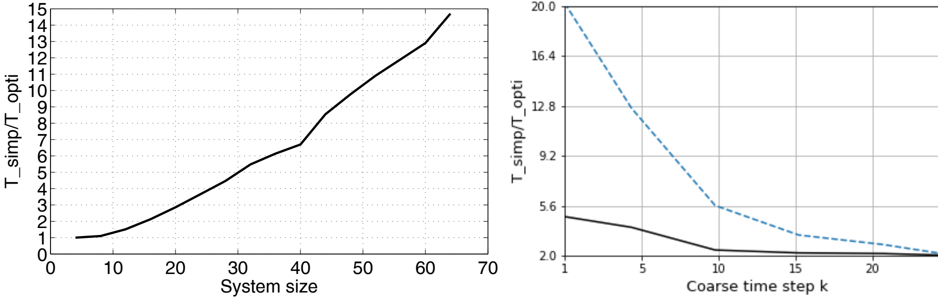
The speedup obtained from using: Sorting defaults (left), Listing defaults (right).



The speedup obtained from the truncation strategy for VaR and ES computations: '- -' 1% largest values, '—' 5% largest values, '· · ·' 10% largest values.



The speedup obtained from optimized regressions. Left: in the case of inner trajectories, by playing optimally with the number of threads n^* devoted to each linear system. Right: in the case of outer trajectories, by limiting the number of eigenvalues accounted for in the regressions. '- - ' 90% of the spectrum, '—' 95% of the spectrum.



Albanese, C., T. Bellaj, G. Gimonet, and G. Pietronero (2011). Coherent global market simulations and securitization measures for counterparty credit risk. *Quantitative Finance 11*(1), 1–20.

Albanese, C., S. Caenazzo, and S. Crépey (2017). Credit, funding, margin, and capital valuation adjustments for bilateral portfolios. *Probability, Uncertainty and Quantitative Risk 2*(7), 26 pages. Available at http://rdcu.be/tHKo.

Albanese, C. and S. Crépey (2017). XVA analysis from the balance sheet. Working paper available at https://math.maths.univ-evry.fr/crepey.

Modigliani, F. and M. Miller (1958). The cost of capital, corporation finance and the theory of investment. *Economic Review 48*, 261–297.

Villamil, A. (2008). The Modigliani-Miller theorem. In *The New Palgrave Dictionary of Economics*. Available at http://www.econ.uiuc.edu/ avillami/course-files/PalgraveRev_ModiglianiMiller_Villamil.pdf.