

# Quantization-Aware Optimization for Cryptographic Neural Network Inference

Mariah Stuart

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,  
OR, USA.

mariah.work@oregonstate.edu

## Abstract

Cryptographic neural network inference seeks to preserve the confidentiality of inputs, models, and outputs while still allowing useful predictions to be computed across trust boundaries. The practical difficulty is that secure computation protocols impose substantial communication, memory, and latency costs, especially when neural networks contain nonlinear activation functions and high-precision parameters. Quantization-aware optimization has emerged as a critical systems strategy because it changes both the arithmetic intensity and the protocol structure of private inference. This paper examines quantization-aware optimization as a socio-technical architecture problem rather than a narrow numerical compression technique. It analyzes how reduced precision interacts with secure multiparty computation, homomorphic encryption, trusted execution, and hybrid deployment models. The discussion emphasizes structural trade-offs among accuracy, privacy, latency, energy use, robustness, fairness, and governance. It argues that quantization can reduce cryptographic overhead by simplifying nonlinear operations, lowering communication volume, and enabling integer-arithmetic execution, yet it can also introduce approximation vulnerabilities, distributional bias, and auditing difficulties. The paper considers infrastructure implications for cloud, edge, and regulated environments, as well as lifecycle sustainability concerns. It concludes that quantization-aware cryptographic inference requires co-design across machine learning, cryptography, hardware, policy, and organizational governance. Without integrated evaluation, optimization may improve benchmark performance while weakening accountability, equity, and long-term maintainability.

## Keywords

cryptographic inference, quantization, private machine learning, secure multiparty computation, homomorphic encryption, systems optimization, fairness, governance, sustainability.

## 1. Introduction

Private neural network inference addresses a fundamental tension in modern machine learning: useful prediction often depends on sensitive data and valuable models, yet centralized computation can expose both. A hospital may wish to query a remote diagnostic model without revealing patient records, while a model provider may wish to protect proprietary weights. These requirements have motivated a rich body of work in secure computation, including foundational privacy definitions [1], privacy homomorphisms [2], secure two-party protocols [3], general secure multiparty computation [4], and fully homomorphic encryption [5]. Each approach offers different assumptions and performance profiles, but all confront the same system-level problem: neural network inference is computationally irregular, data-intensive, and dominated by operations that do not map cleanly onto cryptographic primitives.

Quantization-aware optimization enters this space because precision is not merely an implementation detail. The number of bits used for weights, activations, and intermediate accumulations determines memory traffic, arithmetic throughput, communication payloads, and the complexity of nonlinear function evaluation. In secure inference, these factors are amplified because every multiplication, comparison, and activation may require interaction between parties or expansion within encrypted representations. Techniques such as low-precision training, integer-only inference, and activation reduction have therefore become central to making private inference economically viable [13][14][15]. However, treating quantization as a pure efficiency lever obscures its architectural consequences. Reduced precision changes the protocol selection space, the hardware mapping, the robustness surface, and the governance obligations of deployed systems.

This paper argues that quantization-aware optimization for cryptographic neural network inference should be understood as a co-design discipline. It joins numerical representation, cryptographic protocol design, accelerator architecture, deployment topology, fairness auditing, and policy oversight. The analysis is organized around structural trade-offs rather than isolated benchmark gains. Section two frames cryptographic inference as a systems problem. Section three examines quantization as an architectural lever. Section four analyzes precision, security, and latency trade-offs. Section five discusses hardware and deployment infrastructure. Section six addresses robustness, fairness, and governance. Section seven considers sustainability and lifecycle implications. Section eight provides cross-domain comparisons and case illustrations. Section nine outlines a forward-looking research agenda, and section ten concludes.

## **2. Cryptographic Inference as a Systems Problem**

Cryptographic inference is often described as a privacy-preserving alternative to plaintext model serving, but its operational profile differs sharply from conventional inference. Secure multiparty computation protocols allow parties to jointly evaluate a function without revealing private inputs, using secret sharing, garbled circuits, or hybrid combinations [3][4][7]. Homomorphic encryption permits computation directly over encrypted data, but fully homomorphic evaluation remains expensive for deep networks [5]. Trusted execution environments offer another route, yet they depend on hardware trust assumptions and side-channel resistance. In practice, deployed systems often combine several mechanisms, selecting protocols according to layer type, tensor shape, and latency budget.

