Aceleração de Amplificação de Privacidade via NTT em Sistemas CV-QKD: Desafios e Tendências em Hardware

Ramylla L. B. G. Bezerra [1,3], Henrique N. Teixeira [1,3], José B. de Souza Júnior [2,3], Linton T. C. Esteves [3], Paulo C. M. de A. Farias [1], Nelson A. F. Neto [3]

[1] Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC) Escola Politécnica, Universidade Federal da Bahia (UFBA) Salvador, BA, Brasil

[2]Programa de Pós-Graduação em Modelagem Computacional e Tecnologia Industrial (PPGMCTI) Universidade SENAI CIMATEC Salvador, BA, Brasil

[3]QuIIN - Quantum Industrial Innovation, SENAI CIMATEC Salvador, BA, Brasil

{ramyllabezerra, henrique.teixeira, paulo.farias}@ufba.br

{linton.esteves, jose.juniors, nelson.neto}@fieb.org.br

Abstract. *Continuous-Variable Quantum Key Distribution (CV-QKD) systems require processing massive data blocks (*10^8 *to* 10^{10} *bits) to mitigate finite-size effects and ensure information-theoretic security. This requirement creates a severe computational bottleneck in the Privacy Amplification (PA) stage. This computational delay directly impacts end-to-end network latency, limiting the viability of real-time quantum cryptography applications. This paper presents a broad survey of the literature regarding hardware acceleration for PA, focusing on the technical transition from Fast Fourier Transform (FFT) to Number-Theoretic Transform (NTT). We map the evolution of architectures across Central Processing Unit (CPU), Graphics Processing Unit (GPU), and Field Programmable Gate Array (FPGA) platforms, analyze the mathematical impact of modular arithmetic on bit-to-bit precision, and identify emerging trends in on-the-fly matrix generation using Linear-Feedback Shift Registers (LFSR). This survey establishes the theoretical and technical foundation for future high-speed quantum network implementations.*

Resumo. *Sistemas de distribuição quântica de chaves por variáveis contínuas (CV-QKD) exigem o processamento de blocos massivos de dados (*10^8 *a* 10^{10} *bits) para mitigar efeitos de tamanho finito e garantir segurança teórica da informação, o que gera um gargalo crítico na etapa de amplificação de privacidade (PA). Esse atraso computacional impacta diretamente a latência fim-a-fim da rede, limitando a viabilidade de aplicações de criptografia quântica em tempo real. Este trabalho apresenta uma revisão abrangente da literatura sobre a aceleração em hardware da PA, focando na transição técnica da transformada rápida de Fourier (FFT) para a transformada de teoria dos números (NTT). Mapeamos a evolução das arquiteturas em plataformas de CPU, GPU e FPGA, analisamos o impacto da aritmética modular na precisão bit-a-bit e*

identificamos tendências em geração de matrizes on-the-fly via registradores de deslocamento com realimentação linear (LFSR). Este levantamento fundamenta tecnicamente os requisitos para futuras implementações de redes quânticas de alta velocidade.

1. Introdução

A distribuição quântica de chaves (*Quantum Key Distribution*) (QKD) representa o estado da arte em comunicações seguras, permitindo que duas partes compartilhem chaves criptográficas com segurança incondicional, fundamentada nas leis da mecânica quântica [Luo et al. 2024]. Nesse cenário, os sistemas de distribuição quântica de chaves por variáveis contínuas (*Continuous-Variable Quantum Key Distribution*) (CV-QKD) se consolidam como uma alternativa estratégica, pois viabilizam a integração transparente com as infraestruturas de telecomunicações ópticas padrão já existentes [Pirandola et al. 2020, Usenko and Filip 2016].

No entanto, a viabilidade prática desses sistemas em taxas da ordem de Gigabits por segundo (Gbps) esbarra na altíssima demanda computacional do pós-processamento clássico. Dentre as etapas desse processamento, que tipicamente incluem a reconciliação de bases (*sifting*), estimativa de parâmetros e reconciliação de informações (correção de erros), a amplificação de privacidade (*Privacy Amplification*) (PA) destaca-se como o gargalo mais severo para o processamento em tempo real [Luo et al. 2024]. Ela é a etapa final responsável por destilar a chave secreta definitiva, eliminando qualquer informação residual que possa ter vazado para um espião (Eve) durante a transmissão quântica ou durante as trocas de mensagens no canal clássico [Bennett et al. 1995].

