Expected end-to-end latency of cause-effect chains

Yde Sinnema
Lund University, Sweden
Email: yde.sinnema@control.lth.se

Martina Maggio
Saarland University, Germany & Lund University, Sweden
Email: maggio@cs.uni-saarland.de

Abstract

The literature on end-to-end latency of cause-effect chains is strongly oriented towards bounding the maximum reaction time and maximum data age. When combining the real-time analysis with the control-theoretical design of cyber-physical systems, only considering the worst case yields suboptimal behaviour under normal operating conditions. We want to promote research at the boundary between control design and real-time systems, and in particular derive probability distributions of end-to-end latency under uncertainty caused by jitter and deadline misses. Such results will enable the design of controllers that optimise the expected performance of cyber-physical systems while guaranteeing worst-case compliance.

I. Introduction

The complexity of modern cyber-physical systems has led to control software implementations that split the choice of the control action into several tasks running at different rates, possibly on different processing units. This is common in robotics, where different sensing and actuation modules are combined and must communicate via shared interfaces [1], [2]. As an illustrative example, consider the control pipeline in Figure 1 that consists of two sensor tasks gathering data from a physical system, an estimator task that that applies a sensor fusion algorithm on the sensor data, and a controller task that computes the actuation.

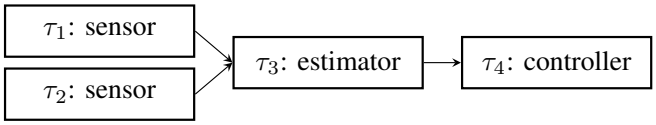

Fig. 1. Example of a DAG representing two cause-effect chains of tasks.

The timing analysis of such systems does not only concern individual instances of tasks, but also chains of tasks that transfer data to each other [3]. The functional dependencies between these tasks are usually represented by cause-effect chains or by directed acyclic graphs (DAGs). End-to-end latency metrics quantify the propagation time of data through a cause-effect chain, from the sampling of a signal by the source task to the actuation imposed by the sink task. The latency depends both on the processing time of each task and on the additional delay introduced by misalignment or rate differences between tasks.

As reviewed in [4], existing research focuses on bounding the worst-case end-to-end latency to ensure that some imposed timing requirements are met. However, from the control design perspective, the maximum latency is not the only metric of interest. While having an upper bound on the latency enables the synthesis of robust controllers that guarantee stability for every possible delay value, these controllers do not perform optimally in the average case [5] (as an example, see the controllers developed in [6]).

The same observation can be made about existing probabilistic analyses that take uncertainty related to execution time or deadline misses into account. Estimations of the execution time distribution aim to construct a safe upper bound rather than an accurate probability distribution [7]. Similarly, work on probabilistic end-to-end latency focuses on exceedance probabilities instead of latency distributions [8].

Regardless of whether probabilistic factors are taken into account, there is variability in latency between data paths. The alignment of tasks with different rates changes throughout the hyperperiod, causing a varying synchronisation delay. Non-determinism in the form of unknown execution times, jitter in read and write instants [9], or deadline miss probabilities [10], [11], [12] additionally generates uncertainty about the propagation path of data through a chain. The difference between data being processed by one job of a task or the next one can have a major effect on the end-to-end latency and subsequently the closed-loop control performance.

We argue that there is a need for research that derives probability distributions for end-to-end latency metrics. This is a crucial prerequisite for the design of control algorithms that optimise for expected performance while still being robust with respect to the worst case. In the following, we will briefly formalise a proposed system model before stating the concrete research problems to be addressed.

II. System model

In this section, we present a possible model to analyse the end-to-end latency distribution of a DAG of periodic tasks that are subject to some uncertainty. We focus on the uncertainty introduced by small discrepancies in read and write instants. We assume that the task model tries to follow the Logical Execution Time (LET) paradigm [13], but the implementation might be imprecise. We believe this to be a model suitable for the analysis of control pipelines in robotics; other models may be better for different application domains.

A task τ is defined as the tuple $(T, \phi, f^{\mathsf{r}}, f^{\mathsf{w}})$. $T \in \mathbb{N}$ is the task period and $\phi \in [0, T)$ is the task phase, i.e., the time offset of the release of the first job of τ. The probability density functions $f^{\mathsf{r}}, f^{\mathsf{w}} : [0, T) \to [0, 1]$ represent the probability distributions of the read and write instants in a job, respectively. There may be a dependence between both, that could for example be inferred from an execution time distribution.

Data dependencies between tasks are represented by a DAG, where a directed edge $\tau_i \to \tau_j$ indicates that the data produced (written) by τ_i is consumed (read) by τ_j. Figure 2 illustrates the task model using the DAG structure in Figure 1.

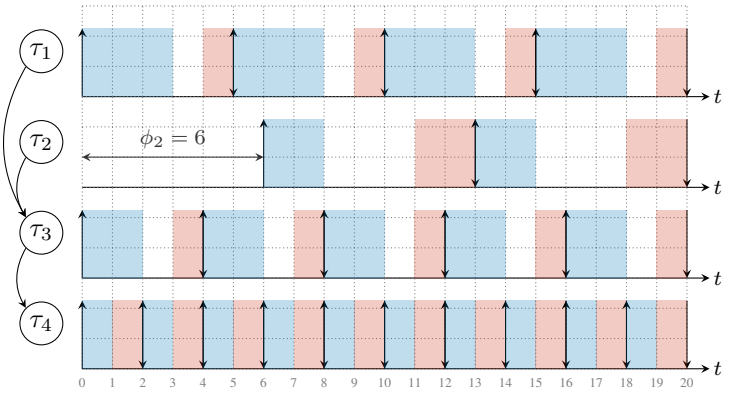

Fig. 2. Illustration of the task model with $\tau_1 = (5, 0, f_1^{\mathsf{r}}, f_1^{\mathsf{w}})$, $\tau_2 = (7, 6, f_2^{\mathsf{r}}, f_2^{\mathsf{w}})$, $\tau_3 = (4, 0, f_3^{\mathsf{r}}, f_3^{\mathsf{w}})$, $\tau_4 = (2, 0, f_4^{\mathsf{r}}, f_4^{\mathsf{w}})$. The read and write jitter is visualised by the blue and red areas which respectively indicate the regions with non-zero read and write probability.

