QoS Coverage Maps for Placement and Readiness in Urgent Digital Twins

Sowmya Yellapragada*, Dante Sanchez-Gallegos†, Rafael Tolosana Calasanz‡, Jesus Carretero Perez†, Manish Parashar*
*University of Utah, Salt Lake City, UT, USA
†Universidad Carlos III de Madrid (UC3M), Madrid, Spain
‡University of Zaragoza (UNIZAR), Zaragoza, Spain

***Abstract*—Urgent continuum workflows may miss deadlines because of poor task placement, delayed artifact readiness, insufficient compute capacity, or excessive full-fidelity execution time. We isolate placement and readiness while holding capacity and fidelity fixed. We introduce *QoS coverage maps*, which vary offered load and readiness delay and classify whether the primary diagnostic target $\tau = 0.50$ is met by the best-performing of three non-learning baselines, learned placement, advance artifact staging, or neither. The threshold separates operating regions. For wildfire, learned placement improves on-time completion by up to 85.5 percentage points, while increasing readiness delay from 0 to 10 s at 1.0 fps reduces completion from 87.7% to 0.4%. For earthquake–tsunami, learned placement adds no coverage relative to the primary baseline because 96.4% of critical-path computation is fixed to HPC. At $\tau = 0.50$, 10 of 30 tested conditions remain uncovered in each workflow, motivating capacity scaling or bounded approximation.**

***Index Terms*—computing continuum, urgent computing, digital twins, data readiness, quality of service, QoS coverage maps, multi-agent reinforcement learning**

I. Motivation and Problem

Urgent workflows must produce results before their operational value expires [1]. Timeliness depends not only on task placement, but also on whether the state required for execution is ready. Although learning-based schedulers can improve placement [2], [3], a dispatched task may still wait for its required state to be materialized. We define this residual preparation time as *readiness delay*, which consumes the same deadline budget as queuing and execution.

A deadline miss may therefore result from poor placement or insufficient readiness, requiring different interventions. We ask which intervention is sufficient to meet a QoS target under a given workload and readiness condition, and use QoS coverage maps to identify regions covered by placement, advance artifact staging, or neither.

II. QoS Coverage-Map Methodology

We introduce a QoS coverage map as a two-dimensional representation of the operating conditions under which an urgent workflow meets a prescribed QoS target. It measures on-time completion across offered load λ and readiness delay d, as follows:

$$Q(\lambda, d) = \frac{\text{incidents completing within } D}{\text{incidents that owed a deadline}}. \tag{1}$$

The deadline budget is

$$T = T_{\text{ready}} + T_{\text{queue}} + T_{\text{transfer}} + T_{\text{exec}}. \tag{2}$$

Contention increases queuing delay, while poor readiness increases reconstruction delay. Both can occur together and consume the same deadline budget, producing the diagonal frontier. We sweep λ against d with capacity K and fidelity f fixed. We use $\tau = 0.50$ as the primary diagnostic threshold, not as a deployment safety target. A deployment-specific threshold should come from the application's QoS contract. Cells are labeled *none* when the best-performing of three non-learning baselines succeeds, *placement-covered* when the baseline fails and learned placement succeeds, *readiness-covered* when learned placement succeeds only after the required artifacts are reconstructed in advance at the same load, and *uncovered* otherwise. Coverage labels are conditional on the threshold, baseline set, fixed capacity, and fixed fidelity.

III. System and Workflows

We evaluate the QoS coverage-map methodology using two urgent workflows with contrasting placement and readiness characteristics: wildfire monitoring and earthquake–tsunami forecasting.

The wildfire workflow detects and verifies smoke, issues an alert within 2 s, and produces a fire-spread forecast within 20 s [5]. Its compute-intensive mapping and forecasting stages can execute on cloud or HPC resources, while elevated wildfire risk may provide time to pre-stage required artifacts. Thus, both placement and readiness can affect deadline satisfaction.

The earthquake–tsunami workflow produces a shaking alert within 2 s and an inundation forecast within 600 s [6]. Under our profiled configuration, 96.4% of the critical-path computation is restricted to HPC, limiting placement freedom. Moreover, earthquakes provide no useful event-specific preparation window, so required artifacts must either already be ready or incur preparation delay after the event.

The experimental continuum comprises five edge nodes, two cloud sites, and an HPC tier with four execution slots, connected through an approximately 85 ms, 20 Mbps uplink. Storage includes hot NVMe replicas and colder remote representations of workflow artifacts.

We define *readiness* as the availability of the data, models, software, dependencies, and runtime state required for exe-

cution. In our experiments, cold data/model artifacts must be materialized before dependent forecasting tasks can execute. If t_{mat} is the materialization time and t_{pre} the available preparation time, the residual readiness delay is

$$d = \max(0, t_{\mathrm{mat}} - t_{\mathrm{pre}}), \tag{3}$$

and its normalized value is $\delta = d/D$, where D is the workflow deadline.

