QoS Coverage Maps for Placement and Readiness in Urgent Digital Twins

Sowmya Yellapragada[1], Dante Sanchez[2], Rafael Tolosana Calasanz[3], Jesus Carretero Perez[2], Manish Parashar[1]

[1] University of Utah, [2] Universidad Carlos III de Madrid (UC3M), [3] University of Zaragoza (UNIZAR)

PROBLEM AND OBJECTIVES

Urgent digital-twin applications, such as wildfire monitoring and earthquake-tsunami forecasting, run deadline-constrained workflows across a shared edge-cloud-HPC continuum. A workflow can miss its deadline for four distinct reasons:

- Tasks **placed** on congested or unsuitable resources
- Required data, models, containers or runtime state **not ready** at the execution site
- Available compute **capacity insufficient**
- The full-fidelity configuration **needs more computation** than the deadline permits

Existing orchestration optimizes placement while treating readiness, capacity and output quality as fixed. It therefore cannot say which constraint caused a miss, or which intervention would fix it.

How can an orchestrator identify the minimum intervention needed to satisfy an end-to-end QoS deadline under changing contention and readiness?

METHODOLOGY

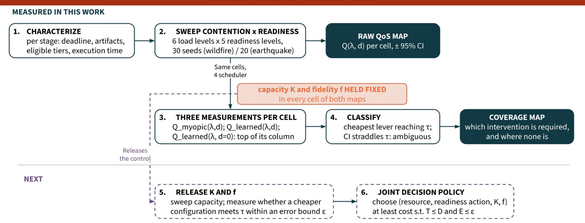


Fig. 0. Current maps hold **capacity and fidelity fixed**. Moving downward changes only residual readiness delay, so vertical changes isolate the storage/readiness effect. Capacity and fidelity are additional intervention dimensions that are not released in the present maps.

CONCEPTUAL MODEL

Q(λ, d) = incidents on time / incidents owing a deadline

(λ) is offered load, (d) residual readiness delay normalized by the workflow deadline (D), and (τ) is the on-time completion threshold. A cell is admissible when (Q(λ, d) ≥ τ). We use τ = 0.50 as the primary diagnostic threshold for separating operating regions, not as a deployment safety target. A deployment-specific threshold should come from the application's QoS contract.

$$T = T_{ready} + T_{queue} + T_{transfer} + T_{exec}$$

Congestion increases T_{queue}, while poor readiness increases T_{ready}. Both can occur together and consume the same deadline budget, producing the diagonal frontier.

EXPERIMENTAL SETUP

- **Scenario A, wildfire.** Cameras spot smoke, a detection alerts responders (2 s) and triggers a spread forecast over terrain and fuel data (20 s). Heavy: Cloud or HPC, can pre-warm.
- **Scenario B, earthquake-tsunami.** Seismometers detect a quake, shaking alert (2 s), then a tsunami forecast over seafloor bathymetry (600 s). HPC-only: 96.4% of compute cannot move, no event-specific pre-warm window.
- **Infrastructure.** 5 edge, 2 cloud, HPC (4 slots), uplink ~85 ms, ~20 Mbps. Storage: hot NVMe, warm, cold erasure-coded.
- **Placement.** Hierarchical CTDE MARL per workflow vs the best of 3 non-learning baselines.
- **Readiness.** `d = max(0, t_rec − prewarm)`, as % of each deadline.
- **Sweep.** 6 loads x 5 readiness x 4 schedulers, capacity fixed, zero-shot. 30/20 seeds per cell.

SCENARIO A: WILDFIRE

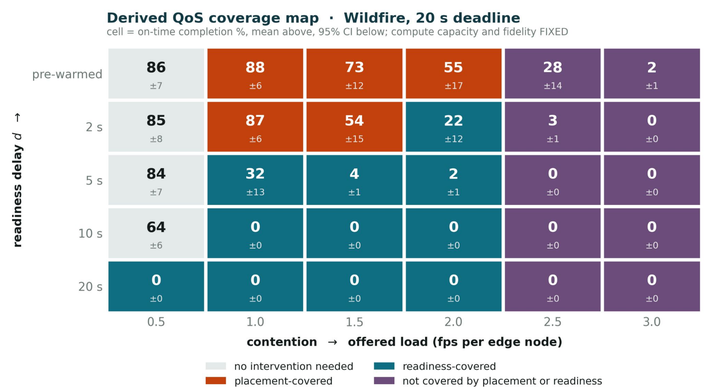


Fig. 1. Wildfire coverage map at the primary diagnostic threshold τ = 0.50, 30 seeds per cell. The number reports learned-policy on-time completion (mean above, 95% CI below). Cell color reports the class: no intervention, placement-covered, readiness-covered, or uncovered. Capacity and fidelity are fixed.

