Ground truth Circuit-breaker: Leveraging Surrogate Models to Achieve Graceful Degradation in Distributed Systems

Benjamin Murray
School of Enterprise Computing and Digital Transformation, TU Dublin, Ireland
X00218726@myTUDublin.ie

Introduction

Modern distributed systems increasingly rely on resilience mechanisms to cope with partial failures, overload, and unpredictable runtime conditions. Traditional fault-tolerance patterns, such as retries and circuit breakers, typically fail fast or degrade functionality when backend services become unavailable. This thesis explores a novel resilience approach, the Ground Truth Circuit Breaker, in which lightweight machine learning models are positioned as middleware in front of microservices to generate approximate responses when downstream services are unavailable or under extreme load. The research investigates the feasibility, performance characteristics, and architectural implications of replacing live service calls with ML-generated responses at runtime. The work evaluates this pattern as an extension to existing resilience strategies, with particular focus on system availability, latency, and applicability in high-load scenarios.

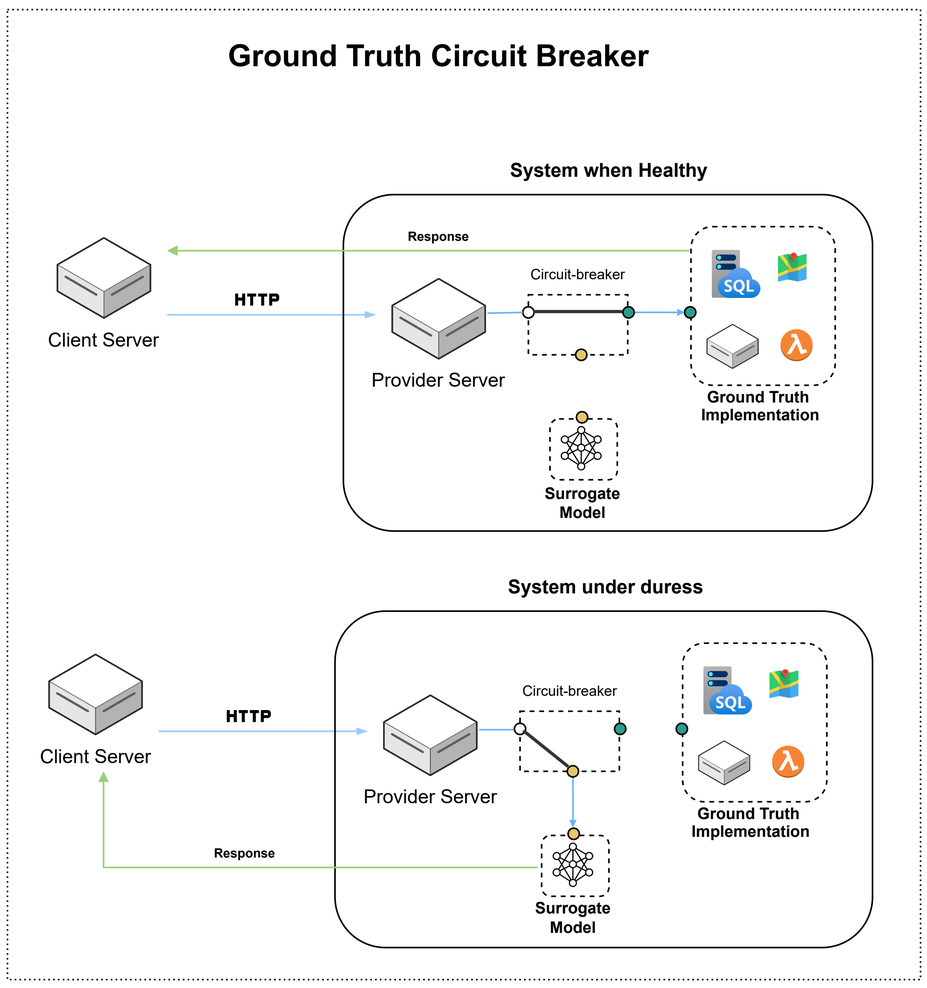
TOPSIS Analysis

RQ1: Can the *Groundtruth Circuit Breaker*, which uses a surrogate model to approximate downstream service behaviour, improve system resilience under partial or total service degradation?

