ABSTRACT

Current distributed systems carry legacy subsystems lacking sufficient instrumentation for monitoring the end-to-end business transactions supported by these systems. Retrofitting these systems with expensive monitoring instrumentation provides high-granularity, precise tracking of transactions, while in the absence of instrumentation, only probabilistic monitoring is possible. Given a limited budget, instrumentation strategies which maximize the effectiveness of monitoring transactions throughout the system are proposed. The operation of the end-to-end system is modeled by a queuing network; each queue represents a subsystem which produces time-stamped log-records as transactions pass through it. Two simple heuristics for instrumentation are proposed which become optimal in different regimes. One heuristic selects transition links (between queues in the queuing model) for instrumentation in the decreasing order of the load factors of transactions they support. Sufficient conditions for this load-factor heuristic to be optimal are proven using the notion of stochastic order. The other heuristic selects transition links based on the approximated tracking accuracy of probabilistic monitoring is shown to be tight at low arrival rates.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Distributed systems—Performance evaluation

General Terms
Performance, Measurements, Theory

Keywords
Probabilistic transaction monitoring, Queuing networks, Stochastic comparison, Bipartite matching.

1. INTRODUCTION

Many enterprise systems in operation have evolved over a long period of time. They started out as small ventures and underwent rapid unplanned expansion in the face of increasing demand and business consolidations. New systems and applications were installed and old systems were never fully upgraded, leading to a ‘lump’ of heterogeneous systems with varied OS and hardware configurations. Complete documentation of these end-to-end systems are hardly available. Troubleshooting errors are usually long-drawn and handled on a case-by-case basis.

Instead, errors can be handled in a more organized fashion by continuously monitoring transactions supported by these systems. By tracking the progress of transactions in different subsystems, we can detect anomalies and sluggish performance. However, in the absence of sufficient instrumentation, e.g., according to the ARM guidelines [1], precise and effective monitoring is hard to come by. Instead, systems may only output limited information in the form of log-records (henceforth referred to as transaction footprints), which usually contain a local identifier for each transaction instance in that subsystem. Unfortunately, since the subsystems have gone through independent development, there is no guarantee that the identifiers of footprints at different subsystems can be uniquely matched with each other to obtain a globally unique trace of each transaction instance. It thus becomes important to find the most-likely footprints of each transaction instance as it travels through the system without relying on a unique global identifier.

An alternative to having a system-wide unique transaction identifier is to provide link-by-link transaction association by instrumenting the consecutive states (processes or operations) along which the transactions progress. This permits transaction footprints generated by one application to be associated the footprints generated by the subsequent application. If all links are instrumented, then starting with the entry point of a transaction into the system, transaction footprints can be located in all the applications traversed by the transaction. However, this is an expensive proposition, and when the instrumentation budget cannot cover all the links, we need to decide which links to instrument. What follows below is a pursuit of easy to use guidelines for spending link instrumentation budget.

2. SUMMARY OF CONTRIBUTIONS

We model the distributed system using a state-transition diagram, where each state represents a subsystem or an application traversed by the transactions of interest. See Fig.1. Each state is modeled as an infinite-server queue /G/∞ with semi-Markovian service times [3]. We assume Poisson arrivals into the system at the unique start state with rate \( \lambda \).
For every transaction that passes through a state, a transaction footprint consisting of the entry timestamp record is available. Crucially, the footprint does not contain the identity of the transaction generating it due to lack of monitoring instrumentation.

In the absence of any monitoring instrumentation, only probabilistic transaction tracking is possible using the footprint records. The difficulty in matching footprints belonging to the same transaction in two consecutive states comes from uncertainties introduced by out-of-order progress of transactions from one state to another. In particular, the optimal maximum-likelihood (ML) and the simple sub-optimal first-in first-out (FIFO) algorithms for matching transactions footprints have been suggested and their properties and trade-offs have been discussed in [2]. These automated probabilistic matching algorithms yield a correct match with a certain probability along each transition link.

On the other hand, after the instrumentation of a particular transition link, transaction footprints in the two linked states can be matched perfectly since instrumentation produces footprints with unique consistent transaction identifiers for the pair of states along the selected transition link. However, there are monetary and human capital costs associated with retrofitting instrumentation to each transition link. Thus, we have an optimization problem where within a given budget constraint, we need to select transition links for instrumentation such that the total accuracy of matching transactions at all the links is maximized.

