Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Physio-Semantic Compute Orchestration in Cyber-Physical Systems: A Mathematically Constrained Multi-Agent Edge-Cloud Architecture for Real-Time Manufacturing Intelligence

Bhupender Sharma[1], Dr. Sandeep Kumar Yadav[2]

1*Research Scholar, Department of Mechanical Engineering, School of Engineering & Technology, Vikrant University, Gwalior, (M.P.), India com

2*Ph.D. Guide, Department of Mechanical Engineering, School of Engineering & Technology, Vikrant University, Gwalior, (M.P.), India

ARTICLEINFO

Received: 02 Oct 2024

Revised: 19 Nov 2024

Accepted: 29 Nov 2024

ABSTRACT

AI-enabled diagnostics in industrial cyber-physical systems (CPS) increasingly combine lightweight edge models with large cloud-hosted language models. Existing model-routing methods choose among models by estimating the difficulty of a query and trading response quality against monetary cost. In a machine that is degrading, however, the consequence of a late answer depends on the physical condition of the machine, which a query-only router cannot observe. This paper proposes a physio-semantic orchestration formulation in which the physical state of the machine enters the routing decision explicitly. Each diagnostic event carries a semantic descriptor and a physical state estimate; the physical state determines a response deadline and an effective age under a Weibull degradation model, and the penalty for delay is the conditional probability of failure during the response interval. Routing is posed as an expected-cost minimization over a heterogeneous model set spanning edge, fog and cloud tiers, with terms for cost, diagnostic quality, state-dependent degradation, deadline risk and live queue state. We state the assumptions under which the formulation holds, show by a standard decision-theoretic argument when physical state has strictly positive value, and describe a three-tier multi-agent architecture in which cloud outputs remain advisory. The formulation is evaluated by discrete-event simulation on three synthetic surrogate scenarios structured after high-speed CNC machining, industrial acoustic anomaly detection (MIMII-structured) and turbofan degradation (C-MAPSS-structured), using 30 replications of 4000 events each and six baselines. In these simulations the proposed router reduced the mean realized objective relative to a semantic-only router by approximately 21% in the CNC scenario and by approximately 4% and 2% in the acoustic and turbofan scenarios, respectively, and the difference vanished when physical state was made uninformative. Queue-blind routing produced unbounded backlog. The pre-specified targets of a 70% reduction in latency and cost with 98% accuracy retention relative to cloud-only routing were not jointly met in any scenario. The results illustrate the behaviour of the formulation under stated synthetic assumptions and are not evidence of performance on real machines; validation on the original datasets and on hardware is identified as future work.

Keywords: cyber-physical systems; compute orchestration; model routing; edge-cloud computing; multi-agent systems; predictive maintenance; Weibull degradation; queueing; manufacturing intelligence

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1. INTRODUCTION

1.1 Industrial CPS and real-time manufacturing intelligence

Industrial cyber-physical systems couple physical processes, such as machining, rotation and thermodynamic conversion, with computation and communication. Lee et al. describe a layered architecture for CPS-based manufacturing in which sensing, analytics and decision functions are organized above the machine level [1]. Three kinds of monitoring problem recur in this setting and are used as representative scenarios throughout this paper. The first is high-speed machining, where spindle vibration, motor current and thermal signals must be interpreted within a short window to avoid workpiece or tool damage. The second is acoustic diagnosis of rotating and reciprocating machinery, for which the MIMII collection provides recordings of valves, pumps, fans and slide rails under normal and anomalous conditions [3]. The third is prognostics for turbofan engines, for which the C-MAPSS simulation produced run-to-failure sensor trajectories in which a health index defined from operational margins reaches zero at failure [2].

These scenarios differ in time scale, sensing modality and the form of the diagnostic output, but they share a structural problem. An event arrives, a diagnostic computation must be performed somewhere in a distributed computing hierarchy, and the value of the result depends both on how well it is computed and on how soon it is available.

1.2 AI- and LLM-enabled industrial diagnostics

Large language models have been examined for design and manufacturing tasks, including converting text prompts into specifications and manufacturing instructions, with both benefits and limitations reported [4]. A broader survey discusses applications in manufacturing and identifies the need for refinement, integration with quantitative models, and attention to interpretability and reliability [5]. Both works are preprints and are cited here for the problem context rather than as validated performance evidence. What these sources establish is that language-model capabilities are being considered for tasks that were previously handled by numerical monitoring pipelines. What they do not establish is that such models can be placed inside a physical control loop, and this paper does not assume so.

1.3 Edge-cloud latency and the computational orchestration problem

Edge computing places processing near the data source and is motivated in part by response-time requirements, bandwidth cost and data privacy [6]. Cloud-hosted models offer greater capability but add network delay, queueing and, for autoregressive generation, inference time that grows with output length. A routing layer that decides where each task is executed is therefore a natural component of a hybrid deployment. Hybrid inference between a small edge-deployable model and a large cloud model, with a router that predicts query difficulty and a tunable quality level, has been proposed for language tasks [7]. In this paper the term orchestration refers to the decision of which model, in which tier, handles each diagnostic event, together with the load-dependent consequences of that decision.

1.4 The latency-quality-cost trade-off

Three quantities are in tension. Larger models give higher expected diagnostic quality but have higher cost and latency. Smaller models answer quickly and cheaply but fail on difficult cases. Cascades and routers aim to move along this trade-off adaptively: FrugalGPT learns which combination of language-model APIs to query for different inputs, and reports matching the best single model at substantially lower cost on the tasks studied [8]. For an industrial machine a fourth quantity is added. A response that arrives late has a physical consequence that depends on the condition of the machine at the time of the event. This fourth quantity is the subject of the paper.

1.5 Research gap

The routing literature cited above optimizes response quality against inference cost. RouterBench, for example, evaluates routers by inference dollar cost and performance over a large set of pre-computed model outputs [9]. RouteLLM learns routers between a stronger and a weaker model from human preference data [10]. Hybrid inference conditions on predicted query difficulty [7]. In these works the input to the router is the query, and, as far

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as the cited descriptions indicate, no variable describing the state of a physical asset enters the objective. Multi-agent frameworks for language-model applications, such as AutoGen, provide general infrastructure for agents that converse with one another [11]; real-time constraints of physical processes are outside the scope described for that framework. The gap addressed here is narrow and specific: a routing formulation in which the physical state of the machine is an explicit input that changes both the cost of delay and the feasible set of models. We do not claim that no work exists on any part of this problem. A systematic literature search to support a stronger priority claim has not been performed, and the claim made is limited to the formulation reported here.

1.6 Research objective

The objective is to formulate, analyse and evaluate in simulation a physio-semantic routing policy for industrial CPS, and to determine under which conditions physical-state-aware routing differs materially from semantic-only routing.

1.7 Research questions

RQ1. How can physical machine state and semantic task complexity be combined into a routing decision for a heterogeneous model set, and what information does each contribute?

RQ2. Under what conditions does conditioning on physical state yield a strictly lower expected cost than a policy that ignores it?

RQ3. How does a router that ignores live queue state behave when load is high, and what load condition separates stable from unstable operation?

RQ4. In simulated CNC, acoustic and turbofan scenarios, does the proposed policy reach the pre-specified targets of at least 70% reduction in latency and cost with at least 98% accuracy retention relative to cloud-only routing?

1.8 Research hypotheses

H1. Under state-dependent deadlines and degradation, the proposed policy yields a lower mean realized objective than a semantic-only policy with the same information about queue state, and the advantage grows with the share of events whose latency ceiling is below the capability demanded by the task.

H2. When physical state carries no information about deadlines or degradation, the proposed policy and the semantic-only policy make identical decisions, so the advantage is zero.

H3. A router that omits live queue state produces unbounded backlog at a tier whose offered load reaches its service capacity, whereas queue-aware routing keeps mean waiting time bounded at the same arrival rate.

H4 (proposed target, not an expectation). The proposed policy achieves a reduction of at least 70% in mean latency and in mean cost relative to cloud-only routing while retaining at least 98% of its accuracy. The target is taken from the research proposal and is tested rather than assumed.

Each hypothesis is tied to measurable quantities in Section 5: the mean realized objective, the deadline violation rate, the mean waiting time and tier utilization, and the ratios of latency, cost and accuracy to those of cloud-only routing.

1.9 Contributions

The contributions are classified by their relation to existing work. Weibull reliability modelling [12] and queueing and Lyapunov analysis of stochastic networks [13] are established theory and are used here as modelling tools. Cost-quality routing between models [7], [8], [10] is established; its adaptation to a setting with hard, state-dependent deadlines is the first adaptation made here. The second is an integration: the conditional failure probability under a Weibull model is used as the delay penalty, with effective age inferred from observed physical state, so that the same delay costs more on a machine that is further along its degradation. The third is a formulation in which the semantic capability demanded by a task and the latency ceiling permitted by the machine state are two separate quantities whose conflict defines when physical state matters. The fourth is an executable simulation protocol with common random numbers across policies, six baselines, four ablations and sensitivity

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

analysis, released as code. A decision-theoretic result on the value of physical state is stated in Section 3, but it is an application of a standard argument and is not presented as a new theorem. No Lyapunov stability guarantee is derived in this paper.

1.10 Paper organization

Section 2 reviews related work and states the gap for each strand. Section 3 formalizes the system model and the optimization problem. Section 4 describes the proposed framework and the three-tier agent architecture. Section 5 describes the simulation and evaluation methodology. Section 6 reports the simulation results. Section 7 discusses interpretation, limitations and threats to validity, and Section 8 concludes.

2. RELATED WORK

This section reviews the strands of work on which the formulation rests. Each subsection states what has been done, the method used, the limitation that remains, why that limitation matters for industrial CPS, and how the present work responds. Only sources that were checked against a publisher, proceedings or repository record during preparation of this draft are cited. Where a strand is represented by few verified sources, the coverage is stated as limited rather than extended from memory; a systematic search is listed as future work in Section 8.

