Latency vs. Privacy: A Theoretical Architecture for Pre-Warmed Trusted Execution Environments in Real-Time Bidding

Sivaramakrishnan Vaidyanathan

University of Cincinnati, USA

Abstract

Real-Time Bidding has a high set of temporal limitations since advertising platforms have to be fast and capable of responding within about 100 milliseconds to compete effectively in online auctions. With the development of privacy-first models, which require Trusted Execution Environments, there grows a basic system incompatibility, with standard instantiations of these secure computational models, where the size of the initialization sequences is orders of magnitude larger than the time required to run a single auction. The cold start phenomenon, including secure boot sequence and remote attestation protocol implementations, causes latencies that make on-demand provisioning incompatible with real-time bidding needs. The Pre-Warmed Enclave Pool architecture mitigates this problem by keeping a persistent collection of fully initialized and attested secure enclaves, practically separating the costly trust establishment steps in the context of a request and its important processing path. Projections based on mathematical modeling of the performance of recorded attributes of a commercial trusted execution platform show that this architectural style is capable of decreasing end-to-end processing latency to the point of achieving the required thresholds of auction participation. The resolution requires one to accept more infrastructure spending, which is a privacy tax, which is the costs that are necessitated by the need to over-provide resources to accommodate variable traffic patterns whilst maintaining sufficient service levels. According to this architectural blueprint, it is possible to combine privacy compliance with performance optimization, but in order to achieve both goals at the same cost, organizations must bear increased computational expenses to fulfill regulatory compliance and protect user information as the economic cost of regulatory compliance and user-data privacy.

Keywords: computational, commercial, protocol, attestation

1. INTRODUCTION

Real-Time Bidding is a distributed auction system in which the advertising systems are competing to have inventory space on a severely time-constrained basis. The response delivery in the OpenRTB specification is limited to narrow windows, usually limited to 100 milliseconds since the transmission of the initial request [1]. This limitation is because digital ad auctions occur within active sequence page loads, and going beyond this limit means the missed opportunity wherein inventory is directed to other respondents that are faster, or the slot remains empty. Introducing Trusted Execution Environments into such latency-critical infrastructure generates an architectural contradiction threatening programmatic advertising viability under emerging privacy frameworks.

Contemporary privacy frameworks, particularly Google's Protected Audience API with accompanying Bidding and Auction Services, require sensitive data and auction algorithms to operate within isolated computational spaces [2]. These environments deliver cryptographic assurances that code and information remain confidential from privileged administrators and infrastructure operators. Although dealing with regulatory requirements, this isolation creates significant amounts of computational overhead due to the mandatory security measures. The construction of trust requires loading special kernel images, setting up cryptographic objects, and running remote attestation chains with off-the-shelf key management infrastructure before the initiation of productive computation.

Traditional cloud designs dynamically allocate the resources, which are instantiated every time a request is received, and cease execution after the processing. This elastic solution is the most efficient at resources and optimizing costs, but is fundamentally incompatible with trusted environments, which need a lot of initialization. Early architectural examination shows that combined secure boot and attestation durations substantially surpass total time allocations for complete bid processing cycles, establishing a temporal impossibility blocking privacy technology adoption in performance-sensitive advertising systems.

The Pre-Warmed Enclave Pool architecture presented here decouples expensive trust establishment from critical request paths. Maintaining continuously refreshed inventories of initialized and attested secure enclaves converts the cold start obstacle from an insurmountable barrier into a manageable resource challenge. Mathematical modeling based on published performance data from commercial trusted platforms demonstrates how this pattern sufficiently reduces processing latency to satisfy real-time bidding demands while quantifying inevitable cost increases from maintaining idle computational reserves.

2. THEORETICAL FRAMEWORK AND PROBLEM DEFINITION

Implementation of trusted execution in real-time bidding requires extensive mathematical analysis of the overall system latency into sub-elements. The knowledge of the contribution of separate components allows determining the areas of optimization and deciding whether the offered solutions are able to meet the high-performance requirements set by digital advertising auctions. The OpenRTB protocol documentations show that exchanges set timeouts between 75 and 120 milliseconds based on market segments and quality requirements, with premium inventory clearing a longer processing window and remnant inventory being cleared with an aggressive limit [1]. These time constraints are highly realistic in the sense that, in practice, the performance of page loads directly affects the user experience and that advertising systems cannot tolerate time delays, which can be perceived.

