Hound: Causal Learning for Datacenter-Scale Straggler Diagnosis

Benjamin C. Lee

P. Zheng and B.C. Lee. Proc. of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS), June 2018.

Stragglers in Datacenter Computation

Task Parallelism

Split jobs into parallel tasks

Aggregate task results

Stragglers

Exhibit atypically poor performance

Delay job completion

Example

Extend completion time by 50%
in 20% of Google jobs

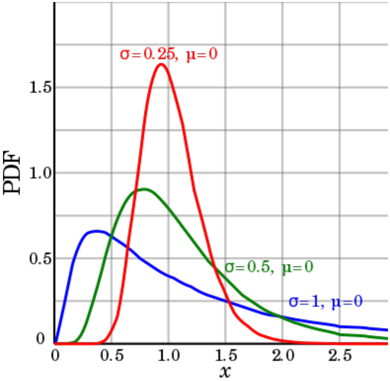


Image: "Log-normal distribution" [Wikipedia

Mitigating Stragglers

Speculative Scheduling

Clone tasks on different machine

Avoid machines predicted to underperform

Inefficient Clones

Consume resources inefficiently

E.g., Data skew across tasks

Causal Diagnoses

Rely on expertise, domain knowledge

Fail to scale, laborious

Dean and Barroso. "The Tail at Scale" **[CACM'13]**

Machine Learning for Diagnosis

Profile Datacenters

System monitors track task, job events

Hardware counters track microarchitectural activity

Reveal Stragglers' Causes

Allocation and scheduling

Colocation and interference

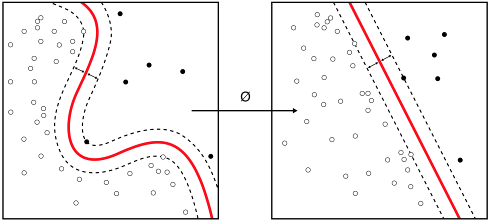


Image: "Statistical classification" **[Wikipedia]**

Desiderata from Machine Learning

Datacenter-Scale Insight

Extract patterns across jobs' disparate models

Interpretable Models

Codify domain expertise, interpretable insight

Unbiased Inference

Reduce risks of false causal explanations

Computational Efficiency

Design models with scalable implementation

Hound Framework

1. Base Learning for Jobs

Associate performance with system conditions

2. Meta Learning for Datacenter

Discover recurring, interpretable causes at scale

3. Ensemble Learning

Reconcile results from independent learners

Hound Framework

1. Base Learning for Jobs

Associate performance with system conditions

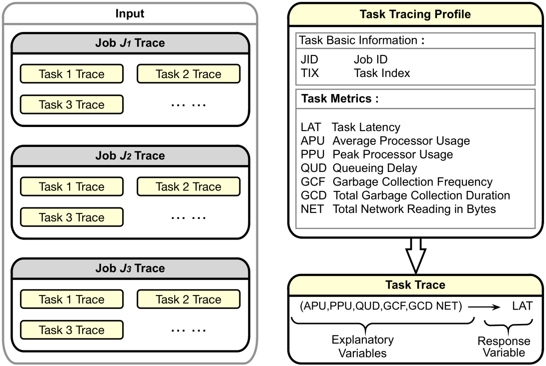
2. Meta Learning for Datacenter

Discover recurring, interpretable causes at scale

3. Ensemble Learning

Reconcile results from independent learners

Base Learning



Dataset

Task profiles in job

Response

Task latency

Predictor

Profiled metrics

Models

Logistic regression

Dependence models

Rubin causal models

Rubin Causal Models

Confounding Bias

Arises when association between two variables explained by third variable

Example

Latency is higher on older processors, but slower memory is confounding

Rubin Causal Models

Confounding Bias

Arises when association between two variables explained by third variable

Example

Latency is higher on older processors, but slower memory is confounding

Rubin Causal Model

Estimates effect of $Z \in \{0, 1\}$ on R while controlling for X

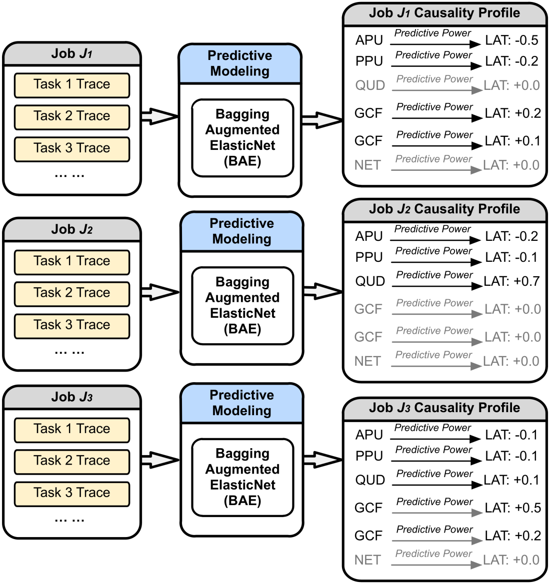
$$\mathbb{E}\left[\frac{ZR}{e(X)}\right] - \mathbb{E}\left[\frac{(1-Z)R}{1-e(X)}\right]$$