2.1 AI-based mechanical fault diagnosis

A review of deep learning for machine health monitoring organizes the field around auto-encoders, restricted Boltzmann machines and deep belief networks, convolutional networks and recurrent networks [14]. The common method is to learn features from sensor signals and map them to fault classes or health indicators. The limitation relevant here is not accuracy but deployment: these models are described and evaluated as signal-to-label mappings, and the review does not treat the placement of computation across an edge-fog-cloud hierarchy as a decision variable. In CPS this matters because the same model can be acceptable at the edge and unacceptable in the cloud solely because of response time. The present work treats the choice of where a diagnosis is computed as the decision to be optimized, with the diagnostic models themselves taken as given.

2.2 Predictive maintenance and RUL estimation

Data-driven remaining-useful-life estimation has used convolutional networks applied along the temporal dimension of multichannel sensor data, with feature learning and regression trained together [15], and deep convolutional networks for prognostics more generally [16]. Evaluation of such methods depends on run-to-failure data, which are scarce for safety-critical equipment. The C-MAPSS simulation addressed this by generating sensor trajectories from a thermodynamic engine model in which flow and efficiency losses grow at an exponential rate from a randomly chosen starting point until a health index, defined as the smallest of several operational margins, reaches zero [2]. The limitation for orchestration is that RUL models output an estimate and, usually, an error metric; the time taken to obtain the estimate and the machine condition at the time of the request do not appear in the evaluation. For prognostics the latency penalty is likely to be small compared with the time scale of degradation, an expectation the simulation in Section 6 examines rather than assumes.

2.3 Industrial acoustic anomaly detection

The MIMII collection contains sounds from four machine types, namely valves, pumps, fans and slide rails, with seven product models per type, recorded in factory conditions and including anomalies such as contamination, leakage, rotating unbalance and rail damage [3]. It was introduced to support unsupervised detection of anomalous sound, and it contains 26,092 files of normal operation, each segment lasting ten seconds [3]. The limitation for the present purpose is that the collection supplies audio and condition labels, with no notion of a response deadline, a queue, or a machine state beyond the recording condition. Acoustic monitoring is nevertheless a representative case of event-triggered diagnosis with moderate time constraints, and it is the second application domain. Other acoustic anomaly detection methods are not reviewed here, because the sources available during drafting were not verified.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.4 LLMs and generative AI in industrial systems

Early assessments of language models in design and manufacturing examine their use in turning text into specifications, producing manufacturing instructions and exploring design variants, and report both promising results and clear limitations [4]. A survey of language models in manufacturing discusses applications and identifies refinement, integration with quantitative models, interpretability and reliability as open challenges [5]. Both are preprints. The limitation is the absence of a treatment of the timing of inference relative to machine dynamics. The consequence for CPS is that the capability of large models cannot be used indiscriminately: a model that can reason about a root cause may return its answer after the window in which the answer is actionable. The proposed formulation makes this explicit by separating the capability a task demands from the latency the machine state permits.

2.5 Edge-cloud AI orchestration

Edge computing was introduced as a paradigm that processes data near its source, motivated by response-time requirements, battery life, bandwidth cost and privacy [6]. Hybrid inference for language models routes each query between a small model that can run on a low-cost edge device and a large cloud model, based on predicted query difficulty and a quality level that can be tuned at test time, and reports up to 40% fewer calls to the large model without a drop in response quality on the tasks studied [7]. The limitation is that the router is conditioned on the query and a desired quality level; the cost being traded is the cost of using the large model. For an industrial machine, the cost of delay is not constant and depends on the condition of the asset. The present formulation retains the edge-cloud routing idea and replaces the constant delay cost by a state-dependent one.

2.6 LLM and model routing

Cascades, benchmarks and learned routers form a coherent line of work. FrugalGPT learns which combinations of language-model APIs to use for different queries, to reduce cost and improve accuracy [8]. RouterBench provides a dataset of more than 405 thousand inference outcomes from representative models and a framework for evaluating routers by inference cost and performance [9]. RouteLLM trains routers between a stronger and a weaker model from human preference data and data augmentation, and reports cost reductions of more than a factor of two on public benchmarks without sacrificing response quality [10]. All three optimize quality against monetary cost for language tasks. The limitations for industrial CPS are that latency is not the object of optimization in the formulations cited, that no state of a physical asset enters the objective, and that there is no notion of a deadline that varies with the event. These limitations matter because the cost of a wrong or late diagnosis in a machine is not monetary API cost. The present work keeps the router-as-decision-rule structure of these methods and extends the objective.

2.7 Multi-agent systems for CPS and manufacturing

AutoGen provides a framework in which multiple language-model agents converse to accomplish tasks, with customizable agent behaviours and conversation patterns [11]. Lee et al. describe an architecture for CPS-based manufacturing systems in which functions are organized in levels from sensing to decision-making [1]. The limitation is a gap between the two: the language-model framework is generic infrastructure, and the manufacturing architecture does not specify how generative agents with very different response times are to be coordinated. The consequence for CPS is that agents on the machine side and agents in the cloud operate on different time scales, and a coordination rule must state which agent may act on the physical process. The proposed architecture assigns physical actuation only to the lowest tier and treats higher-tier output as advisory (Section 4). Literature on holonic and agent-based manufacturing control is not reviewed in this draft because the corresponding sources were not verified.

2.8 Physics-informed and physically constrained AI

Physics-informed neural networks incorporate laws expressed as differential equations into the training of neural networks, so that the learned function respects those laws [17]. The method addresses the fidelity of a learned model to physics. It does not address the question of where the model is run or whether its answer arrives in time. The present work uses physical knowledge differently: not inside the diagnostic model, but in the orchestration

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

objective, through a degradation model that converts observed state and delay into a failure probability. The Weibull distribution that serves this purpose is a flexible family in which the hazard rate may be decreasing, constant or increasing [12], and it is adopted as a modelling assumption subject to sensitivity analysis, not as a physical law.

2.9 Comparative research-gap analysis

Table 1 compares the verified sources along the dimensions relevant to the formulation. The entries describe what each source states in its abstract or main text as read for this draft. A dash indicates that the dimension is not addressed in the source as described.

Table 1. Literature comparison along the dimensions relevant to physio-semantic orchestration.

Ref.	Domain	Method	Edge/cloud	Physical state	Semantic or query complexity	Dynamic routing	Limitation relative to this work
[1]	Manufacturing CPS	Layered CPS architecture	Not specified	Sensing layer	–	–	No routing or timing formulation
[2]	Turbofan RUL data	Thermodynamic simulation with exponential damage	–	Simulated sensors, health index	–	–	Data generator; no compute placement
[3]	Machine sound	Dataset for unsupervised anomaly detection	–	Audio under normal and anomalous operation	–	–	No deadline, queue or state beyond recording condition
[14]	Machine health monitoring	Review of deep learning	–	Sensor signals	–	–	Computation placement not treated
[15]	RUL prognostics	Convolutional regression on multichannel sensors	–	Sensor signals	–	–	Latency and request-time state not evaluated
[16]	RUL prognostics	Deep convolutional networks	–	Sensor signals	–	–	Latency and request-time state not evaluated
[4]	Design	Case	Cloud	–	–	–	Timing against

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ref.	Domain	Method	Edge/cloud	Physical state	Semantic or query complexity	Dynamic routing	Limitation relative to this work
	and manufacturing	studies with GPT-4	model				machine dynamics not treated
[5]	Manufacturing	Survey of LLM use	–	–	–	–	Timing against machine dynamics not treated
[6]	General IoT	Edge computing paradigm	Edge	–	–	–	Conceptual; no routing objective
[7]	Language tasks	Difficulty-based router, small versus large	Edge and cloud	–	Predicted query difficulty	Yes	Constant cost of delay; no asset state
[8]	Language tasks	LLM cascade	Multiple LLM APIs	–	Query-dependent selection	Yes	Monetary cost only; no asset state
[9]	Language tasks	Router benchmark	Multiple models	–	Router inputs are prompts	Evaluated	Cost and quality only; no asset state
[10]	Language tasks	Preference-trained binary router	Strong and weak model	–	Query-based	Yes	Cost and quality only; no asset state
[11]	LLM applications	Multi-agent conversation	Not specified	–	–	Agent-defined	Generic; physical timing outside scope
[17]	Physical simulation	Physics-informed neural networks	–	Physical law in training	–	–	Fidelity to physics; no placement decision

The matrix supports a limited conclusion. Within the sources examined, query-conditioned routing exists [7]–[10], physical state is used in diagnostic and prognostic modelling [2], [3], [14]–[17], and layered or multi-agent architectures exist [1], [11], but no source examined combines query-conditioned routing with a physical-state-dependent delay penalty. Because the search was not systematic, this observation is a statement about the sources examined and not a general claim of absence.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. SYSTEM MODEL AND PROBLEM FORMULATION

This section defines the model on which the router operates. To keep the status of each statement clear, every item is labelled as an established result, an assumption (A), a proposed definition (D), or a derived result (R). Symbols are summarized in Table 2. All quantities denoted as costs, probabilities and weights are dimensionless unless a unit is given.

3.1 Industrial CPS model

The system receives a stream of diagnostic events indexed by i, arriving at times t_i (s). Event i carries a semantic descriptor, a physical state estimate and the current state of the computing hierarchy. A routing policy maps this observation to one of K models, each deployed in one of three tiers (edge, fog, cloud). The model chosen for event i is denoted j_i. The physical process evolves independently of the routing decision except through the consequence of delay, which is formalized in Section 3.8. The framework does not place the cloud tier in the safety loop: physical actuation is reserved for the lowest tier and higher-tier outputs are advisory (Section 4.8).

3.2 Physical-state representation

(D1) The physical state at time t is a vector s(t) in R^p whose components are measured or derived signals, for example vibration RMS, kurtosis of the vibration signal, temperature gradient, spindle speed, motor current or acoustic band energies. A monotone health map g converts the state into a degradation fraction h = g(s), with h in [0, 1). The router observes an estimate $\hat{h} = h + \varepsilon_s$, where ε_s is zero-mean noise with standard deviation σ_s.

(A1) The degradation fraction h summarizes the aspects of the physical state that determine the cost of delay. This is an assumption: real machines require a calibrated map g, and the quality of that map determines how much of the benefit analysed here can be realized. The simulation in Section 6 varies σ_s to examine the dependence.