We propose two heuristics for instrumentation which are optimal under some conditions. One of them is based on the load factors of queues along different transition links; the load factor of a queue is the ratio of the average arrival rate to the average service rate. The other heuristic is based on an approximate expression of monitoring accuracy at each transition link and is tight at low arrival rates.

The load-factor heuristic is based on the intuition that a higher load factor for an infinite-server queue leads to a higher number of transactions being simultaneously serviced, and hence, introduces a higher uncertainty to transaction tracking. Therefore, a simple heuristic is to instrument links in the reverse order of their load factors. But the service times at each queue could follow different distributions and even belong to different distribution families. For example, if one link has a deterministic service time with a high load factor, then the load-factor heuristic makes a wrong prediction since we can track transactions precisely with deterministic service times.

One of the main contributions of the paper is to provide sufficient conditions for the optimality of the load-factor heuristic under different service distributions and automated algorithms for matching log-records. In particular, we consider two simple matching rules, viz., first-in first-out (FIFO) and random matching. Under FIFO matching if the service times, normalized by their arrival rates, and also their “spreads” follow the same stochastic order, then the load-factor heuristic is optimal. Under random matching, a stochastic order on the normalized service times along with a mild condition on their support is sufficient to establish the optimality of the load-factor heuristic. For the special case when all the service distributions belong to the same family, the load-factor heuristic is guaranteed to be optimal, and no additional conditions are needed.

We propose another heuristic based on approximate expressions of accuracy of the probabilistic matching algorithms. In general, computing the exact expressions of matching accuracy is NP-hard and our approximations are tight in the regime of low arrival rates. Intuitively at low arrival rates, over all possible busy period queue sizes, the dominant event is having busy period of size unity. Hence, we approximate tracking accuracy by the probability of having a unit-sized busy period, which is simple to evaluate. We then propose instrumentation strategies based on this approximation.

2.1 Problem Formulation

Let $\gamma$ be the particular matching policy employed at link $e$, e.g., FIFO, random match and so on. Let $P_0^\gamma(e)$ be the probability of correctly matching all footprints generated by the states on both sides of link $e$ that belong to the same busy period of size $B$, and let $E[P_0^\gamma(e)]$ denote its expectation over the arrival and service distributions at link $e$ in steady state.

One possible selective instrumentation strategy is to maximize the total expected matching accuracies along different transition links in the transaction model. Note that matching footprints along a link is precise and fail-proof after it has been instrumented. For each link $e$, let $\pi_e \in \{0,1\}$ be the indicator if link $e$ is selected for instrumentation. Then, the monitoring accuracy at link $e$ after instrumentation is given by

$$E[P_0^{\text{EFF}}(e); \pi_e] = \pi_e + (1 - \pi_e)E[P_0^\gamma(e)]$$

(1)

since monitoring accuracy is unity after instrumentation and $E[P_0^{\text{EFF}}(e)]$ is the accuracy using only footprints.

Therefore, given a set of initially uninstrumented transition links $E$ in the state diagram, a total budget of $E$ for instrumentation, and costs $C_e$ to instrument the link $e \in E$, we have the following 0-1 integer program

$$\begin{aligned}
\pi_e \in [E; \mathcal{E}] := \arg \max_{\pi_e} & \sum_{e \in \mathcal{E}} E[P_0^{\text{EFF}}(e); \pi_e], \\
\text{s.t.} & \sum_{e \in \mathcal{E}} \pi_e C_e \leq E, \quad \pi_e \in [0,1], \quad \pi_e \in [E; \mathcal{E}].
\end{aligned}
$$

3. Summary of Our Solutions

The optimal instrumentation problem in (2) is the classical NP-complete knapsack problem [4, p. 68]. A simple greedy solution of (2) is based on the decreasing order of the ratios.
\[ \rho(e) := \frac{1 - \mathbb{E}[P^B_B(e)]}{C_e}, \]  

(3)

However, in general, computing \( \rho(e) \) for each link \( e \) is itself NP-hard since \( \mathbb{E}[P^B_B(e)] \) involves computing the permanent of a matrix.