SCENARIO B: EARTHQUAKE-TSUNAMI

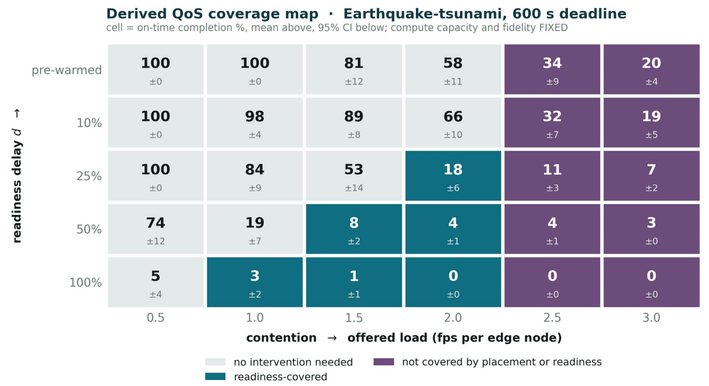


Fig. 2. Earthquake–tsunami coverage map at τ = 0.50, 20 seeds per cell. Readiness is normalized to the 600-second tsunami deadline. Cell numbers and class colors use the same convention as Fig. 1.

OBSERVATIONS

- **Workflow structure limits placement:** Wildfire's heavy stages are movable; 96.4% of earthquake–tsunami's critical path is HPC-bound.
- **Wildfire placement gain:** Up to +85.5 pp, robust across both baselines for τ ≤ 0.75.
- **Earthquake-tsunami placement:** 0/30 added cells with the primary baseline; 3 cells are baseline-sensitive at τ = 0.50.

Scope: Capacity and fidelity are fixed. Readiness excludes persistent cross-incident caching. Validation covers two contrasting workflows.

WHERE LEARNED PLACEMENT HELPS AND WHERE IT STOPS

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

Table 1. Wildfire, pre-warmed row, 30 seeds, ±95% CI. 14 of 30 cells across the full map show an advantage significant at 95%.

WHERE READINESS BINDS

Readiness delay	On-time (%)	Rehydrations
pre-warmed	87.7 ±5.6	0
2 s (10%)	87.5 ±5.7	639
5 s (25%)	32.3 ±13.3	465
10 s (50%)	**0.4** ±0.2	**424**
20 s (100%)	**0.1** ±0.1	**427**

Table 2. Wildfire at 1.0 fps with identical compute. Model re-cooled per incident, so every dormant to active transition re-pays reconstruction. Rehydrations fall as (d) grows because fewer forecasts survive to trigger one

ROBUSTNESS TO QoS THRESHOLD AND BASELINE

U = uncovered cells. We reclassify the same measured cells at τ ∈ {0.50,0.75,0.90}. No new workflow runs were required.

Domain	U at τ = 0.50	U at τ = 0.75	U at τ = 0.90
Wildfire	10/30	20/30	30/30*
Earthquake–tsunami	10/30	15/30	20/30

* At τ = 0.90 no wildfire cell reaches the target because the highest measured on-time completion is 87.7%.

- High load: All 2.5–3.0 fps cells stay uncovered.
- Wildfire: Labels are unchanged with deadline-aware Frontier.
- Earthquake: A deadline-aware Frontier is not available. The alternative agrees on 26/30 cells.

CONCLUSIONS

- Workflow structure strongly affects whether placement is a useful lever: wildfire has movable heavy stages, while earthquake–tsunami is largely HPC-constrained.
- Placement and readiness address different parts of the same deadline budget.
- Coverage boundaries depend on the QoS threshold and baseline. The 2.5–3.0 fps high-load region remains uncovered across all robustness checks.
- At the primary diagnostic threshold τ = 0.50, 10/30 cells in each workflow remain beyond placement and readiness. Stricter thresholds expand this region.
- The remaining region motivates additional levers such as capacity and certified approximation.