3.3 Semantic-task representation

(D2) The semantic content of event i is represented by a feature vector q_i derived from the alarm text or query, for example syntactic depth or the frequency of domain terms. Each event has a latent difficulty d_i in [0, 1], the reasoning demand of the task, and the router observes a scalar proxy z_i = φ(q_i). In the simulation z_i = d_i + ε_z with zero-mean noise of standard deviation σ_z. Neither d_i nor z_i has a physical unit.

(A2) The proxy z_i is informative about d_i. The datasets that define the application domains contain no natural-language queries, so in this study the semantic descriptor is generated synthetically; this limitation is discussed in Section 7.9.

3.4 Computational model set

(D3) The model set is M = {m_1, ..., m_K}. Model j is characterized by its tier, the median service time S_j (s), the log-scale dispersion σ_j of the service time, the median network delay n_j (s), the number of parallel servers N_j, a normalized invocation cost γ_j, and a capability χ_j in [0, 1]. The probability that model j diagnoses a task of difficulty d correctly is

$$p_j(d) = p_{lo} + (p_{hi} - p_{lo}) / (1 + \exp[-\theta (\chi_j - d)]) \quad (1)$$

where p_lo and p_hi are the floor and ceiling accuracies and θ is a slope parameter. Equation (1) is a proposed functional form chosen for monotonicity and smoothness, not an empirical law.

(A3) Capability increases with tier index, $\chi_1 < \chi_2 < \ldots < \chi_K$, so that on average larger models are better. Individual tasks may violate the ordering, and no result below depends on strict dominance.

3.5 Latency model

(D4) The end-to-end latency of event i handled by model j is

$$L_{ij} = L_{route} + L_{net,ij} + L_{q,ij} + L_{inf,ij} \quad (2)$$

where L_route is the router overhead (s), L_net,ij the network delay, L_q,ij the waiting time in the queue of model j, and L_inf,ij the inference time. The inference time is modelled as log-normal with median S_j and log-scale σ_j.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Log-normality is an assumption chosen for positivity and right skew; it is not derived from the architecture of any model.

3.6 Cost model

(D5) Each invocation of model j incurs cost γ_j, normalized so that one cloud invocation has cost 1. The cost stands for a combination of monetary price and resource consumption. Energy is not modelled separately, because no energy figures are available for the surrogate models.

3.7 Diagnostic-quality constraint

The minimum quality requirement is Pr(correct) ≥ α. In the router this hard constraint is relaxed into the penalty $w_q (1 - \hat{p}_j(z))$, where $\hat{p}_j(z)$ is an estimate of the probability that model j is correct given the semantic proxy, learned from labelled history (Section 4.3). The relaxation turns the constraint into a trade-off whose strength is set by w_q and does not guarantee that α is met on every event.

3.8 Machine degradation and urgency model

(Established) The Weibull cumulative distribution, introduced in the engineering literature by Weibull [12], is

$$D(t) = 1 - \exp[-(t/\eta)^{\kappa}] \quad (3)$$

with scale η (s) and shape κ. Its survival function is $S(t) = 1 - D(t)$ and its hazard rate is $\lambda(t) = (\kappa/\eta)(t/\eta)^{(\kappa-1)}$, which is decreasing for $\kappa < 1$, constant for $\kappa = 1$ and increasing for $\kappa > 1$.

(A4) The progression of damage in the monitored process follows the Weibull model (3). This is a modelling assumption. It is justified only to the extent that a monotone or wear-out hazard describes the failure mode of interest, and it is subjected to sensitivity analysis over κ in Section 6.

(D6) The effective age of a machine with degradation fraction h is the age at which the Weibull distribution function equals h,

$$a(h) = \eta\,[-\ln(1 - h)]^{(1/\kappa)} \quad (4)$$

so that $D(a(h)) = h$. The cost of delay is the conditional probability that the machine reaches the failure state during a response interval of length L, given that it has survived to age a(h),

$$\Psi(h, L) = 1 - S(a + L) / S(a) = 1 - \exp\{ -[((a + L)/\eta)^{\kappa} - (a/\eta)^{\kappa}] \} \quad (5)$$

The unconditional increment of the distribution function over the interval is $D(a + L) - D(a) = (1 - h)\,\Psi(h, L)$, which differs from Ψ by the survival factor. The conditional form is used because the machine is observed to be running at the time of the event.

(R1) The delay penalty (5) has the following properties. First, Ψ is strictly increasing in L. Second, for fixed L, the exponent in (5) has derivative with respect to a equal to $-(\kappa/\eta^{\kappa})[(a + L)^{(\kappa-1)} - a^{(\kappa-1)}]$, so Ψ increases with the effective age, and hence with h, when $\kappa > 1$; it is independent of h when $\kappa = 1$, where $\Psi = 1 - \exp(-L/\eta)$; and it decreases with h when $\kappa < 1$. Consequently a given delay is more costly on a machine that is further along a wear-out degradation, and the shape parameter controls whether the physical state affects the penalty at all. The result follows from differentiation and is not a new theorem; it is stated because the sensitivity analysis in Section 6 is organised around it.

(A5) The response deadline depends on the degradation fraction according to

$$\tau(h) = \tau_{max}\,(\tau_{min} / \tau_{max})^{h} \quad (6)$$

so that $\tau(0) = \tau_{max}$ and $\tau(1) = \tau_{min}$, with a geometric interpolation. Equation (6) is a simulation assumption that encodes the idea that a machine closer to failure leaves less time for a response. Real deadlines would be derived from process dynamics and safety analysis. An event misses its deadline when $L > \tau(h)$.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.9 Queueing and load model

(D7) Each model j is served by N_j identical servers with first-come-first-served discipline. Let f_j(t) denote the earliest time at which a server of model j becomes free, which the router can observe. The router's estimate of the waiting time of an event arriving at time t is

$$\hat{W}_j = \max\{0, f_j(t) - t - L_{route} - n_j\} \quad (7)$$

Let ν_j be the long-run rate of events routed to model j and E[S_j] = S_j exp(σ_j²/2) the mean service time. The offered load of tier j is

$$\rho_j = \nu_j E[S_j] / N_j \quad (8)$$

(Established) For a stationary arrival process, ρ_j < 1 is necessary for the queue of tier j to be stable; if ρ_j ≥ 1 the backlog grows without bound. The simulation uses a two-state Markov-modulated Poisson arrival process with bursts, so that instantaneous load can exceed the long-run average. Service times are not assumed exponential, and closed-form M/M/c expressions are not used. Lyapunov drift-plus-penalty methods provide a framework for jointly bounding queues and optimizing time-averaged penalties [13]; applying that framework to this problem to obtain a formal guarantee is outside the scope of this paper.

3.10 Physio-Semantic Complexity

The research proposal defines physio-semantic complexity as a function C_PS = f(C_semantic, C_physical) of the semantic and physical components. This paper gives the function a specific form. The capability a task demands and the speed the machine state permits are two quantities on the same scale, the tier index.

(D8) With a quality threshold α_c, the semantic component is the smallest tier whose estimated accuracy meets the threshold,

$$C_{sem}(z) = \min\{ j : \hat{p}_j(z) \geq \alpha_c \} \quad (9)$$

and the physical component is the largest tier whose nominal latency fits within the deadline,

$$C_{phys}(\hat{h}) = \max\{ j : L_{route} + n_j + S_j \leq \tau(\hat{h}) \} \quad (10)$$

The physio-semantic complexity of an event is the pair C_PS = (C_sem, C_phys) and the orchestration slack is Δ = C_phys − C_sem. When Δ ≥ 0 at least one model both meets the quality threshold and fits the deadline, and the routing problem is mainly one of cost. When Δ < 0 no model does both, and the router must trade quality against the consequence of delay. This descriptor is used for analysis and does not need to be computed by the router. It also explains why semantic-only routing can be adequate in some domains and not in others: if Δ ≥ 0 for almost all events, the physical state rarely changes the best choice.

3.11 Routing policy

(D9) The proposed policy chooses the model that minimizes an estimated expected cost for the observation x = (z, ĥ, queue state),

$$\pi^*(x) = \text{argmin over } j \text{ of } J_j(x), \text{ with } J_j(x) = w_c \gamma_j + w_q [1 - \hat{p}_j(z)] + w_d \Psi(\hat{h}, \hat{L}_j) + w_m \hat{P}_j^{miss}(\hat{h}) \quad (11)$$

where $\hat{L}_j = L_{route} + n_j + \hat{W}_j + S_j$ is the nominal latency of model j and

$$\hat{P}_j^{miss} = 1 - \Phi_N([\ln(\tau(\hat{h}) - L_{route} - n_j - \hat{W}_j) - \ln S_j] / \sigma_j) \quad (12)$$

is the probability that the log-normal inference time exceeds the remaining time budget, with Φ_N the standard normal distribution function; $\hat{P}_j^{miss} = 1$ when the budget is not positive. The weights w_c, w_q, w_d, w_m are dimensionless and express the relative importance of cost, quality, degradation and deadline risk. They are fixed a priori in this study (Section 5.5) and are not tuned.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The variants used as baselines and ablations are obtained from (11) by restricting information. A semantic-only policy replaces $\hat{h}$ by the population median of h. A physical-only policy replaces $\hat{p}_j(z)$ by the population mean accuracy of model j. The variants differ only in the information available, not in the form of the objective.

3.12 Optimization problem

The design problem is the constrained expected-cost problem

(P1) minimize over π of $E[\, w_c \gamma_j + w_d \Psi(h, L) \,]$ subject to $\Pr(\text{correct}) \geq \alpha$, $\Pr(L > \tau(h)) \leq \varepsilon$, $\rho_j < 1$ for every tier j, and $L_{route} \leq \xi \min_j (S_j + n_j)$ (13)

where ε is a tolerated deadline-violation probability and ξ bounds the router overhead relative to the fastest model. Problem (P1) is not solved exactly. The router implements the Lagrangian-style relaxation (11), which moves the quality and deadline constraints into the objective with fixed multipliers and handles the load constraint through the observed queue state in (7). The relaxation is a heuristic: it does not guarantee that the constraints in (13) hold, and the simulation reports the realized violation rates and utilizations instead.