We propose two approaches to efficiently computing the order of \( \rho(e) \) in (3). One approach avoids computation of \( \mathbb{E}[P^B_B(e)] \) altogether and instead infer their order through simple link parameters such as the load factors. The other approach computes \( \mathbb{E}[P^B_B(e)] \) approximately by considering only events under small batch sizes \( B \).

### 3.1 Approach 1: Load-Factor Heuristic

We propose the load-factor heuristic for selection of links for instrumentation in (2) under equal costs of instrumentation \((C_e \equiv C)\). The heuristic selects transition links in the decreasing order of their load factors \((L_e = \frac{\lambda_e}{\mu_e})\) until the budget constraint is met. We can see that the load-factor heuristic is the optimal selection strategy for instrumentation whenever the monitoring accuracies \( \mathbb{E}[P^B_B(e)] \) at different transition links follow the reverse order of their load factors \( L_e \).

We now provide conditions for optimality of load factor heuristic when the monitoring strategy in the absence of instrumentation is FIFO matching. Let \( S_e \) denote the random service time along a transition link \( e \) normalized by the arrival rate \( \lambda_e \) to the link. Note that \( \mathbb{E}[S_e] = L_e \), the load factor. Let the “spread” \( V_e \) of \( S_e \) be given by

\[ V_e := S_e(1) - S_e(2), \]  

(4)

where \( S_e(1) \) and \( S_e(2) \) are independent samples drawn from the distribution of \( S_e \).

**Theorem 1 (FIFO Comparison).** For two transition links with normalized service times \( S_e \) and \( S_{e'} \), we have

\[ V_e^\text{st} \geq V_{e'}^\text{st}, S_e \geq S_{e'} \Rightarrow \mathbb{E}[P^B_B(e)] \leq \mathbb{E}[P^B_B(e')]. \]  

(5)

Hence, if all the normalized service times and their spreads at different links satisfy a stochastic order, then the load factor heuristic is optimal under equal instrumentation costs and FIFO matching rule.

We now provide optimality conditions for load-factor heuristic when the monitoring strategy in the absence of instrumentation is random matching, where we uniformly pick a match over all possible matches.

**Theorem 2 (Comparison Under Random Match).** Under random matching rule at links \( e, e' \) with normalized service times \( S_e \) and \( S_{e'} \) with supports \([\Delta_1, \Delta_2]\) and \([\Delta'_1, \Delta'_2]\),

\[ S_e \geq S_{e'}, \Delta_1 \leq \Delta'_1 \Rightarrow \mathbb{E}[P^B_B(e)] \leq \mathbb{E}[P^B_B(e')]. \]  

(6)

Hence, if all normalized service times satisfy a stochastic order and their supports satisfy the above condition, then the load factor heuristic is optimal under equal instrumentation costs and random matching rule.

### 3.2 Approach 2: Low Rate Approximation

The sufficient conditions to guarantee optimality of load-factor heuristic may not always hold and moreover, the costs for instrumenting different links may vary. For this case, we devise an alternative strategy for instrumentation which is tight in the low arrival rate regime. The simplest approximation is the first order approximation where we consider only the event that the busy period consists of a single transaction. At low arrival rate, this approximation becomes tight

\[ \lim_{\lambda_e \rightarrow 0} \frac{\mathbb{P}[B_e = 1]}{\mathbb{E}[P^B_B(e)]]} = 1. \]  

(7)

Using this approximation, we provide a greedy solution for (2) based on the decreasing order of the ratios

\[ \rho'(e) := 1 - \frac{P[B_e = 1]}{C_e}, \quad \forall e \in \mathcal{E}. \]  

(8)

### 4. Conclusion

In this paper, we formulated the problem of selectively retrofitting monitoring instrumentation to maximize the effectiveness in tracking transactions throughout the end-to-end business system. In contrast to making ad-hoc decisions, our strategy enables an efficient allocation of the instrumentation resources. We proposed two simple heuristics which are shown to be optimal or close to the optimal instrumentation strategy under some conditions. Our heuristics are intuitive and base their instrumentation decisions on simple queueing parameters such as the load factor or the probability of a unit-sized busy period. We formally proved the optimality of the load-factor heuristic for a wide-range of service distributions using the machinery of stochastic comparison. We are currently evaluating the performance of our proposed methods, see Fig.2.

### 5. References