$$e(X) = P\{Z = 1|X\}$$

Observe Z and R from data

Estimate $e(X)$

Causality Profiles



Infer relationship between metrics, job time

Scale to hundreds of metrics, millions of jobs

Hound Framework

1. Base Learning for Jobs

Associate performance with system conditions

2. Meta Learning for Datacenter

Discover recurring, interpretable causes at scale

3. Ensemble Learning

Combine results from independent learners

Meta Learning

Words

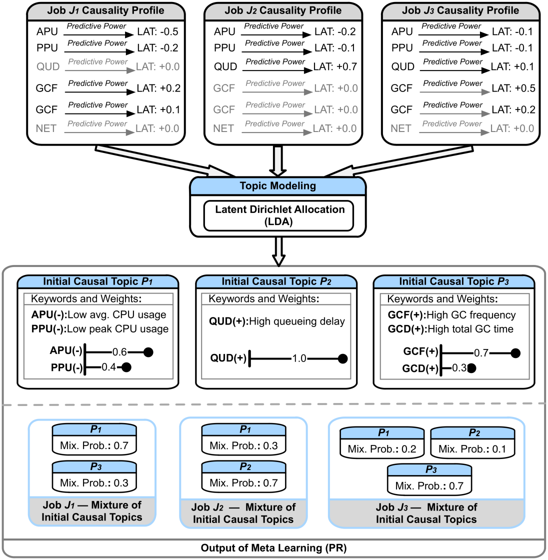
Metrics (+), (-) indicate atypically high, low values

Topics

Recurring word clusters from causality profiles

Diagnoses

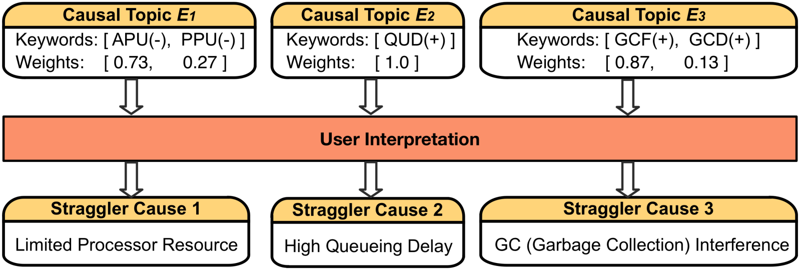
Assign topic mix to jobs



Interpretable Diagnoses

Topic reveals keywords, weights

System architect interprets cause



Hound Framework

1. Base Learning for Jobs

Associate performance with system conditions

2. Meta Learning for Datacenter

Discover recurring, interpretable causes at scale

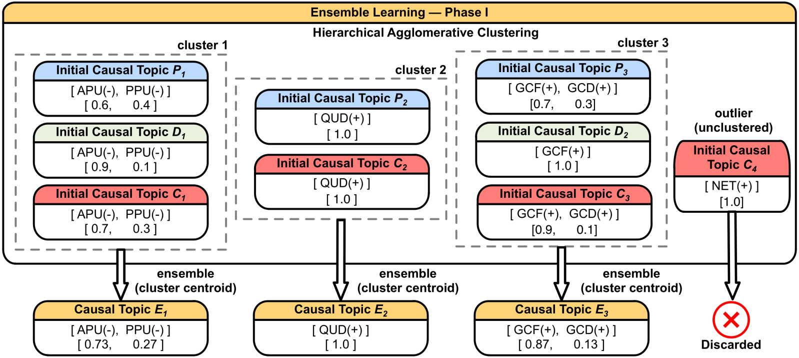
3. Ensemble Learning

Reconcile results from independent learners

Ensemble Learning

Construct learners for prediction (P), dependence (D), causation (C)