(R2) Value of physical state. Let $x_S = (z, \text{queue state})$ be the information available to a semantic-only policy and $x = (z, h, \text{queue state})$ the information available to a policy that also observes the degradation fraction, and let $J(j; x)$ denote the expected realized cost of choosing j given x. A policy that chooses argmin over j of $J(j; x)$ for each x attains an expected cost no greater than any policy that depends only on x_S. The inequality is strict if and only if the minimizing model changes with h on a set of observations of positive probability, because the pointwise minimizer of the conditional expected cost is optimal among all policies measurable with respect to the same information, and a coarser information set can only match it. This is a standard decision-theoretic argument and is stated to make precise what the experiments test: the benefit of physical state arises exactly when the argmin depends on h, which by (5), (6) and (11) requires that τ or Ψ vary with h. For $\kappa = 1$ the penalty Ψ is independent of h (R1), and the dependence can then arise only through the deadline. When neither τ nor Ψ depends on h the benefit is zero, which is hypothesis H2.

(R3) Queue-blind instability. A router that ignores queue state and sends a long-run fraction π_j of events to tier j loads that tier with $\rho_j = \nu \pi_j E[S_j] / N_j$. If this reaches 1 the tier is unstable by the necessary condition stated after (8), irrespective of how well the router estimates quality. This is an elementary consequence of the stability condition and motivates the inclusion of (7) in (11).

Table 2. Principal symbols, definitions and units.

Symbol	Definition	Unit / range
i, j, K	Event index, model index, number of models	–
t_i, ν	Arrival time of event i; long-run arrival rate	s; events/s
s(t), h, $\hat{h}$	Physical state vector; degradation fraction; its estimate	mixed; [0,1); [0,1)
σ_s, σ_z	Standard deviations of state and semantic noise	dimensionless
d, z	Latent task difficulty; observed semantic proxy	[0,1]
χ_j, $p_j(d)$	Capability of model j; probability of correct diagnosis	[0,1]; [0,1]
S_j, σ_j, n_j, N_j	Median inference time; log-scale; median network delay; servers	s; –; s; –
γ_j	Normalized invocation cost	cloud call = 1
L, L_{route}, L_{net}, L_q, L_{inf}	End-to-end latency and its components	s
η, κ	Weibull scale and shape	s; –

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Symbol	Definition	Unit / range
D(t), S(t), λ(t)	Weibull distribution, survival and hazard functions	–; –; 1/s
a(h)	Effective age implied by h	s
Ψ(h, L)	Conditional failure probability during delay L	[0,1]
τ(h), τ_min, τ_max	Response deadline and its bounds	s
f_j(t), Ŵ_j, ρ_j	Earliest free time; estimated wait; offered load of tier j	s; s; –
C_sem, C_phys, Δ	Capability demand; latency ceiling; orchestration slack	tier index
w_c, w_q, w_d, w_m	Objective weights (cost, quality, degradation, deadline risk)	dimensionless
J	Realized per-event objective	dimensionless

4. PROPOSED PHYSIO-SEMANTIC COMPUTE ORCHESTRATION FRAMEWORK

4.1 Overall architecture

The framework places a router between the event sources and the model tiers. Events enter from two sources: alarms or queries that carry text, and streams of sensor data from which a physical state estimate is computed. The router combines the two, consults the current state of each model queue, and dispatches the event to one model. Results return to the requesting supervisory function. Tier-1 outputs may trigger machine-level protective actions through a safety function that is independent of the orchestration layer. Outputs from the fog and cloud tiers are advisory. A detailed specification for the architecture diagram is given in Appendix A (Figure 1).

4.2 Physio-semantic feature engineering

The physical features are computed over a moving window of the sensor stream. Candidates named in the research proposal are the RMS envelope of vibration, the kurtosis of the vibration signal as an indicator of transient shocks, the relative temperature gradient, spindle speed and motor current. For acoustic monitoring the analogous features are band energies; for turbofan data the physical summary is a health index derived from the sensor channels. A learned or engineered map g converts these features into the degradation fraction estimate $\hat{h}$ (Section 3.2). The semantic features are lightweight text descriptors of the alarm or query, such as syntactic depth and the frequency of domain terms, and are mapped to the scalar proxy z (Section 3.3). Feature computation must itself be cheap, since its time is part of the routing overhead L_route; the cost of the feature stage is not measured in this study.

4.3 Complexity estimation

The router needs two estimators: the accuracy estimate $\hat{p}_j(z)$ of each model as a function of the semantic proxy, and the degradation estimate $\hat{h}$. In the simulation $\hat{p}_j(z)$ is a binned empirical table with 20 bins over z, trained on 20,000 labelled synthetic events generated with a seed disjoint from all evaluation seeds. Each cell holds the observed fraction of correct diagnoses of model j in that bin, and the table falls back to the model's overall accuracy where a bin holds fewer than 20 events. With real data, a gradient-boosted tree ensemble over the semantic features is a candidate estimator, as suggested in the research proposal; this paper does not evaluate such an estimator. The estimate $\hat{h}$ is the observed value with additive noise of standard deviation σ_s.

4.4 Routing mechanism

For each event the router executes the following steps. It first obtains z and $\hat{h}$. It then obtains the earliest-free time f_j of each model and computes Ŵ_j by equation (7). For each of the K models it evaluates the nominal latency $\hat{L}_j$, the delay penalty $\Psi(\hat{h}, \hat{L}_j)$ by equation (5), the deadline-miss probability by equation (12), and the objective

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

J_j by equation (11). It dispatches the event to the model with the smallest objective. The load-shedding behaviour described in the research proposal follows from the same rule: as the estimated wait at a tier grows, both the delay penalty and the miss probability for that tier increase, and events move to other tiers without a separate switching mechanism.

4.5 Three-tier edge-fog-cloud multi-agent architecture

The architecture has three tiers whose agents differ in capability, response time and authority. The first tier is at the machine, the second at the facility, and the third in a remote datacenter. Table 3 summarizes the tiers and the models used in the simulation.

4.6 Agent roles

The tier-1 agents supervise the machine. Two models are used in the simulation: a reflex model with a service time of the order of milliseconds that stands for a threshold detector or compact classifier, and an edge small language model with a service time of the order of a hundred milliseconds. The research proposal describes tier-1 agents as one-to-three-billion-parameter models operating at 100 Hz. This paper does not assume that a generative model of that size can respond at that rate; the reflex function is represented by a fast model of unspecified architecture, and the service times of all models are simulation parameters, not measurements of any device. The tier-2 agent is a fog-hosted model that can combine information from several machines and use longer histories. The tier-3 agent is a large cloud-hosted model used for tasks that demand the highest capability, such as interpreting unstructured documentation. It is advisory only.

Table 3. Agent tiers, simulated models and responsibilities.

Tier	Model in simulation	Deployment class	Responsibility	Authority over the physical process
1 (edge)	m1, reflex model	Machine-side controller or gateway	Immediate supervision of thresholds and short-horizon anomalies	May trigger protective action through the independent safety function
1 (edge)	m2, edge small language model	Edge computer	Routine diagnosis of low-complexity events	Advisory, plus tier-1 supervision
2 (fog)	m3, fog model	Facility server	Cross-machine and history-aware diagnosis	Advisory
3 (cloud)	m4, cloud model	Remote datacenter	Highest-capability reasoning on complex or novel cases	Advisory only; never in the safety loop

4.7 Communication and orchestration protocol

Messages between the router and the agents carry the event identifier, the time stamp, the semantic and physical features, the deadline $\tau(\hat{h})$, the selected tier and an advisory flag. The transport is a design choice; industrial protocols such as OPC UA or MQTT are candidates and are not evaluated here. Two rules govern the interaction. First, escalation does not remove tier-1 supervision: an event sent to a higher tier is still monitored locally, so that a late or missing response from a higher tier does not leave the machine unsupervised. Second, a response that arrives after its deadline is retained for logging and operator review but is not applied automatically.

4.8 Safety and human-in-the-loop constraints

The design assumes that safety functions are implemented by certified controllers independent of the orchestration layer, and that the orchestration layer can request but not override them. Outputs from tiers 2 and 3 that propose

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

changes to machine parameters are advisory and require confirmation by an operator or by a deterministic supervisory rule. Loss of connectivity to higher tiers must fall back to tier-1 behaviour. This paper makes no claim of compliance with any functional-safety standard; compliance would require a safety analysis of the complete system that is outside the scope of a routing study.

4.9 Computational complexity of the router

After feature extraction, the router performs a constant number of arithmetic operations per model, so the decision cost is O(K) per event, with an O(1) table lookup for each p^_j(z). Updating the server-free times of the chosen model costs O(log N_j) with a heap. Memory is O(K B) for the accuracy tables with B bins plus O(ΣN_j) for the server states. The simulation charges 0.5 ms of router overhead to every adaptive policy as an assumed value. The requirement stated in the research proposal that the router decide in less than 10 ms is a design target that has not been measured on any hardware in this work, and the simulation does not establish it.

5. SIMULATION AND EVALUATION METHODOLOGY

5.1 Research design

The evaluation is a discrete-event simulation comparing routing policies on synthetic surrogate scenarios, one for each application domain. All policies see the same events: for each replication, the arrival times, latent difficulties, degradation fractions, noise realizations, service-time draws, network-delay draws and the uniform random number that determines whether a model's diagnosis is correct are generated once and shared by all policies (common random numbers). Differences between policies are therefore paired. The experiments address H1 to H4 through the variables and metrics described below. The code is provided with the manuscript (Appendix B).

5.2 NASA C-MAPSS

C-MAPSS is a simulation of turbofan engines. Response surfaces of the sensors are generated by a thermodynamic model as a function of flow and efficiency changes in engine modules, an exponential rate of loss is imposed from a randomly chosen starting point until a failure criterion based on a health index is reached, and the output is a multivariate time series in cycles [2]. It represents degradation-oriented prognostics. In this study the dataset itself is not used. The turbofan scenario is a surrogate in which the degradation fraction h is drawn from a Beta distribution to represent engines at different stages of life, with long process time scale and relaxed deadlines. Descriptive statistics of the real dataset, such as the numbers of engines, operating conditions and sensors, are not reproduced here and will be taken from the official NASA documentation when the real data are used.

5.3 MIMII

MIMII contains sound recordings of valves, pumps, fans and slide rails under normal and anomalous conditions, with seven product models per machine type and ten-second segments [3]. It represents event-triggered acoustic anomaly detection. The acoustic scenario is a surrogate with moderate process time scale and moderate deadlines. As with C-MAPSS, no audio is processed.