To isolate readiness effects, artifacts are returned to the cold state after each incident. Persistent caching and risk-aware retention are outside the scope of this study.

IV. Placement Approach

We use hierarchical multi-agent reinforcement learning with centralized training and decentralized execution [7]. Weight-shared edge agents choose local execution or offload, while a remote agent selects cloud or HPC from queue and congestion observations. We train one policy per workflow and compare it with three non-learning baselines. The primary map uses the best-performing baseline in each cell. For robustness, wildfire labels are also recomputed with a deadline-aware Frontier baseline. The earthquake simulator currently lacks a compatible deadline-aware Frontier implementation.

V. Experimental Design and Results

We evaluate six loads, five readiness levels, and four schedulers with capacity and fidelity fixed. Policies are trained with pre-warmed artifacts and evaluated zero-shot. Wildfire uses 30 seeds per cell and earthquake–tsunami 20, with 95% confidence intervals. Using the same measured cell values, we also reclassify coverage at $\tau \in \{0.50, 0.75, 0.90\}$ and test baseline sensitivity without rerunning the workflows.

For wildfire, learned placement is most effective at intermediate loads, with a statistically significant advantage in 14 of 30 cells. At 1.0 fps, increasing readiness delay from 0 to 10 s reduces on-time completion from 87.7% to 0.4%. At $\tau = 0.50$, 10 of 30 cells in each workflow remain uncovered. At $\tau = 0.75$, this grows to 20 of 30 wildfire cells and 15 of 30 earthquake–tsunami cells. At $\tau = 0.90$, no wildfire cell reaches the target because the maximum observed completion is 87.7%. The 2.5–3.0 fps region remains uncovered across all tested thresholds and baseline choices. Wildfire labels are unchanged with the deadline-aware Frontier baseline at every tested threshold. Relative to the primary baseline, learned placement adds no earthquake–tsunami cells. An alternative simple placement policy agrees on 26 of 30 cells at $\tau = 0.50$.

TABLE I
Wildfire pre-warmed row with 30 seeds. Learned values are mean ±95% CI. Gain is learned placement minus the best-performing non-learning baseline in percentage points.

Load (fps)	Learned (%)	Best baseline (%)	Gain (pp)
0.5	85.8 ± 7.4	85.6	+0.2
1.0	87.7 ± 5.6	2.2	+85.5
1.5	73.2 ± 11.7	0.9	+72.3
2.0	55.1 ± 16.5	0.7	+54.4
2.5	27.9 ± 13.9	0.2	+27.7
3.0	1.6 ± 1.1	0.0	+1.6

TABLE II
Required intervention across each 30-cell tested grid at the primary diagnostic threshold $\tau = 0.50$, relative to the best-performing non-learning baseline.

Intervention	Wildfire	Earthquake–tsunami
None, best baseline suffices	13.3%	46.7%
Placement	16.7%	0.0%
Readiness	36.7%	20.0%
Uncovered	33.3%	33.3%

VI. Implications and Planned Work

Workflow structure, readiness state, and workload jointly determine the binding constraint. Coverage boundaries depend on the QoS threshold and baseline. At the primary diagnostic threshold $\tau = 0.50$, 10 of 30 cells in each workflow remain beyond placement and readiness. Stricter thresholds expand the uncovered region. Future work will release fidelity and capacity as additional intervention dimensions and identify the minimum-cost configuration that restores timeliness while satisfying an explicit quality bound. Current scope fixes capacity and fidelity, excludes persistent cross-incident caching, and validates two contrasting workflows.

References

[1] D. Balouek-Thomert, E. G. Renart, A. R. Zamani, A. Simonet, and M. Parashar, "Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows," *Int. J. High Perform. Comput. Appl.*, vol. 33, no. 6, pp. 1159–1174, 2019.

[2] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. ACM Workshop on Hot Topics in Networks (HotNets)*, 2016, pp. 50–56.

[3] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. NeurIPS*, 2017, pp. 6379–6390.

[4] H. Weatherspoon and J. D. Kubiatowicz, "Erasure coding vs. replication: A quantitative comparison," in *Proc. Int. Workshop on Peer-to-Peer Systems (IPTPS)*, 2002, pp. 328–337.

[5] J. Mandel, J. D. Beezley, and A. Kochanski, "Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011," *Geosci. Model Dev.*, vol. 4, no. 3, pp. 591–610, 2011.

[6] J. Macías, M. J. Castro, S. Ortega, C. Escalante, and J. M. González-Vida, "Performance benchmarking of Tsunami-HySEA model for NTHMP's inundation mapping activities," *Pure Appl. Geophys.*, vol. 174, no. 8, pp. 3147–3183, 2017.

[7] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," in *Proc. NeurIPS*, 2022.