The systems burden is especially severe for nonlinear operations. Convolution and matrix multiplication can be expressed as additive or multiplicative operations that map reasonably well to secure arithmetic. Activation functions, pooling, normalization, and comparisons are more difficult because they require branching, lookup tables, approximations, or interactive protocols. Prior systems such as MiniONN, SecureML, Chameleon, Delphi, and CrypTen illustrate different points in this design space, balancing assumptions, communication rounds, and supported network architectures [7][8][9][10][11]. Their differences show that private inference is not a single protocol but a family of architectures with distinct failure modes and cost structures.

Within this landscape, reduction of nonlinear complexity becomes a first-order optimization objective. A related line of work examines how reducing reliance on rectified linear units can lower the cost of two-party private inference [12]. This is significant because activation functions often force secure protocols into expensive comparison or oblivious selection routines. If a model can be trained or transformed to use fewer costly nonlinearities, the

cryptographic workload may fall substantially. Yet such transformations are not neutral. They alter model behavior, calibration, and interpretability. They also create dependencies between training-time choices and inference-time protocol selection. A quantization-aware pipeline must therefore reason about the whole stack, from loss functions and regularization to communication scheduling and runtime orchestration.

From an infrastructure perspective, cryptographic inference also changes the economics of serving. Latency is not only compute-bound but also network-bound. Communication rounds can dominate wide-area deployments, and memory pressure can determine whether a model fits on an edge device or requires a datacenter. In regulated industries, audit trails, key management, and attestation add further layers. Consequently, optimization must consider availability, observability, and incident response. A private inference service that is fast but opaque may be unacceptable in clinical or financial settings. Quantization-aware design must therefore be evaluated against service-level objectives that include trust, compliance, and reproducibility, not only throughput.

### **3. Quantization as an Architectural Lever**

Quantization reduces the numerical precision used to represent model parameters and activations. Classical deep compression demonstrates that pruning, trained quantization, and encoding can dramatically reduce model size and energy use [13]. Integer-only inference frameworks further show that fixed-point arithmetic can replace floating-point operations without severe accuracy loss when training accounts for quantization effects [14]. Binarized and highly quantized networks push this logic further by constraining weights and activations to very low precision [15]. These techniques are often motivated by edge deployment, but they have profound implications for cryptographic inference because secure protocols often perform better with smaller integers and simpler arithmetic.

The architectural leverage of quantization appears in several layers. First, smaller operands reduce communication volume in secret-shared and homomorphic settings. Second, integer arithmetic can simplify protocol compatibility, especially when addition, multiplication, and comparison are implemented over finite domains. Third, reduced activation precision can make lookup tables or polynomial approximations more tractable. Fourth, lower memory footprints improve cache behavior and enable larger models to fit within constrained secure enclaves. In this sense, quantization is not just compression; it is a redesign of the computational substrate that cryptographic protocols operate upon.

However, quantization-aware optimization must be integrated with training rather than applied after the fact. Post-training quantization can be effective, but it often struggles when activation distributions are long-tailed or when the model relies on fine-grained numerical distinctions. Quantization-aware training introduces simulated quantization during learning, allowing the model to adapt to reduced precision. In cryptographic settings, the training objective may also need to account for protocol-specific costs. For example, a model may benefit from fewer activation boundaries, smoother nonlinearities, or weight distributions that reduce the frequency of expensive operations. This creates a joint optimization problem involving accuracy, protocol cost, and robustness.

System architects must also consider how quantization interacts with model architecture. Depthwise separable convolutions, attention mechanisms, and normalization layers have different quantization sensitivities. Some layers may tolerate aggressive reduction, while others require higher precision to preserve calibration. Secure inference systems may

therefore benefit from mixed-precision plans that allocate precision according to cryptographic cost and accuracy impact. Such plans introduce new scheduling challenges: the runtime must track precision metadata, manage conversions, and ensure that protocol transitions do not introduce leaks or errors. Quantization thus becomes a cross-layer concern that spans model design, compiler optimization, protocol selection, and hardware execution.

#### **4. Trade-offs in Precision, Security, and Latency**

The most visible trade-off in quantization-aware cryptographic inference is accuracy versus efficiency. Lower precision reduces communication and computation, but it can degrade prediction quality, especially on rare classes or out-of-distribution inputs. In private inference, this trade-off is complicated by the fact that the model owner and data owner may have different tolerances for error. A medical application may require high sensitivity, while a recommendation service may prioritize latency. The optimization target is therefore not a single accuracy threshold but a multi-stakeholder negotiation over acceptable risk.