5.4 CNC telemetry scenario

No CNC telemetry is used. The CNC scenario is fully synthetic. It represents high-speed machining in which damage escalates on a short time scale and deadlines for protective responses are short. The parameters were chosen to make the time scales of the three scenarios differ by roughly an order of magnitude, not to reproduce any particular machine.

5.5 Simulation assumptions

The following assumptions apply to all scenarios. Arrivals follow a two-state Markov-modulated Poisson process with a calm rate ν and a burst rate 3ν, with mean dwell times of 8 s and 2 s. Latent difficulty d and degradation fraction h are drawn from Beta distributions whose parameters are listed in Table 8, with h capped at 0.97. Observation noise has standard deviations σ_z = 0.15 and σ_s = 0.08 in the base case. The objective weights are w_c = 1, w_q = 3, w_d = 4 and w_m = 2, chosen a priori so that a cloud invocation, a wrong diagnosis, the delay penalty and a missed deadline all contribute at comparable magnitude; they were not tuned for any policy. The

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

model set has the parameters in Table 9. The accuracy function (1) uses p_lo = 0.30, p_hi = 0.99 and θ = 12. Each adaptive router is charged 0.5 ms of overhead. The realized objective for an event is J = w_c γ + w_q (1 − correct) + w_d Ψ(h, L) + w_m 1[L > τ(h)], evaluated with the true h and the realized latency. This realized objective is the common yardstick for all policies. Because the router objective (11) and the realized objective share the same form and weights, the evaluation favours a policy that estimates (11) well. This is a deliberate property of the design, discussed in Section 7.10.

Table 8. Simulation scenarios (synthetic surrogates; all values are simulation parameters).

Parameter	CNC (high-speed machining)	MIMII-structured (acoustic)	C-MAPSS-structured (turbofan)
Calm arrival rate ν (events/s)	5	4	3
Difficulty d, Beta(a, b)	(2, 3)	(2, 3)	(2.5, 2.5)
Degradation fraction h, Beta(a, b)	(1.5, 4)	(1.2, 6)	(1.5, 2.5)
Weibull scale η (s)	15	60	300
Weibull shape κ	2.5	1.8	3.0
Shortest deadline τ_min (s)	0.05	0.5	2
Longest deadline τ_max (s)	10	30	120

Table 9. Model set parameters used in all scenarios (simulation parameters, not measurements).

Model	Tier	Median inference S (s)	Log-scale σ	Median network n (s)	Servers N	Cost γ	Capability χ
m1 reflex	Edge	0.003	0.25	0	4	0.00	0.15
m2 edge small LM	Edge	0.150	0.30	0	2	0.05	0.42
m3 fog model	Fog	0.600	0.30	0.010	2	0.30	0.68
m4 cloud model	Cloud	1.600	0.35	0.060	16	1.00	0.92

Network delays are log-normal around the stated medians with log-scale 0.3 for the fog tier and 0.4 for the cloud tier.

Table 4. Application-domain and dataset characteristics.

Aspect	CNC / high-speed machining	Industrial acoustic (MIMII)	Turbofan degradation (C-MAPSS)
What it represents	High-speed manufacturing and machine-state monitoring	Industrial acoustic anomaly detection	Degradation and RUL-oriented prognostics
Source in this study	Synthetic scenario; no telemetry	Surrogate structured after MIMII [3]; no audio used	Surrogate structured after C-MAPSS [2]; no engine data used

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Aspect	CNC / high-speed machining	Industrial acoustic (MIMII)	Turbofan degradation (C-MAPSS)
Verified facts about the real source	Not applicable	Valves, pumps, fans, slide rails; seven product models per type; 26,092 normal files; ten-second segments [3]	Thermodynamic simulation; exponential flow and efficiency loss; health index defined from operational margins [2]
Natural-language queries in source	None	None	None
Relative time scale in the surrogate	Shortest	Intermediate	Longest

The three sources are not interchangeable. They differ in modality (telemetry, audio, multivariate cycles), in whether degradation is observed (a run-to-failure trajectory in C-MAPSS, discrete anomaly conditions in MIMII), and in time scale. The common element is the orchestration problem of Section 3.

5.6 Baseline methods

Table 5 lists the baselines and the information each uses. The adaptive baselines B4 and B5 use the same objective, the same queue information and the same overhead as the proposed method, and differ only in the information about the event. The fixed policies B1 and B2 are standard reference points rather than tuned competitors. A reference policy with access to the true difficulty and degradation fraction is reported as an upper reference and is not a baseline.

Table 5. Baseline methods and the information and assumptions each uses.

ID	Policy	Information used	Assumptions
B1	Cloud-only	None	Every event goes to m4; no router overhead
B2	Edge-only	None	Every event goes to the edge small language model m2; no router overhead
B3	Rule-based	Thresholds on $\hat{h}$ and z	If $\hat{h} > 0.60$ use m1; else if $z > 0.70$ use m4; else if $z > 0.45$ use m3; else use m2; thresholds fixed a priori, not optimized; no queue information
B4	Semantic-only	z, queue state	Objective (11) with $\hat{h}$ replaced by the population median of h
B5	Physical-only	$\hat{h}$, queue state	Objective (11) with p^_j(z) replaced by each model's population mean accuracy
B6	Proposed physio-semantic	z, $\hat{h}$, queue state	Full objective (11)
REF	Oracle reference	True d and h, queue state	Objective (11) evaluated with true values; not attainable in practice

5.7 Proposed method

The proposed method is policy B6, the minimizer of (11) described in Section 4.4, using the learned accuracy table of Section 4.3.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.8 Experimental variables

The independent variables are the routing policy, the scenario, the Weibull shape κ, the arrival-rate multiplier, the state-noise level σ_s, the degradation weight w_d, and whether the physical state is relevant to deadlines and degradation. The dependent variables are the metrics of Section 5.9. Seeds are the unit of replication.

5.9 Evaluation metrics

Table 6 lists the metrics. Precision, recall and F1-score are not computed, because the surrogate generates a single correct or incorrect outcome per diagnosis and not a class label set; accuracy is reported instead. Remaining-useful-life error is not computed because the surrogate does not produce RUL estimates. ROUGE-L is not used, because the diagnostic outputs are not generated text. Energy is not computed, because no energy figures exist for the surrogate models.

Table 6. Evaluation metrics.

Group	Metric	Definition
Diagnostic	Accuracy	Fraction of events diagnosed correctly, irrespective of timing
Diagnostic	Timely accuracy	Fraction diagnosed correctly and within the deadline
Latency	Mean, median, P95, P99 (s)	Statistics of end-to-end latency L over events
Computational	Mean cost	Mean invocation cost γ per event
Computational	Invocation frequency	Fraction of events sent to each model
Computational	Utilization	Total service time divided by N_j times the arrival horizon, per tier; values above 1 indicate backlog at the end of the horizon
CPS	Deadline violation rate	Fraction of events with $L > \tau(h)$
CPS	Degradation penalty	Mean $\Psi(h, L)$ with the true h
CPS	Queue behaviour	Mean waiting time and tier utilization
CPS	Routing overhead	Assumed 0.5 ms per adaptive decision
Composite	Mean realized objective J	Mean of the per-event objective defined in Section 5.5
Ratios to B1	Latency reduction, cost reduction, accuracy retention	One minus the ratio of mean latency, one minus the ratio of mean cost, and the ratio of accuracy, each relative to B1

5.10 Statistical analysis

Each scenario is simulated for 30 independent replications of 4000 events (seeds 1000 to 1029). For each metric the mean and standard deviation across replications are reported. Differences between the proposed policy and each comparator are computed per replication and analysed as paired samples. Confidence intervals come from a percentile bootstrap over replications with 10,000 resamples, at a significance level of 0.05 divided by six for the six comparisons made in each scenario (Bonferroni correction). A sign-flip permutation test with 20,000 random sign assignments gives a Monte Carlo p-value, whose smallest attainable value is about 5×10^{-5}. Replications are repeated draws from the same synthetic generator, so these statistics quantify simulation noise only and carry no information about real systems.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.11 Sensitivity analysis

For ten replications of 3000 events per setting and the policies B1, B4, B5 and B6, the following are varied one at a time from the base case: the Weibull shape κ in {1.0, 1.5, 2.5, 3.5}, the arrival-rate multiplier in {0.5, 1.0, 1.5, 2.0}, the state-noise standard deviation σ_s in {0.02, 0.08, 0.20, 0.40}, and the degradation weight w_d in {0, 2, 4, 8}. A null setting makes the physical state irrelevant: the deadline is constant at the geometric mean of τ_min and τ_max and w_d = 0 in both the router and the realized objective.

5.12 Ablation experiments

The ablations isolate the contribution of each information source and objective term (Table 7). A1 and A2 coincide with the baselines B4 and B5, and A4 with the proposed method B6. A5 is a supplementary variant that removes queue awareness, added after the first run indicated that load handling matters; it is reported as a supplementary result and not as a pre-specified ablation.

Table 7. Ablation design.

ID	Semantic features	Physical features	Queue state	Deadline-risk term	Same as
A1	Yes	No	Yes	Yes	B4
A2	No	Yes	Yes	Yes	B5
A3	Yes	Yes	Yes	No	–
A4 (full objective)	Yes	Yes	Yes	Yes	B6
A5 (supplementary)	Yes	Yes	No	No	–

6. RESULTS

All results in this section are outcomes of the simulation described in Section 5, run on synthetic surrogate generators. They show how the formulation behaves under the stated assumptions. They are not measurements of real machines, real datasets or real hardware, and no statement below should be read as a performance claim about C-MAPSS, MIMII or any CNC system. Values are means over 30 replications of 4000 events unless stated otherwise.

6.1 Structure of the evaluation

Section 6.2 reports the per-scenario metrics for all policies. Section 6.3 tests the proposed targets (H4). Section 6.4 reports paired statistical comparisons (H1). Section 6.5 reports the ablation. Section 6.6 reports queue behaviour (H3). Section 6.7 reports the sensitivity analysis and the null scenario (H2). Section 6.8 reports the orchestration-slack analysis, and Section 6.9 reserves the structure for experiments on the original datasets, which have not been performed.

