Using Influence to Understand Complex Systems

Adam Oliner
Alex Aiken
Complex Systems
Complex Systems

- Many heterogeneous components
- Noisy, incomplete instrumentation
Goal

To understand the behavior of complex systems so we can diagnose bugs, optimize performance, and build better systems.
Example #1

- Thunderbird supercomputer
- `kernel: Losing some ticks...`
  checking if CPU frequency changed
- Alerts were correlated with job scheduling groups
- Kernel was skipping interrupts under heavy network activity
Example #2

- Stanley (autonomous vehicle)
- Swerved around non-existent obstacles
- Swerving preceded by unusual laser sensor behavior
- Stale data resulted in hallucinated obstacles
Hard Constraints

• Cannot modify or perturb
• Cannot assume source code access
• Cannot assume correctness predicates or specifications
• Cannot require a specific log format

Raw Data

- Component i
  
  kernel: Uhuh. NMI received. Dazed and confused, but trying to continue

  network: Shutting down loopback interface: succeeded

- Component j

  0.051480 -0.011220 -0.967890 -0.018017 0.011043
  -0.096482 1128770317.424981 xyz 0.047278

  0.048510 -0.007590 -0.962610 -0.020342 0.009299
  -0.095897 1128770317.435014 xyz 0.057273

2011/04/012
Anomaly Signals

- Surprise over time: $\Lambda_j(t)$
- Compare heterogeneous components
- Encode knowledge
Some Signals

How do we compute and represent these patterns?
Cross-Correlation

\[(\Lambda_i \ast \Lambda_j)(t) \equiv \int_{-\infty}^{\infty} \frac{[\Lambda_i(\tau) - \mu_i][\Lambda_j(t + \tau) - \mu_j]}{\sigma_i \sigma_j} d\tau\]
Cross-Correlation

\[ i \rightarrow j \]

![Graph showing cross-correlation with correlation coefficient and delay axis. The graph displays peaks and troughs indicating the correlation between two signals at different delays.](image)

- \( \alpha \)
- \( -\alpha \)
- \( 0 \)
- \( \varepsilon \)
Influence

Two components share an influence if they tend to exhibit surprising behavior around the same time.

Influence Graph

- Adjacent components share an influence
- \texttt{fan7} precedes \texttt{disk32}
- \texttt{fan7} and \texttt{fan6} are \texttt{\simultaneous}

![Diagram of Influence Graph]

\texttt{fan7} \quad \texttt{fan6} \quad \texttt{disk32} \quad \texttt{job failures}
A Typical Question

“The nodes in rack 5 are performing poorly. Is there a pattern to the misbehavior? Does it have anything to do with the disk errors?”
A Problem

- Too many cross-correlations
  - $O(n^2)$ for $n$ components

- Use query to constrain the problem
  - Compute only a subset of the pairs

Influence Matrix

- Each entry contains correlation & delay
- Each cross-correlation fills 2 entries
Scaling: Focal Set

- Computes all pairs in set
Scaling: Peripheral Set

- Each focal to each peripheral component
Synthetic Components

- Metacomponents
  - “All the nodes in a rack”
- Binary components
  - “When the log contains ‘ERR’”
- Masked components
  - “Ignore the morning reboot”
Example Query

top=5 periph=disk ^r5

• Components matching ^r5 are in focus
• Disks are the periphery
• Plot only the 5 strongest influences
Example Query

top=5 periph=disk ^r5

^r5} disk
Anomaly Signals

- Text logs
  - Surprise = unusual term distribution
- Sensor logs
  - Surprise = unusual timing distribution

Isolating a Problem

- Job failures on Blue Gene/L (BG/L)
  - Production supercomputer at LLNL
  - 131,072 processors
- Full-system jobs: everything is suspect
- Clue: time of the crash
Isolating a Problem

\texttt{top=1 periph=meta CRASH<-174:175} \\
(1.65 seconds)

Rack 3 shares the strongest influence: Recurse!
Isolating a Problem

$$\text{top=5 periph=R03 CRASH}$$

(4.81 seconds)

Most suspicious component out of more than 131,000
Isolating a Problem

• Suspect node generated two messages
• One explains the crash:

\texttt{ddr: Unable to steer [...] consider replacing the card}

• Ninety other components generated hundreds of messages (noise)
Stanley Swerving
What Happened?

1. Bump
2. Stall
3. Evade

(IMU) + Blocked = Evaded

(Lasers)  (Decision)
The Bug

- Implicit timing dependency
- Nondeterministic
- Two month search
Why So Elusive?
Isolating a Problem 2

periph=all SWERVE<-60:100
(20.37 seconds)

Several components seem to share an influence with the swerving; how are they related to each other?
Isolating a Problem 2

last
(35.60 seconds)

Points directly at a likely cause; sufficient to isolate the problem.
Online System

- Easier to deploy in a production system
- Fewer resource requirements
- Delayed influence $\rightarrow$ prediction
Key Challenge

- Compression sufficient to compute the cross-correlations in real time
- Idea: Summarize multiple signals as a single signal
- Solution scales to 100,000+ signals
SQL Cluster

- 9-machine SQL database cluster
- Multiple numerical metrics per machine
- Used directly as anomaly signals
SQL Cluster Bug

- Bug: some machines would occasionally slow down and hang
- Clue: “The main time when we had the problem [was]: [Tuesday] 6:30:00 PM to [Wednesday] 12:22:00 AM [and] another incident the following morning (10am or so, I think)”
Online Analysis

Is this information actionable?

DISK/FORKS → SWAP → CRASH

~30 mins  ~210 mins
Case Study Summary

• Real, unmodified production systems
• No expert knowledge
• Found tough bugs, revealed interesting interactions, admins finding it useful
• Fast enough to be interactive
Conclusion

- View systems as producing signals
- Analyze via large-scale signal processing
- A new class of tools for understanding complex systems
Fin

You have reached the end of the presentation. Please turn back.