The Mystery Machine: End-to-end performance analysis of large-scale Internet services

Michael Chow
David Meisner, Jason Flinn, Daniel Peek, Thomas F. Wenisch
Internet services are complex

Scale and heterogeneity make Internet services complex
Internet services are complex

Scale and heterogeneity make Internet services complex
Step 1: Identify segments
Step 2: Infer causal model
Step 3: Analyze individual requests
Step 4: Aggregate results
Challenges

- Previous methods derive a causal model
  - Instrument scheduler and communication
  - Build model through human knowledge

Need method that works at scale with heterogeneous components
Opportunities

• Component-level logging is ubiquitous
  
  Tremendous detail about a request’s execution

• Handle a large number of requests
  
  Coverage of a large range of behaviors
1) Infer causal model from large corpus of traces
   – Identify segments
   – Hypothesize all possible causal relationships
   – Reject hypotheses with observed counterexamples

2) Analysis
   – Critical path, slack, anomaly detection, what-if
Step 1: Identify segments
Define a minimal schema

Task

Segment 1

Segment 2

Event

Event

Event
Define a minimal schema

Aggregate existing logs using minimal schema
Step 2: Infer causal model
## Types of causal relationships

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Counterexample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Happens-Before</strong></td>
<td></td>
</tr>
<tr>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Mutual Exclusion</strong></td>
<td></td>
</tr>
<tr>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Pipeline</strong></td>
<td></td>
</tr>
<tr>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
</tr>
</tbody>
</table>

Michael Chow
Producing causal model

Causal Model

S1 → N1 → C1
S2 → N2 → C2

Michael Chow
Producing causal model

Causal Model

Trace 1

Time
Producing causal model

Causal Model

Trace 1

Time
Producing causal model

Causal Model

Trace 2

Time

S1 → N1 → C1
S2 → N2 → C2

S1 → N1
S2 → N2

C1 → N2
C2 → N1

Michael Chow
Producing causal model

Causal Model

Trace 2

S1 N1 C1
S2 N2 C2

Time
Producing causal model

Causal Model

Trace 2

Time
Step 3: Analyze individual requests
Critical path using causal model

Trace 1

S1 → N1 → C1 → S2 → N2 → C2

Time
Critical path using causal model

Trace 1
Critical path using causal model

Trace 1
Step 4: Aggregate results
Inaccuracies of Naïve Aggregation
Inaccuracies of Naïve Aggregation

Michael Chow

![Diagram showing percent of end-to-end latency for Summed Delay and Critical Path, with categories for Javascript, DOM, CSS, Network, and Server]
Inaccuracies of Naïve Aggregation

Need a causal model to correctly understand latency
High variance in critical path

- Breakdown in critical path shifts drastically
  - Server, network, or client can dominate latency
High variance in critical path

- Breakdown in critical path shifts drastically
  - Server, network, or client can dominate latency

20% of requests, server contributes 10% of latency
High variance in critical path

- Breakdown in critical path shifts drastically
  - Server, network, or client can dominate latency

20% of requests, server contributes 50% or more of latency
Diverse clients and networks
Diverse clients and networks
Diverse clients and networks
Differentiated service

Slack in server generation time
- Produce data slower
- End-to-end latency stays same

No slack in server generation time
- Produce data faster
- Decrease end-to-end latency

Deliver data when needed and reduce average response time
Additional analysis techniques

• Slack analysis

• What-if analysis
  – Use natural variation in large data set
What-if questions

• Does server generation time affect end-to-end latency?
• Can we predict which connections exhibit server slack?
Server slack analysis

Slack < 25ms

Slack > 2.5s
Server slack analysis

Slack < 25ms

End-to-end latency increases as server generation time increases

Slack > 2.5s

Server generation time has little effect on end-to-end latency
Predicting server slack

- Predict slack at the receipt of a request
- Past slack is representative of future slack

![Graph showing the relationship between first and second slack](image-url)

Michael Chow
Predicting server slack

- Predict slack at the receipt of a request
- Past slack is representative of future slack

Type II Error
- 9%

Type I Error
- 8%

Classifies 83% of requests correctly

Michael Chow
Questions