INFERENCE OPTIMIZATION

Accelerate

Reduce latency and cost with speculative decoding and custom kernels. 274x speedup potential for production workloads.

Inference is too slow and too expensive

LLM inference at scale means high latency and high costs. Users wait. Bills grow. And optimizing inference requires deep GPU expertise most teams don't have. Accelerate handles the optimization so you can focus on building.

274x	60%	0%
Speedup potential	Typical cost reduction	Quality degradation

Speculative decoding
Draft-then-verify for faster generation

Custom kernels
Hand-optimized Triton for max throughput

Quality preserved
Faster without degrading outputs

Drop-in integration
Minimal changes to your pipeline

BUILT ON

rotalabs-accel open source toolkit. Inspect methods, benchmark yourself, verify claims.

OPTIMIZES

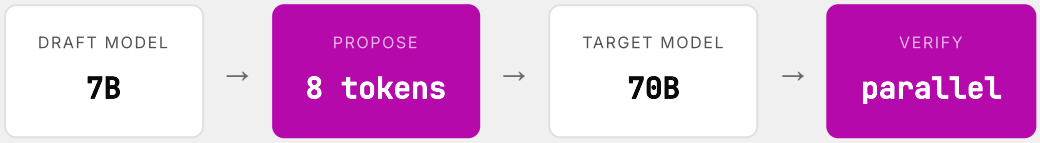
Llama, Mistral, custom fine-tuned models, any transformer architecture

01 - HOW SPECULATIVE DECODING WORKS

Draft-then-verify for dramatic speedups

Speculative decoding uses a small, fast draft model to propose tokens, then verifies them with your target model in parallel. Same outputs, dramatically faster.

Speculative Decoding Flow



Why It Works

Small models generate tokens fast. Large models verify multiple tokens simultaneously. When draft tokens match what the large model would have produced, you skip the slow generation entirely.

Mathematically Equivalent

Speculative decoding produces the exact same outputs as standard decoding. The verification step ensures no quality loss. Faster without compromise.

Custom Triton Kernels

We write low-level GPU kernels optimized for your specific workload. Attention, MLP, and memory operations tuned for maximum throughput on your hardware.

Workload-Specific Tuning

Draft model selection, speculation length, and kernel parameters tuned for your use case. What works for code generation differs from what works for chat.

VERIFIED SYNTHESIS

274x

Speedup for code generation with verification

CONVERSATIONAL

3-5x

Typical speedup for chat applications

Research Foundation

Speculative decoding is a well-established technique with strong theoretical guarantees. Our implementation builds on this research while adding production-grade optimizations.

Open source: rotalabs-accel toolkit available at rotalabs.ai. Benchmark yourself.

02 - ENGAGEMENT MODEL

From audit to optimization

We analyze your inference workload and implement optimizations tailored to your use case.

01

Audit

We profile your inference pipeline to identify bottlenecks. Detailed breakdown of compute, memory, and transfer costs. Clear understanding of optimization opportunities.

02

Optimize

Custom speculative decoding setup for your models. Hand-tuned Triton kernels for your hardware. Draft model selection and speculation length optimization.

03

Validate

Rigorous quality testing to ensure no regression. We prove that outputs are equivalent to baseline. Statistical validation across your test suite.

04

Deploy

Integration into your production pipeline with minimal code changes. Monitoring dashboards for performance tracking. Documentation and knowledge transfer.

03 - USE CASES

Where Accelerate delivers value

CODE GENERATION

Verified Synthesis

274x speedup for code generation with verification. Draft code, verify with execution. Dramatically faster while maintaining correctness guarantees.

REAL-TIME CHAT

Conversational AI

3-5x latency reduction for chatbots and assistants. Users notice the difference. First token and total response time dramatically improved.

BATCH PROCESSING

High-Volume Inference

Process documents, analyze data, generate content at scale. Same GPU budget, 3x more throughput. Or same throughput, 60% lower cost.

EDGE DEPLOYMENT

Resource-Constrained

Run larger models on smaller hardware. Memory and compute optimizations that make deployment feasible where it wasn't before.

04 - TECHNICAL SPECIFICATIONS

What we optimize

CAPABILITY	SPECIFICATION
Speculative Decoding	Draft model selection, speculation length tuning, acceptance rate optimization
Custom Kernels	Triton-based attention, MLP, and memory kernels optimized for your hardware
Supported Models	Llama, Mistral, custom fine-tuned models, any transformer architecture
Hardware Support	NVIDIA A100, H100, L40S, RTX 4090; AMD MI250, MI300
Quality Validation	Statistical equivalence testing, regression detection, output comparison
Monitoring	Latency dashboards, throughput metrics, acceptance rate tracking
Integration	Python SDK, vLLM integration, custom inference servers

05 - ENGAGEMENT OPTIONS

Work with us

Audit

$10K

Comprehensive profiling and recommendations. Understand your optimization potential before committing.

Optimization

$50K

4-week implementation. Custom speculative decoding and kernel optimization for your workload.

Retainer

$5K/mo

Ongoing optimization as your models and workloads evolve. Continuous improvement.

Pricing is indicative. Contact us for custom requirements and volume engagements.

Start with an audit

Understand your optimization potential before committing to implementation.

Request Audit

Learn More