Traditional serverless computing is efficient by providing on-demand precision in the instantiation of the resources and release them immediately after the completion. This pattern of provisioning is best when stateless request processing is required, where the cost of initialisation remains in single-digit milliseconds. The concept of trusted execution environments is fundamentally different since there is a lot of security-oriented initializations that occur before any workload is accepted. Research on secure distributed architectures shows that establishing cryptographic trust between isolated enclaves and remote attestation infrastructure introduces latency dominating overall processing timelines [3]. Secure boot sequences verify cryptographic signatures of loaded code, initialize protected memory with specific access controls, and generate attestation evidence proving enclave integrity to external validators.

Cold start latency models for on-demand trusted environments accumulate duration across sequential phases that resist parallelization or elimination without compromising security.The early stages are loading enclave image files with special operating system kernels and application code executing in restricted memory segments. This loading diverges from standard application startup because every component undergoes cryptographic verification before execution commences, preventing unauthorized code injection. On successful image verification, enclave kernels start runtime environments, which configure memory protections and open up communication channels with untrusted host systems. Attestation stages then produce cryptographic measures of loaded programs and at runtime states and send evidence to remote key management services and wait verification responses before enclaves are authorized to accept encrypted information.

Recent studies in confidential computing indicate that these initialization sequences take a long time even on existing hardware that is optimized to execute enclave operations [4]. The attestation workflows especially add much latency due to network round-trips to centralized key management infrastructure, adding layering on the external systems and network status. An insurmountable barrier shown by mathematical analysis, adding these components, is where the initialisation needs greater budgets than all the available request processing budgets. Adding execution time for bidding logic and network transmission delays produces aggregate latency several multiples beyond maximum acceptable thresholds, guaranteeing auction attempt timeouts and failures.

Component	Impact
Secure Boot (T_boot)	Eliminated from the critical path
Remote Attestation (T_attest)	Eliminated from the critical path

Inter-process Communication (T_vsock)	Minimal overhead introduced
Bidding Logic Execution (T_exec)	Unchanged between models
Network Transport (T_net)	Unchanged between models
Orchestrator Overhead (T_overhead)	Additional scheduling cost

Table 1: Latency Component Breakdown [3, 4]

3. PROPOSED ARCHITECTURE: THE PRE-WARMED ENCLAVE POOL

To make execution of trusted programs possible in systems with latency constraints, it is necessary to understand that the cost to initialize a program need not be immediately incurred when a request is being processed on the server, provided the cost is amortized over asynchronous background tasks. The Pre-Warmed Enclave Pool uses this idea by maintaining permanent stocks of secure enclaves that are fully initialised and attested and are waiting to accept incoming requests. This changes temporal issues of the impossible serialization of slow operations into capacity planning problems where enough resources exist to meet predicted workload curves without using up available pools during traffic surges.

The focus of architecture design is on orchestrator elements that run on unreliable parent instances that coordinate the lifecycle of multiple child enclaves. Orchestrators keep track of the pool depths, and a new enclave is instantiated when the inventory is less than the set thresholds, and incoming request is routed to available enclaves by using high-performance inter-process channels. Queue management theory gives mathematical tools for calculating the best pool sizes by assumptions of the expected arrival rates and processing time [5]. Orchestrators introduce disciplines based on first-in-first-out queuing, where enclaves are selected according to the order of completion of initialization, according to the available pools, so that attestation credentials are not too old and risks of enclave usage with obsolete cryptography material are minimized.