6.2 Per-scenario results

Tables 10 to 12 report the metrics for each scenario. Latency is end-to-end latency in seconds, J is the mean realized objective (lower is better), and the number in parentheses is the standard deviation of J across replications.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 10. CNC scenario (synthetic). Mean over 30 replications of 4000 events.

Policy	Accuracy	Timely accuracy	Latency mean (s)	Median (s)	P95 (s)	P99 (s)	Cost	Deadline violation	Mean Ψ	J (sd)
B1 cloud-only	0.976	0.548	2.560	2.142	5.55	7.33	1.000	0.439	0.219	2.828 (0.115)
B2 edge-only	0.677	0.640	0.401	0.219	1.32	2.07	0.050	0.055	0.034	1.263 (0.052)
B3 rule-based	0.813	0.745	0.625	0.225	2.27	3.63	0.225	0.077	0.050	1.140 (0.034)
B4 semantic-only	0.833	0.719	0.764	0.424	2.09	2.47	0.203	0.132	0.065	1.228 (0.019)
B5 physical-only	0.769	0.762	0.698	0.366	2.01	2.54	0.160	0.008	0.045	1.051 (0.024)
B6 proposed	0.809	0.796	0.643	0.294	2.02	2.63	0.190	0.015	0.044	0.969 (0.021)
A3 no deadline term	0.843	0.777	0.786	0.483	2.20	2.77	0.231	0.078	0.058	1.091 (0.021)
A5 queue-blind	0.881	0.266	106.965	86.949	280.54	297.77	0.249	0.702	0.695	4.789 (0.065)
REF oracle	0.835	0.830	0.588	0.254	1.92	2.58	0.194	0.006	0.040	0.862 (0.018)

Table 11. MIMII-structured acoustic scenario (synthetic surrogate). Mean over 30 replications of 4000 events.

Policy	Accuracy	Timely accuracy	Latency mean (s)	Median (s)	P95 (s)	P99 (s)	Cost	Deadline violation	Mean Ψ	J (sd)
B1 cloud-only	0.975	0.971	2.039	1.860	3.71	4.81	1.000	0.005	0.028	1.194 (0.011)
B2	0.676	0.676	0.254	0.187	0.66	1.05	0.050	0.000	0.003	1.037

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Policy	Accuracy	Timely accuracy	Latency mean (s)	Median (s)	P95 (s)	P99 (s)	Cost	Deadline violation	Mean Ψ	J (sd)
edge-only										(0.021)
B3 rule-based	0.836	0.836	0.596	0.232	2.08	3.03	0.238	0.000	0.008	0.761 (0.017)
B4 semantic-only	0.909	0.896	2.232	1.763	6.11	7.37	0.337	0.014	0.031	0.762 (0.016)
B5 physical-only	0.805	0.805	3.754	4.831	8.29	9.37	0.197	0.001	0.043	0.955 (0.019)
B6 proposed	0.905	0.903	2.138	1.708	5.88	7.20	0.334	0.002	0.027	0.731 (0.014)
A3 no deadline term	0.907	0.902	2.198	1.749	6.01	7.29	0.336	0.005	0.028	0.740 (0.014)
A5 queue-blind	0.910	0.286	83.519	75.254	202.61	214.64	0.321	0.688	0.615	4.428 (0.223)
REF oracle	0.926	0.926	1.418	0.973	4.54	6.27	0.318	0.000	0.019	0.614 (0.013)

Table 12. C-MAPSS-structured turbofan scenario (synthetic surrogate). Mean over 30 replications of 4000 events.

Policy	Accuracy	Timely accuracy	Latency mean (s)	Median (s)	P95 (s)	P99 (s)	Cost	Deadline violation	Mean Ψ	J (sd)
B1 cloud-only	0.957	0.956	1.808	1.703	2.99	3.78	1.000	0.001	0.011	1.176 (0.011)
B2 edge-only	0.563	0.563	0.189	0.166	0.37	0.54	0.050	0.000	0.001	1.367 (0.022)
B3 rule-based	0.722	0.722	0.605	0.235	2.09	2.90	0.288	0.000	0.003	1.135 (0.021)
B4 semantic-only	0.895	0.885	2.149	1.584	6.10	8.54	0.478	0.011	0.013	0.867 (0.019)

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Policy	Accuracy	Timely accuracy	Latency mean (s)	Median (s)	P95 (s)	P99 (s)	Cost	Deadline violation	Mean Ψ	J (sd)
B5 physical-only	0.844	0.837	6.453	6.687	12.33	14.08	0.436	0.008	0.034	1.057 (0.023)
B6 proposed	0.894	0.893	1.972	1.528	5.33	7.40	0.479	0.002	0.012	0.847 (0.014)
A3 no deadline term	0.896	0.890	2.075	1.559	5.78	8.01	0.480	0.006	0.013	0.856 (0.016)
A5 queue-blind	0.889	0.805	6.495	3.904	19.93	24.73	0.455	0.094	0.041	1.139 (0.248)
REF oracle	0.908	0.908	1.438	1.166	3.88	5.82	0.451	0.000	0.009	0.761 (0.013)

In the CNC scenario the proposed policy has the lowest realized objective of all non-oracle policies, and its deadline violation rate of 0.015 is far below that of cloud-only (0.439), semantic-only (0.132) and the rule-based policy (0.077). Its accuracy (0.809) is lower than that of semantic-only routing (0.833) and much lower than that of cloud-only (0.976), because the physical state drives some events to faster, less capable models. Timely accuracy, which counts only correct diagnoses delivered within the deadline, is the highest of the non-oracle policies (0.796). The oracle reference, which observes the true difficulty and degradation fraction, attains J = 0.862, so a gap of 0.107 remains between the proposed policy and the information bound in this scenario.

In the acoustic and turbofan scenarios the differences between adaptive policies are small. The proposed policy has the lowest realized objective among non-oracle policies in both, but its margin over the semantic-only policy is 0.031 and 0.021, and over the rule-based policy in the acoustic scenario 0.030. In these two scenarios the mean latency of the proposed policy (2.138 s and 1.972 s) is slightly above that of cloud-only (2.039 s and 1.808 s), because the policy sends most events to the fog model, which operates at a high utilization (Section 6.6), so that waiting time, not inference time, dominates. In both scenarios the deadlines are rarely binding and the cost saving is the main benefit.

6.3 Assessment of the proposed targets (H4)

Table 13 compares the proposed policy with cloud-only routing against the targets set in the research proposal.

Table 13. Proposed policy (B6) relative to cloud-only (B1). Targets: at least 70% latency reduction, at least 70% cost reduction, at least 98% accuracy retention.

Scenario	Latency reduction	Cost reduction	Accuracy retention	Timely-accuracy ratio	All targets met?
CNC	74.9%	81.0%	82.9%	1.45	No (accuracy)
MIMII-structured	−4.9%	66.6%	92.8%	0.93	No (latency, cost, accuracy)

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Scenario	Latency reduction	Cost reduction	Accuracy retention	Timely-accuracy ratio	All targets met?
C-MAPSS-structured	−9.1%	52.1%	93.4%	0.93	No (latency, cost, accuracy)

A negative latency reduction means that the mean latency was higher than for cloud-only routing. The latency and cost targets are met in the CNC scenario only, where cloud-only routing is penalized by long queueing during bursts and many missed deadlines. The accuracy-retention target is not met in any scenario, and the targets are not jointly met anywhere. In the CNC scenario the proposed policy delivers 1.45 times as many correct and timely diagnoses as cloud-only routing, which illustrates why accuracy irrespective of timing may be the wrong yardstick for deadline-bound events. H4 is therefore not supported in these simulations.

6.4 Statistical comparison

Table 14 reports the paired difference in the realized objective between the proposed policy and each comparator, with Bonferroni-adjusted bootstrap intervals. Negative values favour the proposed policy. All sign-flip p-values are at the Monte Carlo floor (below 10^{-4}). As stated in Section 5.10, these statistics describe simulation noise only.

Table 14. Paired difference in mean realized objective, B6 minus comparator (30 replications). Brackets show the Bonferroni-adjusted 95% bootstrap interval for the mean difference.

Comparator	CNC	MIMII-structured	C-MAPSS-structured
B1 cloud-only	−1.859 [−1.915, −1.809]	−0.462 [−0.468, −0.458]	−0.329 [−0.335, −0.323]
B2 edge-only	−0.294 [−0.318, −0.274]	−0.305 [−0.315, −0.296]	−0.520 [−0.530, −0.510]
B3 rule-based	−0.171 [−0.182, −0.160]	−0.030 [−0.036, −0.024]	−0.288 [−0.297, −0.280]
B4 semantic-only	−0.259 [−0.265, −0.253]	−0.031 [−0.035, −0.027]	−0.021 [−0.026, −0.016]
B5 physical-only	−0.082 [−0.087, −0.076]	−0.224 [−0.230, −0.217]	−0.211 [−0.219, −0.203]
A3 no deadline term	−0.123 [−0.126, −0.119]	−0.009 [−0.010, −0.007]	−0.009 [−0.011, −0.007]

Expressed relative to the semantic-only policy, the proposed policy lowers the mean realized objective by about 21% in the CNC scenario, 4% in the acoustic scenario and 2% in the turbofan scenario. Relative to the physical-only policy the reductions are about 8%, 23% and 20%. H1 is supported in all three scenarios, with a large effect in the CNC scenario and small effects in the other two. Differences in the deadline violation rate are mixed: the proposed policy has a slightly higher violation rate than physical-only routing in the CNC and acoustic scenarios (differences of 0.0066 and 0.0010) and a lower one in the turbofan scenario (−0.0068), the physical-only policy buying its low violation rate with lower accuracy.

6.5 Ablation results

Table 15 reports the mean realized objective for the ablation variants.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 15. Mean realized objective J for the ablation variants (30 replications).