Different communication semantics can be considered. We believe that the following are of most direct relevance to the problem at hand.

- Implicit communication: the read and write instants of a job are coupled to the start and end times of its execution. Therefore, variations in scheduling and execution time affect the timing of read and write operations.
- Logical Execution Time (LET) [13]: the read and write instants of a job are fixed to predetermined points in time. These usually correspond to the job release time and the end of the period (i.e., the release time of the next job), but different offsets are also possible [14]. This deterministic model decouples communication timing from task execution.

We consider the usual end-to-end latency metrics, but focus on their distribution and/or expected value instead of the maximum value.

- Reaction time: the time counted from a sample arriving at the source of a chain to an output corresponding to this sample being written at the end of the chain.
- Data age: the time counted (backwards) from an actuation generated by the sink task of a chain to the corresponding input sample being read at the start of the chain.

From the perspective of control design, the data age is more relevant. Under the reasonable assumption that the control action is determined by the last task in a chain, which communicates with the actuators of the physical system, the data age corresponds exactly to the delay experienced by the controller.

III. Research problems

Given a cause-effect chain and a communication model, the core problem is to compute the corresponding probability distribution and from this the expected value of both end-to-end latency metrics defined above.

An important consideration when solving this problem will be the trade-off between scalability and practical applicability. Stronger assumptions on, for example, independence between variables may yield simpler methods with a lower computational complexity, but at the same time reduce the realism of the model.

Extensions to the model should include DAGs of tasks, for which the notion of end-to-end latency distributions will need to be extended to include the interplay of multiple cause-effect chains in a single system. Another important direction is the study of the effect of deadline misses on end-to-end latency, given for example a per-job miss probability.

The final challenge is to exploit this knowledge when designing embedded estimation and control algorithms.

REFERENCES

[1] Y. Tang, Z. Feng, N. Guan, X. Jiang, M. Lv, Q. Deng, and W. Yi, "Response time analysis and priority assignment of processing chains on ros2 executors," in *2020 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, Dec. 2020, pp. 231–243.

[2] H. Teper, M. Gunzel, N. Ueter, G. von der Bruggen, and J.-J. Chen, "End-to-end timing analysis in ROS2," in *2022 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, Dec. 2022.

[3] N. Feiertag, K. Richter, J. Nordlander, and J. Jonsson, "A compositional framework for end-to-end path delay calculation of automotive systems under different path semantics," in *International Workshop on Compositional Theory and Technology for Real-Time Embedded Systems*, 2008.

[4] M. Günzel, H. Teper, G. v. d. Brüggen, and J.-J. Chen, "End-to-end latency of cause-effect chains: A tutorial," *ACM Transactions on Embedded Computing Systems*, vol. 24, no. 1, pp. 1–18, Dec. 2024.

[5] B. Lincoln and A. Cervin, "JITTERBUG: a tool for analysis of real-time control performance," in *41st IEEE Conference on Decision and Control, CDC 2002, Las Vegas, NV, USA, December 10-13, 2002*. IEEE, 2002, pp. 1319–1324. [Online]. Available: https://doi.org/10.1109/CDC.2002.1184698

[6] M. Seidel, M. Maggio, and F. Allgöwer, "A controller synthesis framework for weakly-hard control systems," 2026. [Online]. Available: https://arxiv.org/abs/2603.20146

[7] R. I. Davis and L. Cucu-Grosjean, "A survey of probabilistic timing analysis techniques for real-time systems," *Leibniz Transactions on Embedded Systems (LITES)*, vol. 6, pp. 03:1–03:60, 2019.

[8] M. Günzel, N. Ueter, K.-H. Chen, G. von der Brüggen, and J.-J. Chen, "Probabilistic reaction time analysis," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 5s, pp. 1–22, Sep. 2023.

[9] S. Wang, E. Bini, Q. Deng, and M. Maggio, "Jitter propagation in task chains," in *2025 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, Dec. 2025, pp. 96–108.

[10] G. von der Brüggen, N. Piatkowski, K.-H. Chen, J.-J. Chen, K. Morik, and B. B. Brandenburg, "Efficiently approximating the worst-case deadline failure probability under EDF," in *2021 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, Dec. 2021, pp. 214–226.

[11] A. Friebe, F. Marković, A. V. Papadopoulos, and T. Nolte, "Efficiently bounding deadline miss probabilities of Markov chain real-time tasks," *Real-Time Systems*, vol. 60, no. 3, pp. 443–490, Sep. 2024.

[12] M. Zini, F. Marković, D. Casini, A. Biondi, and B. B. Brandenburg, "In search of butterflies: Exceedance analysis for real-time systems under transient overload," in *2024 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, Dec. 2024, pp. 229–242.

[13] C. M. Kirsch and A. Sokolova, *The Logical Execution Time Paradigm*. Springer Berlin Heidelberg, Oct. 2011, pp. 103–120.

[14] L. Maia and G. Fohler, "Reducing end-to-end latencies of multi-rate cause-effect chains in safety critical embedded systems," in *12th European Congress on Embedded Real Time Software and Systems (ERTS 2024)*, Toulouse, France, Jun. 2024.