Lifecycle management is an important orchestrator task due to the inability of enclaves to survive forever without undermining the security assurance. Attestation cryptographic keys normally have expiration dates that restrict the validity times, and they have to be renewed after a specific time. Also, old enclaves build up state that may spill information across request boundaries when cleared incorrectly, forming side-channels in which the bidding strategy of one advertiser unwillingly affects another. Orchestrators deal with issues by setting up recycling policies that end with the disposal of enclaves once they have processed a set number of requests or once they have reached their limit on the number of years that they can operate. Retired enclaves are gracefully terminated and replaced with new, freshly initialized ones, doing the entire attestation processes, preserving pool capacity but with fresh cryptographic semantics.

Inter-process communication between untrusted orchestrators and trusted enclaves uses the special inter-process mechanisms that are optimized to operate in virtualized environments. Virtual machine performance research demonstrates traditional TCP/IP networking introduces unnecessary overhead when communication occurs between processes on identical physical hosts [6]. Virtual socket implementations bypass substantial network stack portions, providing memory-mapped channels that achieve microsecond-scale latency for small message transfers typical of bid requests. This communication efficiency proves essential because, although initialization costs are removed from critical paths, request routing and data transfer overhead between orchestrators and enclaves still contribute to overall response times and must be minimized to preserve latency budgets for actual bidding computations.

Warm start models enabled by this architecture reduce request processing latency to the sum of inter-process communication duration, execution time for bidding logic, and minimal orchestrator overhead for scheduling and serialization. Elements of initialisation that contribute to the dominance of cold start models are technically removed from the critical paths since they are asynchronous and take place before the arrival of requests. This is due to the fact that systems, with this transformation, can respond within the requirements of real-time bidding, with the overall latency being significantly more dependent on the effectiveness of the bidding algorithm and less related to the overhead of provisioning infrastructure.

Component	Function	Key Parameters	Design Consideration
Enclave Orchestrator	Manages lifecycle and routing	Pool depth monitoring, threshold triggers	Resides in an untrusted parent instance
FIFO Queue	Maintains available enclave inventory	Queue depth (N), arrival rate (λ)	Ensures cryptographic material freshness
Recycling Policy	Retires aged enclaves	Request count limit, maximum lifetime	Prevents side-channel information leakage
Virtual Socket (vsock)	Inter-process communication channel	Throughput 8-12 Gbps, latency <1ms	Bypasses TCP/IP stack overhead
Replenishment Thread	Background enclave initialization	Initialization rate, pool minimum	Maintains capacity during active processing
Request Router	Assigns requests to available enclaves	Scheduling algorithm, load balancing	Minimizes wait time and queue depth

Table 2: Enclave Pool Architecture Components [5, 6]

4. METHODOLOGY OF ANALYSIS

The analysis of Pre-Warmed Enclave Pool feasibility must be based on stringent analytical models that may be based on the established performance features of modern trusted execution platforms. It is a methodology that places more importance on mathematical models and less on empirical data since it allows one to test the feasibility and estimate costs until one decides to invest substantial engineering resources into the implementation process. The methodology is based on available technical specifications and published studies that define secure computing technology performance envelopes in real workload conditions.

Latency projection models synthesize data out of various authoritative sources, creating representative timing estimations on each architectural element. The technical documentation of Software Guard Extensions offered by Intel contains an in-depth examination of issues that influence enclave performance, such as computational overhead caused by the encryption of memory, latency caused by the context switch between trusted and untrusted code, and the time taken to initialize different enclave sizes and configurations [7]. These materials highlight the non-linear relationship between enclave initialization time and either the size of the protected code and data region, since larger memory allocations can only be satisfactorily measured with more extensive cryptography and increased attestation evidence generation. Documentation further indicates that attestation workflow deployment can bring in variable latency based on the state of the network and responsiveness of the remote key management service, and states that systems need to consider high variance in initialisation times rather than assuming optimistic minimum values.

In the modern world, distribution systems studies point to the critical significance of examining the performance percentiles at high levels instead of just looking at the median and average values [8]. Median latency is a measure of the experience of a typical request when operating at normal levels; however, in real-world systems, there is a lot of variance due to different factors such as garbage collection stalls, operating system scheduling choices, and temporary network congestion. In a service where latency is highly constrained, 99th percentile latency can be used as a practical parameter since even a small number of timeouts can have severe business consequences. In real-time bidding, advertisers whose systems are not working on one percent of the auction offers successfully give over that market portion to other advertisers and this is a significant and unacceptable level of campaign reach loss.