Variant	CNC	MIMII-structured	C-MAPSS-structured
A1 semantic features only	1.228	0.762	0.867
A2 physical features only	1.051	0.955	1.057
A3 both features, no deadline-risk term	1.091	0.740	0.856
A4 full objective (proposed)	0.969	0.731	0.847
A5 both features, queue-blind (supplementary)	4.789	4.428	1.139

Combining the two feature sets lowers the realized objective relative to using either alone in the acoustic and turbofan scenarios (A3 and A4 against A1 and A2), and the full objective is best in all three. In the CNC scenario, combining both features without the deadline-risk term (A3, 1.091) is worse than physical features alone with that term (A2, 1.051), so most of the benefit in that scenario comes from the deadline-risk term and not from the presence of both feature sets alone. The supplementary variant A5 performs worst everywhere, which relates to the load analysis that follows.

6.6 Queue behaviour and load (H3)

Table 16 reports mean waiting time and fog-tier utilization for selected policies.

Table 16. Mean waiting time (s) and utilization of the fog tier (m3) for selected policies.

Policy	CNC wait	CNC fog utilization	MIMII wait	MIMII fog utilization	C-MAPSS wait	C-MAPSS fog utilization
B4 semantic-only	0.324	0.94	1.555	0.97	1.230	0.85
B5 physical-only	0.330	0.95	3.314	1.01	5.598	1.00
B6 proposed	0.229	0.86	1.466	0.97	1.050	0.84
A5 queue-blind	106.431	1.52	82.858	1.29	5.610	0.92

The queue-blind variant reaches mean waiting times of 107 s and 83 s in the CNC and acoustic scenarios, with fog utilization above 1, which indicates that work remained queued at the end of the arrival horizon. The queue-aware proposed policy keeps mean waiting times to 0.23 s, 1.47 s and 1.05 s at the same arrival rates. H3 is supported within the finite horizon simulated; a statement about unbounded growth rests on the elementary stability condition (R3) and on the observed backlog, not on an infinite-horizon simulation. The table also shows that the queue-aware policies run the fog tier at utilizations of 0.84 to 0.97, close to saturation. This is a consequence of the cost weights and of the capacity assigned to the fog tier in Table 9, and it leaves little margin for load increases, as the sensitivity analysis shows.

6.7 Sensitivity analysis and null scenario (H2)

Table 17 reports the difference in the mean realized objective between the proposed policy and the semantic-only and physical-only policies (B6 minus B4 and B6 minus B5; negative favours the proposed policy), with ten replications of 3000 events per setting.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 17. Sensitivity of the advantage of the proposed policy: difference in mean J versus B4 / versus B5 (10 replications of 3000 events).

Setting	CNC	MIMII-structured	C-MAPSS-structured
κ = 1.0	−0.264 / −0.106	−0.019 / −0.266	−0.029 / −0.235
κ = 1.5	−0.267 / −0.097	−0.029 / −0.230	−0.026 / −0.231
κ = 2.5	−0.253 / −0.077	−0.035 / −0.199	−0.021 / −0.221
κ = 3.5	−0.228 / −0.072	−0.034 / −0.183	−0.019 / −0.201
Arrival multiplier 0.5	−0.191 / −0.059	−0.000 / −0.097	−0.000 / −0.066
Arrival multiplier 1.0	−0.253 / −0.077	−0.033 / −0.217	−0.020 / −0.210
Arrival multiplier 1.5	−0.316 / −0.102	−0.060 / −0.191	−0.148 / −0.193
Arrival multiplier 2.0	−0.394 / −0.126	−0.056 / −0.204	−0.182 / −0.173
σ_s = 0.02	−0.284 / −0.089	−0.039 / −0.221	−0.023 / −0.199
σ_s = 0.08	−0.253 / −0.077	−0.033 / −0.217	−0.020 / −0.210
σ_s = 0.20	−0.169 / −0.090	−0.026 / −0.236	−0.018 / −0.229
σ_s = 0.40	−0.029 / −0.167	−0.006 / −0.248	−0.013 / −0.220
w_d = 0	−0.309 / −0.117	−0.400 / −0.192	−0.066 / −0.151
w_d = 2	−0.302 / −0.131	−0.110 / −0.291	−0.032 / −0.213
w_d = 8	−0.142 / −0.067	−0.011 / −0.182	−0.013 / −0.195
Physical state irrelevant (null)	0.000 / −0.041	0.000 / −0.176	0.000 / −0.131

Four observations follow. First, the advantage over the semantic-only policy is not sensitive to the Weibull shape: it changes little as κ moves from 1.0 to 3.5, and it remains substantial in the CNC scenario at κ = 1, where by (R1) the penalty Ψ does not depend on h. In this simulation the benefit of physical state therefore arises mainly through the state-dependent deadline and not through the degree of wear-out. Setting w_d = 0, which removes the delay penalty from both the router and the realized objective, does not remove the advantage either. Second, the advantage grows with load, especially in the turbofan scenario (from −0.020 at the base rate to −0.182 at twice the rate). Third, in the CNC scenario the advantage over semantic-only routing falls from −0.253 to −0.029 as the noise in the state estimate rises from 0.08 to 0.40, which shows that the benefit depends on a reasonably accurate state estimate. Fourth, in the null scenario, where neither the deadline nor the penalty depends on the physical state, the proposed and semantic-only policies make identical decisions and the difference is exactly zero in all scenarios, as H2 predicts. The proposed policy remains better than the physical-only policy in that scenario because that policy discards semantic information. H2 is supported.

Cloud-only routing deteriorates sharply as the arrival rate rises. Its mean realized objective in the CNC scenario increases from 2.848 to 6.270 at 1.5 times the base rate and to 6.827 at twice the rate, and in the acoustic scenario from 1.196 to 3.839 at twice the rate, whereas the adaptive policies change far less.

6.8 Orchestration-slack analysis

Using the definitions of Section 3.10 with α_c = 0.90, 20,000 events per scenario (seed 555) and the observed z and $\hat{h}$, the share of events with negative slack, that is, with semantic capability demand above the latency ceiling, is shown in Table 18.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 18. Share of events with negative orchestration slack (Δ < 0), and mean deadline.

Scenario	Events with Δ < 0	Mean deadline τ (s)	Largest tier within deadline is the cloud model
CNC	17.3%	3.32	66.3% of events
MIMII-structured	0.1%	16.99	99.8% of events
C-MAPSS-structured	0.0%	36.22	100% of events

The ordering of these shares matches the ordering of the advantage over semantic-only routing in Table 14 (CNC largest, then acoustic, then turbofan). Three scenarios cannot establish a relationship, and the correspondence is reported as consistent with H1 and not as a test of it. It does indicate that in the acoustic and turbofan scenarios physical state seldom changes which model is feasible, so the advantage there comes mainly from load interactions and deadline tails.

6.9 Experiments on the original datasets (not performed)

Table 19 reserves the structure for experiments on the original datasets and on hardware. No values are reported because those experiments have not been performed.

Table 19. Reserved structure for validation on real data (no results available).

Dataset or setting	Required inputs	Metrics to be reported	Status
NASA C-MAPSS (original files)	Official data, a health-index map g, RUL or health model outputs per tier	RUL error, timely accuracy, latency percentiles, J	Not performed; placeholder
MIMII (original audio)	Official data, anomaly detectors per tier, measured inference times	Accuracy, F1-score, latency percentiles, J	Not performed; placeholder
Real CNC telemetry	Machine data, calibrated degradation map and deadlines	Deadline violation rate, J, router overhead	Not performed; placeholder
Hardware-in-the-loop	Edge, fog and cloud deployment, measured router overhead	End-to-end latency, router decision time	Not performed; placeholder

7. DISCUSSION

7.1 Interpretation of physio-semantic routing

The formulation separates two questions that a query-only router merges: which model is capable enough for this task, and which model is fast enough for this machine at this moment. In the simulation, the first question is answered by the semantic proxy and the second by the physical state through the deadline and the delay penalty. When the answers agree (non-negative slack), routing is a cost problem and the physical state adds little, as in the acoustic and turbofan scenarios. When they disagree (negative slack), the choice depends on the physical state, and a policy that cannot see it must either over-spend on slow models or accept missed deadlines. The results are consistent with this reading, with the qualification stated in Section 6.8 that three scenarios cannot establish a general relationship.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7.2 Comparison with existing orchestration approaches

The semantic-only baseline B4 is a cost-quality router that also observes queue state and is evaluated on the same objective. It is therefore a stronger comparator than a router that ignores load, but it is not a re-implementation of the published routers [7]–[10], whose inputs, training data and objectives differ. The results do not show that the proposed policy outperforms those methods. They show that, within a common simulated environment and objective, adding the physical state to a router of the same structure changes decisions and outcomes in the conditions identified above.

7.3 Edge versus fog versus cloud trade-offs

The three tiers fail in different ways. Cloud-only routing delivers the highest accuracy but misses many deadlines in the CNC scenario, and its mean realized objective increases several-fold as the arrival rate rises, because the cloud queue builds up during bursts. Edge-only routing meets deadlines but forfeits accuracy, particularly where tasks are difficult. The fog tier offers an intermediate trade-off and is where the adaptive policies concentrate most events, at the price of operating close to saturation. The utilizations in Table 16 show that the benefit of intermediate tiers depends on their provisioned capacity, a parameter that was fixed in this study (Table 9) and not optimized.

7.4 Effect of machine degradation

Two features of the degradation model need to be distinguished. The Weibull shape κ determines whether the delay penalty depends on the machine's state (R1), and the deadline function (6) determines whether the feasible set depends on it. The sensitivity analysis shows that the advantage persists when $\kappa = 1$ and when $w_d = 0$, so in this simulation the deadline channel carries most of the benefit. This does not make the degradation model unnecessary: for a deployment in which deadlines are fixed by contract and not by machine state, the Weibull penalty would be the only channel, and the present results do not speak to that case. The strong dependence on state-estimate noise (Table 17) shows that the quality of the health map g is a prerequisite.

7.5 Effect of semantic complexity

The physical-only policy, which ignores the semantic proxy, is clearly worse than the proposed policy in the acoustic and turbofan scenarios (about 23% and 20% higher realized objective), because it cannot match model capability to task difficulty. This is the converse of the argument for physical state and shows that neither information source is sufficient alone. The tier distribution of the capability demand (Section 6.8) shows that a majority of events in all scenarios demand at least the fog model for the 0.90 accuracy threshold, so that capability and not speed constrains many decisions.