Security introduces another dimension. Quantization does not directly weaken cryptographic assumptions, but it can affect the observable behavior of the system. Timing, memory access patterns, and communication volumes may vary with input precision or activation sparsity, creating side channels. Secure protocols often aim for data-oblivious execution, yet quantization-aware optimizations such as dynamic precision or conditional layer skipping can reintroduce input-dependent behavior. Similarly, approximate activation functions may leak information through error patterns if not carefully designed. Security analysis must therefore cover not only the cryptographic core but also the numerical approximations and runtime adaptations surrounding it.

Latency and throughput are also shaped by protocol structure. In two-party computation, round trips can dominate, so reducing the number of nonlinear operations may matter more than reducing total arithmetic. In homomorphic encryption, multiplicative depth and ciphertext expansion may dominate, so quantization must be evaluated alongside packing and batching strategies. In trusted execution, memory bandwidth and enclave paging may be the limiting factors. A quantization scheme that improves one deployment model may be suboptimal or even harmful in another. This variability argues against universal quantization recipes and in favor of architecture-aware optimization.

Finally, privacy-utility trade-offs cannot be reduced to technical metrics alone. Differential privacy offers a formal framework for quantifying disclosure risk [1], but it is not a substitute for governance. Secure computation protects data in use, while differential privacy protects against inference from outputs. Quantization may interact with both by changing the information content of outputs and the precision of gradients or updates. A mature system must document these interactions, test them across threat models, and provide meaningful transparency to users and regulators. Without such evaluation, latency improvements may conceal new forms of risk.

#### **5. Hardware, Deployment, and Infrastructure Considerations**

Quantization-aware cryptographic inference cannot be separated from hardware. General-purpose processors are inefficient for both low-precision arithmetic and cryptographic operations, leading to interest in accelerators that support integer matrix multiplication, modular arithmetic, and cryptographic primitives. Efficient deep learning accelerators already exploit reduced precision, reuse, and parallelism [16]. Cryptographic inference adds requirements for constant-time execution, secure key storage, and high-bandwidth

communication. The resulting hardware stack is heterogeneous and often distributed across edge devices, private clouds, and public clouds.

Deployment topology matters because private inference involves multiple parties with different trust assumptions. A hospital may run a local quantizer and encryption client, a cloud provider may host the model and secure compute, and a regulator may require audit logs. Each boundary introduces latency, key management, and policy enforcement. Edge deployment can reduce communication but constrains memory and energy. Cloud deployment can provide scale but raises concerns about data residency and multi-tenancy. Hybrid designs may split layers across environments, but they complicate quantization planning because precision must be consistent across partitions.

Infrastructure orchestration also affects reliability. Cryptographic inference services must handle key rotation, protocol versioning, fallback paths, and partial failures. Quantized models may be compiled into different artifacts for different hardware, creating a matrix of configurations that must be tested and certified. Model updates introduce another challenge: a new quantized model may change both prediction behavior and protocol cost. Continuous integration must therefore include privacy tests, accuracy tests, and performance tests. In regulated settings, reproducibility requires that the exact quantization configuration, protocol version, and hardware attestation be recorded.

The operational cost of cryptographic inference is not only monetary. It includes cognitive load on developers who must understand both machine learning and cryptography. Tooling that hides complexity can improve adoption, but it can also obscure assumptions. A compiler that automatically selects quantization and protocol parameters may produce efficient binaries while making it difficult to audit why a particular precision was chosen. Governance therefore depends on explainable infrastructure, where optimization decisions can be inspected, reproduced, and challenged. Quantization-aware systems should expose provenance metadata and support independent evaluation.

## **6. Robustness, Fairness, and Governance**

Quantization can alter robustness in ways that are not obvious from clean accuracy. Low-precision models may be more vulnerable to adversarial perturbations, distribution shift, and rare-input instability. Cryptographic inference may amplify these effects because inputs are hidden and debugging is limited. If a model misclassifies a private input, the data owner may not be able to inspect intermediate representations. The model owner may not see the input either. This opacity complicates error analysis and redress. Robustness testing must therefore occur before deployment and include adversarial, out-of-distribution, and corner-case evaluations.