The approach of cost modeling takes into account internal inefficiency of having a pre-warmed computational capacity. In contrast to elastic serverless systems, which exactly match resource allocation to real time demand, pool architectures need to maintain enough enclave inventory to sustain peak traffic loads even when the utilization is at a low point. This over-provisioning is idle capacity during off-peak periods when idle enclaves use computational resources without working on productive workloads. Economic analysis is used to compute the premiums that such inefficiency entails by comparing real infrastructure costs to the imaginary minimums that would be obtained assuming that resources can be distributed with perfect elasticity and no initialization overhead.

The queue theory offers mathematical frameworks for establishing factors of over-provisioning needed to attain defined service level goals in fluctuating traffic patterns. Analysis models consider the arrivals of bid requests to be Poisson processes, which represent the random and independent process of web browsing of users in a large population [5]. Under the conditions of arrival processes which are Poisson-distributed and exponentially distributed service times, classical queuing formulae give the relationship between utilisation rates, queue length, and blocking probability of requests due to lack of enclaves. These equations point out the fact that the capacity buffers necessary to maintain very high service levels are large, and the amount of over-provisioning required to maintain a desired target level of availability is non-linear as target availability approaches perfection.

[redacted]

Table 3: Performance Projection Parameters [7, 8]

5. PROJECTED RESULTS AND DISCUSSION

An analytical approach to recommended architectures gives quantitative estimates that cast a light on the feasibility, as well as expected trade-offs, of implementing trusted execution environments for real-time bidding workloads. Cold start scenarios representing baseline on-demand enclave instantiation performance demonstrate why this approach cannot succeed under current technological constraints. Aggregating documented initialization times for secure boot sequences, remote attestation protocols, bidding algorithm execution, and network communication produces total latencies exceeding auction timeout thresholds by factors exceeding five. This quantitative result confirms the architectural impossibility of on-demand provisioning and establishes the necessity for alternative approaches, decoupling initialization from request processing.

Warm pool architectures transform impossible scenarios into viable systems by relocating initialization overhead to asynchronous background processes. Under this model, incoming bid requests encounter fully initialized enclaves requiring only inter-process communication and execution time to generate responses. Projected latency combines microsecond-scale virtual socket communication overhead, bidding algorithm execution duration, including machine learning inference operations, and minimal orchestrator scheduling overhead. The combination results in aggregate response times that are close yet not too high as defined by real-time bidding protocols, generating solutions where none existed before [9].

The analysis further indicates that there is a lot of variation in the projected performance, especially at high percentiles, where the tail latency effects are the dominant control on the system performance. Even the optimized services portrayed by distributed systems research show slowdowns occasionally due to reasons out of the control of the application [8]. In managed runtime systems, garbage collection can slow down tens of milliseconds, rearranging memory. The scheduling decisions made by the operating system may postpone the execution of a task when temporary allocations of physical cores of the CPU to other tasks occur. Network communication may temporarily congest, or when corrupted, it may need to retransmit packets. These are the reasons why latency variance is a contributing factor to high-percentile response times being significantly above median values, and tail latency may push above acceptable limits even though overall performance may be satisfactory.

Economic analysis measures infrastructure cost premiums of the main pre-warmed capacity to sustain pre-warmed capacity, as privacy taxes companies need to pay to gain regulatory compliance without compromising operational performance. Such cost calculation indicates inherent inefficiency of storing the computational resources that do not consume much energy at the time of low traffic and are in stand-by to serve peak loads. Premium magnitude depends directly on over-provisioning factors required to maintain acceptable service levels during traffic bursts, which in turn depends on arrival pattern statistical characteristics and organizational risk tolerance. European data protection regulations impose substantial penalties for privacy requirement non-compliance, including fines calculated as percentages of global annual revenue reaching tens of millions of euros for large enterprises [10]. When evaluated against these potential penalties, infrastructure cost premiums associated with trusted execution environments represent rational economic decisions trading increased operational expenditure for reduced regulatory and reputational risk.