7.6 Cross-domain applicability

The three scenarios give three regimes. In the CNC scenario deadlines are short compared with large-model latency and slack is often negative. In the acoustic scenario, deadlines are moderate and the choice is mainly one of cost and load. In the turbofan scenario, deadlines are long relative to inference, and the physical state matters mainly in the tails and under load. The common orchestration problem is retained across them, and the formulation reduces gracefully to cost-quality routing when the physical state is irrelevant (the null scenario). Whether the real datasets fall into these regimes depends on their actual time scales and on how deadlines are defined, which this study could not determine.

7.7 Theoretical implications

Result (R2) states when physical state has value: when the best model depends on it. The formulation makes that dependence explicit through two channels, the deadline and the delay penalty, and the experiments show that the two channels can behave differently. The slack descriptor (9)–(10) offers a way to predict, from a task distribution and a machine-state distribution, where physical state will matter, without running a router. These are modest results. No optimality or stability guarantee is claimed for the relaxed router, and the Lyapunov-based analysis suggested by [13] remains to be carried out.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7.8 Practical manufacturing implications

Three practical points follow from the formulation, subject to the limitations below. A deployment needs a calibrated map from machine signals to a degradation fraction and a documented rule for deadlines derived from process dynamics. The routing layer should observe the queue state of every tier, since queue-blind routing produced severe backlog in the simulation. Fog capacity should be planned with explicit utilization targets, because the best-performing policies pushed the fog tier close to saturation. Tier-1 supervision must remain operational independently of the higher tiers (Section 4.8).

7.9 Limitations

The study has several limitations that bound its conclusions. All data are synthetic surrogates, and neither the original C-MAPSS and MIMII files nor any CNC telemetry were used. The datasets contain no natural-language queries, so the semantic descriptor was generated from a latent difficulty with additive noise; real semantic features will relate to difficulty differently. The accuracy table of the router was trained on data from the same generator as the evaluation, so no distribution shift is represented. Service times, network delays, costs and capabilities are assumed values and not measurements; real language-model latency depends on output length, hardware and batching. The deadline function (6) and the Weibull parameters are assumptions, and the weights of the objective were fixed in advance and not tuned for any policy, including the baselines. The rule-based baseline used fixed thresholds that were not optimized, and a tuned rule could narrow the margins, as the near-parity of B3 and B4 in the acoustic scenario suggests. The model set has four models, arrival processes are Markov-modulated Poisson, and fleets of machines sharing queues are not modelled. Energy and RUL error are not evaluated. The router overhead is an assumed 0.5 ms and was not measured, and the simulation was written in Python, so no timing claim about the router can be drawn from it.

7.10 Threats to validity

Internal validity is threatened by the shared form of the router objective and the realized objective: a router that estimates the same function that is used for scoring is favoured over one that does not, and the baselines that share the form (B4, B5) are therefore the fair comparators, while the rule-based and fixed policies are disadvantaged by construction. Common random numbers remove noise from policy comparisons but do not remove this structural bias. Construct validity is limited because the realized objective is a weighted sum with weights chosen a priori, and different weights would change the rankings, as the w_d sensitivity results indicate. External validity is the main threat: conclusions about real manufacturing systems require validation on real data and hardware. Statistical conclusion validity is limited by the nature of the replications, which are draws from one generator and not independent systems, so the very small p-values reflect low simulation variance and not strength of evidence about the world. These threats are the reason this paper presents its results as an illustration of the formulation and not as evidence of its performance.

8. CONCLUSION AND FUTURE WORK

This paper formulated compute orchestration in industrial CPS as an expected-cost routing problem in which the physical state of the machine determines the response deadline and the cost of delay, through a Weibull degradation model and the conditional failure probability. It separated the capability a task demands from the latency ceiling a machine state allows, stated when physical state has value for routing, and described a three-tier architecture in which only the lowest tier acts on the physical process. In simulations on three synthetic surrogate scenarios, the proposed router reduced the mean realized objective relative to a semantic-only router by about 21% in a short-deadline CNC scenario and by about 4% and 2% in acoustic and turbofan scenarios; the advantage vanished when physical state was uninformative, depended on the accuracy of the state estimate, and arose in this setting mainly through state-dependent deadlines. Queue-blind routing produced severe backlog. The proposed targets of a 70% reduction in latency and cost with 98% accuracy retention were not met.

Future work falls into five groups. First, the evaluation should be repeated on the original C-MAPSS and MIMII data, using measured inference times for real models and a documented health map, and on real CNC telemetry. Second, the router overhead and the multi-tier deployment should be measured on hardware. Third, the learned

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

estimators, for example gradient-boosted trees over semantic features, and the weights of the objective should be evaluated, including Pareto analysis in place of a fixed weighted sum. Fourth, a formal stability analysis based on Lyapunov drift-plus-penalty methods [13] should be attempted. Fifth, a systematic literature search should be carried out to test the gap stated in Section 1.5, including the strands of work on acoustic anomaly detection and on agent-based manufacturing control that this draft did not cover.

DECLARATIONS

AI assistance: this draft was prepared with the assistance of an AI system. The simulation code was executed in a sandbox and its outputs are reported as produced. The author is responsible for verifying the content, the references and the code before submission. Funding, conflicts of interest and data availability statements are to be completed by the author. The simulation code and result files accompany the manuscript.

REFERENCES

[1] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for Industry 4.0-based manufacturing systems," Manufacturing Letters, vol. 3, pp. 18–23, Jan. 2015.
[2] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in Proc. Int. Conf. Prognostics and Health Management (PHM), Denver, CO, USA, Oct. 2008, doi: 10.1109/PHM.2008.4711414.
[3] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in Proc. Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York, NY, USA, Oct. 2019, pp. 209–213.
[4] L. Makatura et al., "How can large language models help humans in design and manufacturing?," arXiv:2307.14377, 2023 (preprint).
[5] Y. Li et al., "Large language models for manufacturing," arXiv:2410.21418, 2024 (preprint).
[6] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.
[7] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Rühle, L. V. S. Lakshmanan, and A. H. Awadallah, "Hybrid LLM: Cost-efficient and quality-aware query routing," in Proc. Int. Conf. Learning Representations (ICLR), 2024.
[8] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to use large language models while reducing cost and improving performance," Transactions on Machine Learning Research, 2024 (arXiv:2305.05176).
[9] Q. J. Hu, J. Bieker, X. Li, N. Jiang, B. Keigwin, G. Ranganath, K. Keutzer, and S. K. Upadhyay, "RouterBench: A benchmark for multi-LLM routing system," arXiv:2403.12031, 2024 (preprint).
[10] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in Proc. Conf. Language Modeling (COLM), 2024 (arXiv:2308.08155).
[11] W. Weibull, "A statistical distribution function of wide applicability," Journal of Applied Mechanics, vol. 18, no. 3, pp. 293–297, Sep. 1951, doi: 10.1115/1.4010337.
[12] M. J. Neely, Stochastic Network Optimization with Application to Communication and Queueing Systems. San Rafael, CA, USA: Morgan & Claypool, 2010.
[13] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," Mechanical Systems and Signal Processing, vol. 115, pp. 213–237, 2019.
[14] G. S. Babu, P. Zhao, and X.-L. Li, "Deep convolutional neural network based regression approach for estimation of remaining useful life," in Lecture Notes in Computer Science, Springer, 2016.
[15] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," Reliability Engineering & System Safety, vol. 172, pp. 1–11, 2018, doi: 10.1016/j.ress.2017.11.021.
[16] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," Journal of Computational Physics, vol. 378, pp. 686–707, 2019, doi: 10.1016/j.jcp.2018.10.045.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Appendix A. Figure Specifications

No figure has been drawn. The following specifications are intended for conversion into publication-quality diagrams.

Figure 1. Overall physio-semantic orchestration architecture. A left-to-right block diagram. On the left, two input sources: an alarm and query source (text) and a sensor stream (vibration, current, temperature, acoustic or engine channels). Both feed a feature stage with two sub-blocks, semantic features producing the proxy z and physical features producing the degradation estimate ĥ. These feed the central router block, which also receives queue state from the three model tiers. The router outputs a dispatch arrow to one of the model blocks m1 to m4, grouped into edge, fog and cloud regions. A separate safety function block, independent of the router, connects the tier-1 region to the machine. Return arrows lead from the models to a supervisory block, with advisory outputs drawn as dashed arrows. Caption below the figure.

Figure 2. Physio-semantic routing pipeline. A vertical flow chart with the steps of Section 4.4: obtain z and ĥ; read the earliest-free times; estimate waits by equation (7); for each model compute L^_j, Ψ, the miss probability and J_j; select the minimum; dispatch; execute; log outcome. A side annotation shows the O(K) cost of the loop.

Figure 3. Three-tier multi-agent edge-fog-cloud architecture. Three horizontal bands labelled edge, fog and cloud, each containing the agent of Table 3. Vertical arrows show events moving up and advisory outputs moving down. A boundary marks the region where physical actuation is permitted (tier 1 only), with the safety function drawn beside it.

Figure 4. Mathematical optimization and routing workflow. A diagram that connects the constrained problem (13) to its relaxation (11): the four constraints on the left (quality, deadline, load, overhead), arrows into the four weighted terms of the objective, and the argmin output. A second panel shows the slack construction: the capability demand C_sem and the latency ceiling C_phys on a common tier axis, with the conflict region where $\Delta < 0$ shaded.

Figure 5. Simulation and evaluation framework. A block diagram of the experiment: three surrogate generators feed common random numbers to the eight policies; a discrete-event engine with four server pools produces per-event outcomes; metrics are aggregated over 30 replications; paired statistics and the sensitivity and null experiments follow.

Appendix B. Reproducibility

The simulation is implemented in Python with NumPy. The file pso_sim.py contains the generators, the policies and the discrete-event engine, and is run with the argument main for the 30-replication experiments (seeds 1000 to 1029, 4000 events per scenario) and sens for the sensitivity analysis (10 replications of 3000 events). The file analyze.py computes the tables and the paired statistics, and conflict.py computes the orchestration-slack shares. The files main_results.json and sens_results.json contain the raw per-replication metrics that underlie Tables 10 to 18. The training data of the accuracy table use seed 987654, and the slack analysis uses seed 555.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.