Fairness is similarly entangled with numerical representation. Quantization may disproportionately affect subgroups whose feature distributions occupy narrow numerical ranges. A model that performs well on average may exhibit larger disparities after aggressive precision reduction. Research on causal reasoning and discrimination shows that fairness cannot be reduced to statistical parity alone [17], and work on equality of opportunity highlights the importance of group-aware evaluation [18]. In cryptographic settings, the challenge is heightened because sensitive attributes may be hidden. Fairness auditing may require privacy-preserving computation or trusted third parties, raising questions about consent and purpose limitation.

Security and privacy governance must also address the broader machine learning attack surface. Prior surveys document threats such as model extraction, membership inference, and adversarial examples [19]. Quantization may mitigate some attacks by reducing information leakage, but it may create new vulnerabilities through approximation artifacts. A quantized model may also be harder to watermark or verify, affecting intellectual property protection. Policymakers and standards bodies need technical evidence about these interactions. Without it, regulation may either over-restrict beneficial privacy technologies or under-protect affected populations.

Fairness and machine learning literature emphasizes that technical fixes alone cannot resolve structural inequities [20]. Quantization-aware cryptographic inference should therefore be embedded in organizational processes that include impact assessments, stakeholder consultation, and ongoing monitoring. Governance must define who selects precision levels, who bears the cost of errors, and how affected parties can contest outcomes. These questions are not peripheral to optimization. They determine whether a system is deployable in high-stakes domains. A fast private inference service that exacerbates bias or evades accountability is not a successful system.

## **7. Sustainability and Lifecycle Implications**

Sustainability is an increasingly important dimension of systems design. Quantization can reduce energy consumption by lowering memory traffic and arithmetic intensity, especially in edge deployments [13][16]. Cryptographic inference, however, adds substantial overhead through encryption, communication, and redundant computation. The net environmental effect depends on where computation occurs, how long models remain in service, and how often protocols are retrained or recompiled. A quantized model that fits on a local device may avoid datacenter cooling and network transmission, but it may also require more frequent battery charging or hardware replacement. Lifecycle assessment must consider embodied carbon, not only operational energy.

Model updates and protocol evolution also affect sustainability. Frequent retraining can consume significant resources, while infrequent updates can leave systems vulnerable or inaccurate. Quantization-aware pipelines should support incremental recompilation and modular upgrades. If a new precision scheme requires retraining the entire model, the environmental cost may outweigh efficiency gains. Similarly, cryptographic agility requires that protocols can be replaced without discarding the model. Sustainable design favors separation of concerns: model architecture, quantization policy, protocol suite, and hardware target should be versioned independently where possible.

From an infrastructure perspective, shared services can improve utilization but may increase data movement. A regional private inference service might aggregate demand, yet it must respect data residency and trust boundaries. Energy-aware scheduling could shift non-urgent inference to times or locations with cleaner electricity, but privacy constraints may limit flexibility. Quantization can help by reducing payloads and enabling lower-power hardware, but it does not eliminate the fundamental cost of secure computation. Sustainability claims should therefore be based on transparent measurement and third-party verification rather than vendor assertions.

## **8. Cross-Domain Comparisons and Case Illustrations**

Different domains illustrate distinct quantization-aware cryptographic inference requirements. In healthcare, patient data is highly sensitive, model interpretability is often mandated, and

errors can have severe consequences. A hospital might use secure inference to query a diagnostic model without revealing records, but quantization must preserve sensitivity for rare conditions. Governance must address consent, clinical validation, and liability. In finance, fraud detection and credit scoring involve regulatory scrutiny, adversarial actors, and real-time constraints. Quantization can reduce latency, but fairness auditing and explainability remain essential. Cryptographic inference may protect customer data, yet it can complicate dispute resolution.

Public sector applications, such as benefits eligibility or identity verification, raise even stronger concerns about due process and equity. Private inference might protect vulnerable populations from unnecessary data exposure, but it can also obscure decision logic. If quantization introduces errors that disproportionately affect marginalized groups, the system may violate legal or ethical standards. Cross-domain comparison shows that the same technical optimization can have different acceptability depending on institutional context. There is no universal precision policy; there are only context-specific trade-offs that must be negotiated and documented.

Industrial internet and autonomous systems provide another contrast. Here, latency and reliability may dominate, and models may run on constrained devices. Quantization is often necessary for real-time operation, while cryptographic inference may be used for cross-organizational collaboration, such as sharing predictive maintenance models without revealing proprietary sensor data. The deployment environment may be disconnected or intermittently connected, requiring asynchronous protocols and robust fallback modes. In such settings, quantization-aware design must prioritize deterministic behavior and graceful degradation. These examples reinforce the argument that cryptographic inference is a socio-technical infrastructure, not merely a cryptographic protocol.