Para garantir uma segurança teórica rigorosa contra os efeitos de tamanho finito (*finite-size effects*), a literatura demonstra que o processamento de dados em blocos massivos é imprescindível. Conforme a distância do canal aumenta, o tamanho do bloco n deve crescer proporcionalmente para mitigar as penalidades impostas à taxa de chave secreta [Leverrier et al. 2010]. Blocos na ordem de 10^8 bits são o padrão exigido para redes metropolitanas (cerca de 50 km), podendo escalar para 10^{10} bits em conexões de longa distância [Wu et al. 2025]. Além da distância, a eficiência da PA é determinante para o *throughput* efetivo da rede; sem uma aceleração que acompanhe o ritmo de recepção dos qubits, o acúmulo de dados brutos causaria um aumento exponencial na latência, inviabilizando aplicações que exigem chaves sob demanda (*key-on-demand*).

O custo computacional para realizar a extração por meio de matrizes de Toeplitz escala de forma quadrática ($\mathcal{O}(n^2)$) na sua implementação direta, tornando-se proibitivo em arquiteturas de propósito geral [Wang et al. 2018]. Embora o uso da transformada rápida de Fourier (*Fast Fourier Transform*) (FFT) em matrizes circulantes consiga reduzir essa complexidade para $\mathcal{O}(n \log n)$, a sua inerente dependência da aritmética de ponto flutuante introduz erros de arredondamento numérico que comprometem a exatidão da chave final [Luo et al. 2024].

Diante desse desafio, este trabalho apresenta um levantamento sobre a transição algorítmica para a transformada de teoria dos números (*Number-Theoretic Transform* (NTT), uma solução operada sobre corpos finitos modulares que garante precisão absoluta (bit-a-bit) sem perdas. O objetivo é fundamentar teoricamente os desafios atuais e mapear as tendências para o desenvolvimento de futuras arquiteturas de aceleração em *hardware*

de alto desempenho, visando a operacionalidade de redes quânticas de alta velocidade [Wu et al. 2025].

2. Mapeamento de Plataformas

Esta seção apresenta um levantamento das principais abordagens encontradas na literatura, partindo das implementações clássicas em software até o estado da arte em hardware dedicado.

2.1. Limitações Técnicas

As implementações iniciais de PA em unidades centrais de processamento (do inglês, *Central Processing Unit*) (CPUs) evidenciaram os limites da arquitetura de von Neumann. Em 2014, o uso de transformadas rápidas em CPU atingiu apenas 14,86 Mbps para blocos de 10^6 bits [Zhang et al. 2014]. O gargalo principal é a largura de banda de memória, que impede o processamento em escala de Gbps.

Para elevar o *throughput*, unidades de processamento gráfico (*Graphics Processing Unit*) (GPUs) foram adotadas devido ao paralelismo massivo, superando 1,3 Gbps [Wang et al. 2018]. Contudo, as GPUs apresentam obstáculos: (1) alto consumo de potência térmica (*Thermal Design Power*) (TDP); (2) latência no barramento *PCI Express* entre *host* e memória de vídeo (*Video RAM*) (VRAM); e (3) a dependência da FFT, que é suscetível a erros de arredondamento no domínio complexo [Wu et al. 2025].

2.2. O Paradigma NTT em Lógica Programável

Para superar as limitações de latência e precisão, os arranjos de portas programáveis em campo (*Field Programmable Gate Arrays*) (FPGAs) tornaram-se o paradigma dominante. FPGAs permitem a implementação de *pipelines* aritméticos para operações sobre corpos finitos com baixo consumo de energia. A adoção da NTT pura em hardware atingiu taxas de 10 Gbps a 18 Gbps com processamento em tempo real (*on-the-fly*) [Wu et al. 2025, Bai et al. 2021].

Tabela 1. Comparativo técnico entre paradigmas mapeados na literatura para PA.

Métrica de Análise	**CPU**	**GPU**	**FPGA (Estado da Arte)**
Matemática Subjacente	FFT	FFT	NTT
Precisão Numérica	Ponto Flutuante	Ponto Flutuante	Exata (Modular)
Throughput Médio	< 50 Mbps	> 1 Gbps	1 Gbps – 18 Gbps
Eficiência Energética	Baixa	Muito Baixa	Alta
Gargalo de I/O	Memória RAM	Barramento PCIe	Largura de Banda BRAM
Escalabilidade de Bloco	Limitada	Média	Alta (On-the-fly)
Referências Base	[Zhang et al. 2014]	[Wang et al. 2018]	[Bai et al. 2021, Wu et al. 2025]

3. Análise da Fundamentação Técnica: FFT vs NTT

A fundamentação da PA baseia-se na multiplicação densa $K = T \times W$, onde W é o vetor de dados parciais reconciliados, T é uma matriz de Toeplitz gerada a partir de uma semente pública, e K é a chave final amplificada. O mapeamento da literatura demonstra que embutir a matriz de Toeplitz em uma matriz circulante para aplicar o teorema da convolução cíclica é a abordagem canônica para a redução de complexidade.