This primary question is supported by the following subsidiary questions:

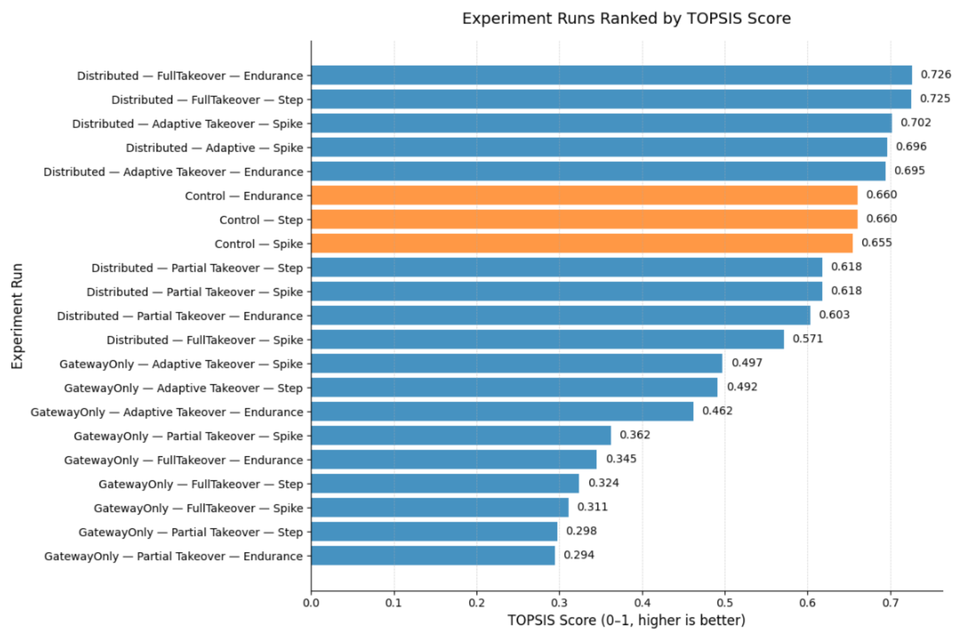
- **RQ1.1:** How accurately can a surrogate model approximate the outputs of a real service under normal operating conditions?
- **RQ1.2:** How does system behaviour vary under different traffic patterns, surrogate placements, and takeover policies?
- **RQ1.3:** How does surrogate-based fallback compare to traditional resilience mechanisms in terms of latency, error rates, and stability?
- **RQ1.4:** What limitations, risks, and trade-offs arise when using surrogate models for degraded response generation?

A New Resilience Pattern



When the system is healthy, surrogate models remain dormant while continuously learning from live production traffic. Under duress, the circuit trips and traffic is fully or partially routed to these models until normal operating conditions are restored. The pattern introduces a controlled trade-off, sacrificing some response fidelity in exchange for lower latency and higher throughput.

Results



Analysis

A weighted TOPSIS analysis showed that multiple configurations of the Groundtruth Circuit Breaker pattern achieved a favourable trade-off between performance and accuracy. The decision model prioritised output fidelity ($\times 3$) over throughput ($\times 1.2$), p95 latency ($\times 1.4$), and CPU utilisation ($\times 1.01$), reflecting the importance of response correctness under degradation. Under this weighting, several surrogate-based configurations not only reduced end-to-end latency but also outperformed the control baselines once their strict 100% ground-truth fidelity requirement was considered, demonstrating measurable resilience benefits with bounded approximation error.

Conclusions and Future Work

This thesis demonstrated that the Groundtruth Circuit Breaker, which uses surrogate models to generate approximate responses during service degradation, can significantly reduce latency and increase throughput in distributed microservice systems. The results show that effectiveness depends on architectural and operational choices, with service-local surrogate placement and adaptive takeover strategies providing the best balance between performance and fidelity, achieving accuracy levels of approximately 85–90%. While surrogate outputs do not fully match ground-truth responses, the resulting trade-off is acceptable in latency-sensitive systems that can tolerate reduced response fidelity. Future work includes extending the pattern to tolerate network-level failures and exploring generative model types to support complex, high-dimensional response formats.

QR Code for Recording