Drop topics found by one learner



Hound Framework

1. Base Learning for Jobs

Associate performance with system conditions

2. Meta Learning for Datacenter

Discover recurring, interpretable causes at scale

3. Ensemble Learning

Reconcile results from independent learners

Topics from Google Datacenter

29-day trace of production system

12K servers for 13K jobs, 3.3M tasks

Topic	Keywords	Weights	Cluster	Interpretation
E_0	MEM_ASSIGN(+), MEAN_MEM(+), PEAK_MEM(+)	0.5, 0.25, 0.25	P_0, P_3, D_0, C_0	Data Skew
E_1	PAGE_CACHE(+), PAGE_CACHE_UM(+), MEM_ASSIGN(+)	0.45, 0.38, 0.17	P_1, D_1, C_1	Data Skew
E_2	DISK_SPACE(+)	1.0	P_2, D_2, C_2	Data Skew
E_3	MEAN_CPU(+), PEAK_CPU(+)	0.52, 0.48	P_4, D_3, C_3	Computation Skew
E_4	PEAK_IO(+), MEAN_IO(+)	0.51, 0.49	P_5, D_4, C_4	I/O Skew
E_5	MEAN_CPU(-), PEAK_CPU(-)	0.8, 0.2	P_6, D_5, C_5, C_6	Limited Processor
E_6	MEAN_MEM(-), PEAK_MEM(-)	0.83, 0.17	P_7, D_6, D_7, C_7	Limited Memory
E_7	MEAN_IO(-)	1.0	D_8, C_8	Limited I/O
E_8	PEAK_IO(-), MEAN_IO(-)	0.83, 0.17	P_8, D_9	Limited I/O
E_9	CACHE_MISS(+), CPI(+)	0.54, 0.46	P_9, D_{10}, C_9	Cache Bottleneck
E_{10}	SCHED_DELAY(+)	1.0	P_{10}, D_{11}, C_{10}	Scheduler (Queueing) Delay
E_{11}	EVICT(+)	1.0	P_{11}, C_{11}	Eviction Delay
P_{12}	FAIL(+)	1.0	unclustered	✖
C_{12}	MACHINE_RAM(+)	1.0	unclustered	✖

Reiss and Wilkes. "Google cluster-usage traces: format+ schema" **[TR'11]**

Causal Coverage

(Dominant) Coverage

Measures how often cause explains (majority of) stragglers

Differentiates major, minor diagnoses

Cause	Coverage	Dominant Coverage
Data Skew (E_0, E_1, E_2)	73.6%	55.0%
Limited Processor (E_5)	39.2%	12.1%
Cache Misses (E_9)	32.6%	7.0%
Limited I/O (E_7, E_8)	36.7%	6.6%
Queueing Delay (E_{10})	20.0%	5.1%
Limited Memory (E_6)	13.6%	2.7%
Computation Skew (E_3)	31.2%	2.2%
Eviction Delay (E_{11})	3.80%	0.90%
I/O Skew (E_4)	5.60%	0.60%

Computational Efficiency

Complexity is O(NM)

N is number of jobs

M is number of tasks per job

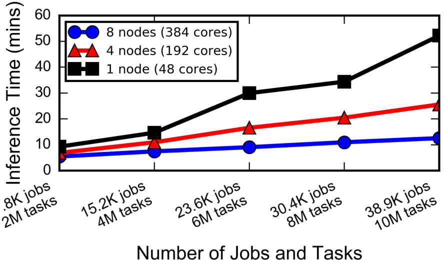
Implementation

Apache PySpark

Spark cluster with eight nodes

Parallel Analysis

40K jobs, 10M tasks



Also in the paper...

Modeling Methods

Prediction – ElasticNet with Bagging

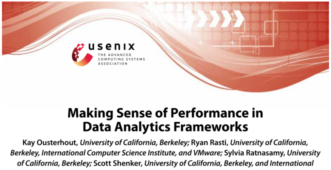
Dependence – Signed Schweizer-Wolff

Causation – Inverse Probability Weighting with AdaBoost

Evaluation and Validation

Visualizing case studies for Google

Comparing to domain expertise from Berkeley



usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

Making Sense of Performance in Data Analytics Frameworks

Kay Ousterhout, *University of California, Berkeley;* Ryan Rasti, *University of California, Berkeley, International Computer Science Institute, and VMware;* Sylvia Ratnasamy, *University of California, Berkeley;* Scott Shenker, *University of California, Berkeley, and International*

Hound: Causal Learning for Datacenter-Scale Straggler Diagnosis

Benjamin C. Lee

P. Zheng and B.C. Lee. Proc. of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS), June 2018.