3.1. O Problema do Arredondamento Numérico na FFT

A convolução via FFT ocorre no domínio dos números complexos $\mathbb{C}$, operando através das chamadas raízes da unidade expressas por $e^{-i2\pi/n}$. Por envolverem funções trigonométricas (senos e cossenos irracionais), esses cálculos não podem ser representados de maneira finita em silício. O *hardware* comercial é forçado a truncar a mantissa utilizando o formato de ponto flutuante, inserindo ruído de discretização cumulativo no *pipeline*. O erro computacional propagado na reconstrução (via transformada inversa) altera a integridade final da chave. Uma perturbação de um único bit na extração de Alice e Bob destrói a paridade exigida na chave final simétrica.

3.2. Determinismo via Corpos Finitos e NTT

A literatura atual consagra a NTT como a substituta exata. A NTT resolve a inexatidão algébrica da FFT substituindo o domínio contínuo $\mathbb{C}$ pelo anel dos inteiros módulo p ($\mathbb{Z}_p$), operando através de uma raiz primitiva ω estrita da unidade de ordem N. A operação no tempo discreto e a convolução são matematicamente idênticas à FFT, mas definidas como:

$$X[k] = \left(\sum_{n=0}^{N-1} x[n]\omega^{nk} \right) \pmod{p} \tag{1}$$

Como a integralidade matemática é preservada no ciclo das variáveis de base, nenhum ruído de ponto flutuante é gerado. Contudo, em termos de implementação lógica, a divisão modular convencional (*modulo p*) é uma operação densa, proibitivamente lenta em unidades lógico-aritméticas (ALUs).

Para contornar o gargalo das divisões no *hardware*, a literatura de vanguarda mapeia o uso extensivo de "Primos de Solinas"(como $p = 2^{64} - 2^{32} + 1$) [Van Assche 2006]. A característica especial desses números na base binária permite que o cálculo da redução modular se restrinja apenas a simples operações de deslocamento de bits (*bit-shifting*) e somas, dispensando completamente o uso de divisores lentos no circuito interno.

4. Tendências em Arquiteturas de Hardware de Alto Desempenho

O levantamento sistemático sobre as propostas recentes em FPGA aponta que o sucesso de aceleradores de taxa de bits superior a 5 Gbps repousa inteiramente sobre dois pilares de projeto microarquitetural: paralelismo massivo configurável e eliminação da latência de armazenamento.

4.1. Paralelismo Radix-r e Redes Butterfly Otimizadas

A NTT tradicional baseada em Radix-2 exige $\log_2 N$ estágios de memória para processar os dados na rede de borboleta (*butterfly network*). Em blocos longos, a latência de

trânsito em estágios profundos colapsa a temporização do FPGA (*timing constraints*). O estado da arte consolida a tendência de processadores NTT de alta ordem (como Radix-8 ou Radix-16). Ao processar simultaneamente múltiplos coeficientes e raízes por ciclo de *clock*, o caminho da transformada colapsa para $\log_r N$ estágios [Wu et al. 2025]. Apesar de consumir significativamente mais módulos de processamento de sinais digitais (*Digital Signal Processor*) (DSPs) embarcados no *chip*, o ganho drástico em largura de banda justifica amplamente a escolha arquitetural de FPGAs modernos focados em DSP intensivo.

4.2. Geração Dinâmica de Matriz via LFSR

Outro avanço elementar mapeado neste levantamento responde ao maior impeditivo dos processadores embarcados: o esgotamento dos recursos de memória. Mesmo utilizando a propriedade simplificada de Toeplitz, armazenar uma semente de dezenas de *Megabits* requereria centenas de instâncias de SRAM/BRAM, exaurindo recursos que deveriam ser empregados em FIFOs de rede.

A solução consolidada pela literatura para o gargalo de memória é a geração on-the-fly via registrador de deslocamento com realimentação linear (*Linear-Feedback Shift Register*) (LFSR) [Krawczyk 1994]. Contudo, essa abordagem introduz um compromisso crítico entre eficiência de área e segurança. Enquanto geradores congruenciais lineares tradicionais podem apresentar correlações lineares exploráveis, a literatura de vanguarda aponta para o uso de LFSRs baseados em polinômios primitivos de alto grau ou combinados com funções de hashing para garantir que a matriz de Toeplitz gerada mantenha as propriedades de extração de aleatoriedade exigidas pelo teorema de Leftover Hash Lemma. A principal limitação atual reside na rigidez do hardware: uma vez fixado o polinômio no silício, a adaptação para diferentes sementes ou tamanhos de bloco pode exigir a reconfiguração parcial do FPGA, um campo de pesquisa ainda em aberto para sistemas de rede dinâmicos.

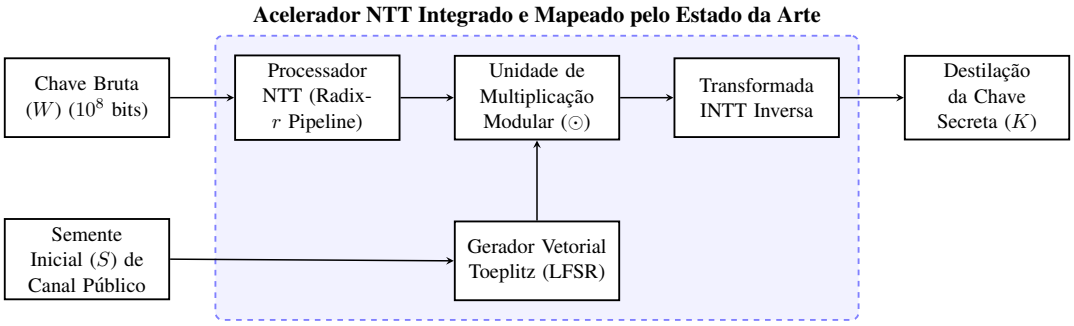

Figura 1. Fluxo de processamento de dados para futuras implementações no *hardware* de CV-QKD.

5. Considerações Finais

Este trabalho efetuou um rigoroso levantamento técnico abordando os gargalos históricos de Amplificação de Privacidade enfrentados pelas metodologias em CV-QKD no limite de latências exigidas por redes metropolitanas. A análise fundamenta de forma clara os motivos pelos quais o esgotamento do poder algorítmico da CPU e a ineficiência energética e numérica das GPUs levaram os FPGAs paralelos dedicados a se tornarem as plataformas canônicas de escolha. Em contrapartida aos desvios flutuantes da arquitetura FFT padrão,

a matemática orientada aos domínios finitos sob a teoria da NTT desponta de maneira unânime como o caminho mais viável, escalável e de acurácia extrema. A combinação entre as arquiteturas dinâmicas Radix de alta ordem e a geração em tempo real através dos processadores de retroalimentação linear (LFSR) estabelece, sob esta ótica de levantamento, o alicerce e as diretrizes definitivas a serem adotadas na construção física dos novos processadores aceleradores clássicos em redes e protocolos de comunicação quântica segura na ordem dos Gbps.

Agradecimentos

Este trabalho foi totalmente financiado pelo projeto "HW DSP: Desenvolvimento e Prototipagem de SoC Multicore com Aceleradores Dedicados e DSP RISC-V" suportado pelo QuIIN - Quantum Industrial Innovation, Centro de Competência EMBRAPII CIMATEC em Tecnologias Quânticas, com recursos financeiros oriundos do PPI IoT/Manufatura 4.0 do MCTI, através do Termo de Cooperação 053/2023, firmado com a EMBRAPII.

Referências

Bai, B., Huang, J., Qiao, G.-R., Nie, Y.-Q., Tang, W., Chu, T., Zhang, J., and Pan, J.-W. (2021). 18.8 gbps real-time quantum random number generator with a photonic integrated chip. *Applied Physics Letters*, 118(26):264001.

Bennett, C. H., Brassard, G., Crépeau, C., and Maurer, U. M. (1995). Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6):1915–1923.

Krawczyk, H. (1994). Lfsr-based hashing and authentication. In *Advances in Cryptology — CRYPTO '94*, pages 129–139. Springer.

Leverrier, A., Grosshans, F., and Grangier, P. (2010). Finite-size analysis of a continuous-variable quantum key distribution. *Physical Review A*, 81(6):062343.

Luo, Y., Cheng, X., Mao, H.-K., and Li, Q. (2024). An overview of postprocessing in quantum key distribution. *Mathematics*, 12(14):2243.

Pirandola, S., Andersen, U. L., Banchi, L., Berta, M., Bunandar, D., Colbeck, R., Englund, D., Gehring, T., Lupo, C., Ottaviani, C., et al. (2020). Advances in quantum cryptography. *Advances in Optics and Photonics*, 12(4):1012–1236.

Usenko, V. C. and Filip, R. (2016). Continuous-variable quantum key distribution. *Entropy*, 18(1):20.

Van Assche, G. (2006). *Quantum Cryptography and Secret-Key Distillation*. Cambridge University Press.

Wang, X., Zhang, Y., Yu, S., and Guo, H. (2018). High-speed implementation of length-compatible privacy amplification in continuous-variable quantum key distribution. *IEEE Photonics Journal*, 10(3):1–9.

Wu, X. et al. (2025). High-speed number theoretic transform for quantum cryptography applications. *Journal of Lightwave Technology*.

Zhang, J. et al. (2014). Fast privacy amplification in quantum key distribution. *Scientific Reports*.