## **9. Research Agenda and Forward-Looking Perspectives**

A forward-looking research agenda should integrate quantization, cryptography, and governance from the beginning. One priority is the development of benchmarks that report accuracy, latency, communication, energy, robustness, and fairness under realistic threat models. Current evaluations often isolate one dimension, making it difficult to compare architectures. Another priority is compilers that can jointly optimize model architecture, precision allocation, and protocol selection. Such compilers must expose their decisions to auditors and support reproducible builds. They should also account for heterogeneous hardware and network conditions.

A second priority is the study of approximation-aware security. Quantization introduces numerical error, but the relationship between that error and information leakage is not well understood. Researchers should develop formal and empirical methods for measuring side channels, inference risks, and failure modes introduced by reduced precision. A third priority is privacy-preserving fairness auditing. Techniques are needed that allow regulators or third parties to verify equitable treatment without accessing raw data or proprietary models. These techniques must themselves be efficient and accountable.

Finally, policy and governance research should examine how quantization choices are made in practice. Who decides that a model can be reduced to lower precision? What evidence is required? How are affected communities involved? Standards bodies can provide guidance, but they must avoid freezing technology or creating compliance burdens that favor large firms. The required reference on activation reduction for two-party private inference [12] illustrates

how technical advances can change cost structures, but it also highlights the need for integrated evaluation. Lower cost is valuable only if it supports trustworthy, equitable, and sustainable systems.

## 10. Conclusion

Quantization-aware optimization is a powerful but double-edged strategy for cryptographic neural network inference. It can reduce communication, simplify nonlinear operations, enable integer arithmetic, and improve hardware utilization. Yet it also alters accuracy, robustness, fairness, and auditability in ways that cannot be ignored. This paper has argued that quantization should be treated as an architectural and governance concern, not merely a numerical compression step. Secure inference systems must co-design precision, protocol, hardware, deployment topology, and oversight. They must evaluate trade-offs across technical and social dimensions, including latency, energy, privacy, equity, and accountability. The path forward requires interdisciplinary research, transparent tooling, and policy frameworks that reward trustworthy optimization. Without such integration, quantization may accelerate private inference while quietly shifting risk to users and communities least able to contest it.

## References

1. Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming* (pp. 1-12). Springer.
2. Rivest, R. L., Adleman, L., & Dertouzos, M. L. (1978). On data banks and privacy homomorphisms. *Foundations of Secure Computation*, 4(11), 169-180.
3. Yao, A. C. (1982). Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science* (pp. 160-164). IEEE.
4. Goldreich, O., Micali, S., & Wigderson, A. (1987). How to play any mental game. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing* (pp. 218-229). ACM.
5. Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing* (pp. 169-178). ACM.
6. Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). GAZELLE: A low latency framework for secure neural network inference. In *27th USENIX Security Symposium* (pp. 1651-1669). USENIX.
7. Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017). Oblivious neural network predictions via MiniONN transformations. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 619-631). ACM.
8. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy* (pp. 19-38). IEEE.
9. Riazi, M. S., Weinert, C., Tkachenko, O., Songhori, E. M., Schneider, T., & Koushanfar, F. (2018). Chameleon: A hybrid secure computation framework for machine learning applications. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security* (pp. 707-721). ACM.
10. Mishra, P., Lehmkuhl, R., Srinivasan, A., Zheng, W., & Popa, R. A. (2020). Delphi: A cryptographic inference service for neural networks. In *29th USENIX Security Symposium* (pp. 2505-2522). USENIX.

11. Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., Ibrahim, M., & van der Maaten, L. (2021). CrypTen: Secure multi-party computation meets machine learning. In *Advances in Neural Information Processing Systems* 34.
12. Peng, H., Zhou, S., Luo, Y., Xu, N., Duan, S., Ran, R., ... & Ding, C. (2023). Rrnet: Towards relu-reduced neural network for two-party computation based private inference. *arXiv preprint arXiv:2302.02292*.
13. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*.
14. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2704-2713). IEEE.
15. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*.
16. Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12), 2295-2329.
17. Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* 30.
18. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 29.
19. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy* (pp. 399-414). IEEE.
20. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. <https://fairmlbook.org>