Table 4: Cost Efficiency Comparison [9, 10]

CONCLUSION

The model proves that the existence of cold start penalties in Trusted Execution Environments is are blocking constraint in the adoption of Real-Time Bidding under traditional on-demand provisioning models. Quantified delay dissimilarity between secure enclave setup demands and the auction time constraints establishes failure rates that make regular deployment architecture infeasible in terms of accepting advertising technology workloads. The Pre-Warmed Enclave Pool design provides a mathematically feasible design by moving the overhead of the initialization to asynchronous background tasks, but keeping dynamically controlled pools of certified enclaves. Speculated latency decreases turn TEE-based bidding into a technically impossible to operationally feasible task, and the response times become acceptable for participating in an auction. The economic consequences are still big, and the premiums of infrastructure costs are the economic indications of inherent inefficiencies of keeping pre-warmed capacity over elastic on-demand allocation. In advertising systems where the amount of queries being processed is large, the annual expenditure growth due to an over-provisioning requirement is the direct financial impact of privacy compliance. The infrastructure premium is a logical risk reduction when compared to possible regulatory fines that can be imposed by modern data protection systems, especially the European regulations that could impose fines as a proportion of overall revenue. Companies that are planning privacy-enhanced advertising systems should make these cost structures part of financial planning models, as they understand that balancing privacy requirements with performance needs will require them to accept high baseline computational costs. The architectural roadmap confirms the existence of the two objectives of privacy protection and optimization of latency as compatible goals that can be achieved with a careful design of the system, but it must be remembered that they must be achieved at the same time, as the direct cost of regulatory compliance and by building trust among the users of digital advertising systems.

REFERENCES

[1] Interactive Advertising Bureau Technology Laboratory, "OpenRTB Version 2.6". [Online]. Available: https://iabtechlab.com/wp-content/uploads/2022/04/OpenRTB-2-6_FINAL.pdf

[2] Amazon Web Services, "AWS Nitro Enclaves," [Online]. Available: https://aws.amazon.com/ec2/nitro/nitro-enclaves/

[3] Stefan Brenner et al., "SecureKeeper: Confidential ZooKeeper using Intel SGX," Middleware' '16: Proceedings of the 17th International Middleware Conference, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2988336.2988350

[4] Weitong Ou et al., "A Survey on Bid Optimization in Real-Time Bidding Display Advertising," ACM Transactions on Knowledge Discovery from Data, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3628603

[5] Leonard Kleinrock, "Queueing Systems," New York: Wiley-Interscience. [Online]. Available: https://ia601403.us.archive.org/13/items/in.ernet.dli.2015.134547/2015.134547.Queueing-Systems-Volume-1-Theory.pdf

[6] Arun Raj Kaprakattu, "Performance Optimization in NUMA and Multi-Socket Virtual Machine Environments: A Technical Analysis," Journal of Computer Science and Technology Studies, 2025. [Online]. Available: https://www.researchgate.net/publication/392005740_Performance_Optimization_in_NUMA_and_Multi-Socket_Virtual_Machine_Environments_A_Technical_Analysis

[7] Intel Corporation, "Performance Considerations for Intel® Software Guard Extensions (Intel® SGX) Applications," 2023. [Online]. Available: https://www.intel.com/content/www/us/en/content-details/671502/performance-considerations-for-intel-software-guard-extensions-intel-sgx-applications.html

[8] Jeffrey Dean and Luiz André Barroso, "The tail at scale," Communications of the ACM, 2013. [Online]. Available: https://dl.acm.org/doi/10.1145/2408776.2408794

[9] "Bidding and Auction Services," Privacy Sandbox for Android. [Online]. Available: https://privacysandbox.google.com/private-advertising/protected-audience/android/bidding-and-auction-services

[10] European Union law, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons about the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)," EUR-Lex, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng