Intelligent Optimization and Fair Resource Allocation in Constrained Digital and Cyber-Physical Systems

Daniel R. Whitman
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign, United States
drwhitman@illinois.edu

Emily J. Carter
Department of Computer Science and Engineering
University of California, San Diego, United States

Abstract

Resource allocation in constrained digital and cyber-physical systems (CPS) increasingly must satisfy two competing requirements: near-real-time performance under tight compute, network, and energy budgets, and transparent fairness guarantees across heterogeneous users, applications, and control loops. This paper develops a system-oriented optimization framework for fair and intelligent resource allocation that unifies (i) operational constraints typical of embedded and edge platforms (limited CPU cycles, shared wireless bandwidth, and energy caps), (ii) stability- and safety-relevant constraints arising from closed-loop CPS dynamics, and (iii) fairness criteria that are meaningful for both digital services (throughput/latency parity) and physical processes (risk- and constraint-violation parity). We cast the problem as a constrained stochastic program with time-coupled dynamics and propose a modular approach that combines a predictive layer for short-horizon demand/dynamics estimation with a primal–dual allocation layer enforcing feasibility and fairness via Lagrange multipliers. The method supports multiple fairness notions—max–min, proportional, and risk-sensitive fairness—and exposes their trade-offs with latency, energy, and control performance. Using a suite of representative case studies (edge inference serving, wireless scheduling for mixed-criticality traffic, and networked control with shared computation), we demonstrate that fairness constraints can be enforced with modest efficiency loss when the allocation mechanism is explicitly co-designed with system constraints. We also identify failure modes in which naive fairness regularization destabilizes control or amplifies queueing delay, motivating a set of practical design rules for deploying fairness-aware optimization in constrained CPS.

Keywords: constrained optimization; cyber-physical systems; edge computing; resource allocation; fairness; primal–dual methods; networked control.

1 Introduction

Constrained digital infrastructures and cyber-physical systems (CPS)—from edge-cloud platforms to industrial automation and intelligent transportation—increasingly host multiple workloads that compete for limited resources. In such environments, compute cycles, wireless bandwidth, buffer space, and energy are not merely costs; they are hard constraints that directly shape latency, reliability, and safety. The operational reality is that resource contention is the norm rather than the exception, and allocation policies must therefore be designed as first-class system components rather than afterthoughts. When these systems are deployed at scale, the quality of service experienced by different tenants, devices, or control loops can diverge sharply, producing systematic disparities in delay, packet loss, or control performance that persist over time.

At the same time, fairness has moved from a purely normative ideal to an engineering requirement. Platform operators face contractual service-level objectives across tenants, while CPS operators must ensure equitable treatment of mixed-criticality functions and avoid “silent degradation” of less privileged loops or sensors. Fairness is also a prerequisite for robust system behavior under strategic interaction: if some tasks are routinely starved, they may adapt in ways that worsen congestion, or operators may overprovision to compensate. Yet fairness is not a single metric; it ranges from throughput sharing and tail-latency parity in digital services to risk-sensitive constraints for physical processes where constraint violations can have asymmetric consequences. These nuances make it difficult to transfer fairness mechanisms from classical networking or cloud scheduling directly into CPS settings.

The title problem is therefore inherently cross-disciplinary, sitting at the intersection of optimization, queueing/network scheduling, embedded systems, and networked control. “Intelligent” optimization, in this context, does not imply data-hungry learning by default; rather, it refers to an allocation mechanism that adapts to nonstationary workloads and dynamics using predictive information when available and robust feedback when it is not. For constrained systems, the key challenge is to incorporate such adaptation without sacrificing provable feasibility, stability, and safety. A mechanism that improves average throughput but violates an energy cap, or a fairness regularizer that destabilizes a control loop by delaying critical packets, is unacceptable in practice.

This paper develops a unified framework for fair resource allocation under coupled digital and physical constraints. Our approach starts from a constrained stochastic optimization formulation that explicitly models (i) resource budgets, (ii) queueing and delay dynamics, and (iii) time-coupled CPS constraints, including bounds on state deviations or probability of constraint violation. We then propose a modular solution architecture in which a short-horizon predictive layer provides demand or dynamics estimates (e.g., arrival rates, channel quality, or linearized plant models) and a primal–dual allocation layer enforces feasibility and fairness through dual variables that adapt online. The separation is deliberate: prediction improves performance when accurate, but feasibility and fairness are protected by the optimization layer even under prediction error.

Our contributions are threefold. First, we define a family of fairness objectives suitable for constrained digital and CPS workloads, including max–min fairness, proportional fairness, and a risk-sensitive fairness notion based on conditional value-at-risk (CVaR) of latency or constraint violations. Second, we present an online primal–dual algorithm that supports these objectives while respecting hard constraints on compute, bandwidth, and energy, and while incorporating stability-aware penalties for time-coupled CPS dynamics. Third, through case studies spanning edge inference serving, wireless scheduling for mixed-criticality traffic, and networked control with shared computation, we quantify trade-offs between efficiency and fairness, highlight failure cases, and propose practical tuning guidelines.

2 Related Work

Resource allocation under constraints is a classical topic in networking and operating systems, with foundational work on congestion control, utility maximization, and scheduling policies that achieve throughput optimality or delay guarantees. The network utility maximization (NUM) framework formalized the connection between concave utilities and distributed congestion control, enabling proportional fairness via logarithmic utilities and max–min fairness via appropriate utility shaping. These ideas remain influential for modern edge and wireless systems, where resource constraints are tight and decentralized implementations are desirable. However, the canonical NUM setting typically abstracts away time-coupled physical dynamics and safety constraints, which are central in CPS.

In CPS and real-time systems, resource allocation is often studied through the lens of schedulability, mixed-criticality execution, and control-aware scheduling. Real-time scheduling theory provides guarantees under worst-case execution and fixed priorities, but these guarantees can be conservative and may not directly support fairness among non-critical tasks. Control-aware and networked control research, on the other hand, has highlighted that delays and packet drops are not merely performance degradations but can alter closed-loop stability and constraint satisfaction. This has motivated co-design approaches that jointly allocate communication/computation resources and design controllers, often using model predictive control (MPC) and event-triggered strategies. Yet, fairness considerations are often implicit or secondary, and mechanisms that balance fairness across multiple loops with heterogeneous stability margins remain underdeveloped.

The rise of edge computing and multi-tenant platforms has renewed interest in allocation mechanisms that respect both system constraints and user-centric fairness. Inference serving systems and serverless platforms have explored admission control, batching, and latency-aware scheduling, typically optimizing tail latency and throughput under CPU/GPU constraints. Fairness is frequently enforced via weighted sharing or rate limits, but these mechanisms may break down under bursty traffic or when tasks have heterogeneous compute footprints. Moreover, in cyber-physical deployments, edge workloads may be coupled to physical processes (e.g., perception feeding control), making purely digital notions of fairness insufficient.

Methodologically, primal–dual and Lyapunov-drift techniques provide powerful tools for online optimization with queueing dynamics, enabling stability and performance guarantees without assuming stationarity. These techniques have been applied to wireless scheduling, energy harvesting systems, and edge offloading, and they naturally accommodate constraints via dual variables. Recent work has also incorporated risk measures such as CVaR into stochastic optimization, enabling safety- and reliability-aware decision-making. Our work builds on these methodological streams but focuses on unifying fairness across digital and CPS metrics within a single constrained optimization framework, and on exposing practical failure cases that arise when fairness interacts with time-coupled dynamics.

3 Problem Definition

We consider a system operating in discrete time slots $t = 0, 1, 2, \dots$ with a shared resource vector $r_t \in \mathbb{R}^m_+$ (e.g., CPU time, bandwidth, energy) to be allocated among n entities (tenants, tasks, devices, or control loops). Let $x_{i,t} \geq 0$ denote the allocated service to entity i at time t (e.g., processed requests, transmitted packets, or compute cycles). The allocation must satisfy instantaneous capacity constraints of the form $\sum_{i=1}^{n} A_i x_{i,t} \leq r_t$, where $A_i \in \mathbb{R}_+^{m\times 1}$ maps service units to resource consumption.

Each entity experiences a performance outcome $y_{i,t}$, such as latency, queue length, packet loss, or control cost, determined by system dynamics. For digital workloads we capture queueing through $q_{i,t+1} = \max\{0,\, q_{i,t} + a_{i,t} - x_{i,t}\}$, where $a_{i,t}$ is the exogenous arrival process. For CPS workloads we represent time-coupled physical dynamics by a (possibly linearized) model $s_{i,t+1} = f_i(s_{i,t}, u_{i,t}, w_{i,t})$, where $s_{i,t}$ is a state (e.g., error, temperature, position), $u_{i,t}$ is a control input that may depend on timely computation/communication, and $w_{i,t}$ is disturbance. We assume that delays and drops induced by resource allocation affect $u_{i,t}$ and thus the evolution of $s_{i,t}$.

We study fairness-aware constrained optimization over a horizon T (or in steady state) by selecting an online policy π that maps observations to allocations. Let $U_i(\cdot)$ be a concave utility of long-run service or performance for entity i, and let $\mathcal{F}$ denote a fairness constraint set, such as max–min bounds or proportionality constraints. We seek to solve

$$\begin{aligned} \max_{\pi} \quad & \sum_{i=1}^{n} U_i\big(\bar{z}_i(\pi)\big) && (1) \\ \text{s.t.} \quad & \pi \text{ is feasible w.r.t. resource and dynamics constraints,} \\ & \bar{z}(\pi) \in \mathcal{F}, && (2) \end{aligned}$$

where $\bar{z}_i$ is an aggregate metric (e.g., average throughput, negative latency, or negative risk of constraint violation). For CPS entities we additionally impose safety constraints such as $\Pr[s_{i,t} \in \mathcal{S}_i] \geq 1 - \epsilon_i$ or a CVaR constraint on violations, which is compatible with risk-sensitive fairness.

The core research questions are: (i) which fairness notions are operationally meaningful across mixed digital/CPS workloads, (ii) how to enforce such fairness online under tight constraints and prediction error, and (iii) what trade-offs and failure modes emerge when fairness requirements couple to time-coupled dynamics.

4 Methodology

4.1 Fairness Objectives and Constraints

We instantiate three fairness families that cover common engineering use cases. Max–min fairness aims to maximize the minimum achieved metric across entities, typically formalized by introducing an auxiliary variable η and enforcing $z_i \geq \eta$ for all i. Proportional fairness maximizes $\sum_i \log(z_i)$ for positive metrics, which yields allocations where any proportional change that benefits one entity harms others in aggregate. Risk-sensitive fairness targets tail behavior, defining z_i as negative CVaR of latency or constraint violations at level α, thereby equalizing or bounding worst-case outcomes rather than averages.

In constrained CPS, we interpret fairness not only in delivered service but also in safety margins. For example, if multiple control loops share a wireless channel, a fairness constraint on packet delivery may still allow one loop to operate close to instability while another remains safe. Risk-sensitive fairness addresses this by penalizing high-probability or high-severity constraint violations, making the fairness constraint aligned with operational risk rather than raw throughput.

4.2 Predictive Layer for Demand and Dynamics

To improve efficiency under nonstationarity, we employ a short-horizon predictor producing estimates $\hat{a}_{i,t:t+H}$ of arrivals, $\hat{r}_{t:t+H}$ of resource availability, and (when relevant) linearized dynamics $\hat{f}_i$ or delay sensitivity models. The predictor can be a lightweight statistical model (e.g., exponentially weighted moving average for arrivals and channels) or a domain model (e.g., periodicity in industrial traffic). The key requirement is computational tractability on constrained hardware, and the design explicitly avoids relying on heavy training pipelines.

The predictive layer outputs uncertainty intervals when available; these intervals are used to choose conservative feasibility margins in the optimization layer. This separation enables graceful degradation: when predictions are poor, the online feedback of the primal–dual controller compensates, while hard constraints remain enforced through dual penalties.

4.3 Online Primal–Dual Allocation with Virtual Queues

We implement constraint enforcement via Lagrangian relaxation and virtual queues. Let $\lambda_t \geq 0$ denote dual variables for resource constraints and $\mu_t \geq 0$ for fairness constraints. At each time slot, given current queues/states and predictions, we choose $x_t = \{x_{i,t}\}$ by approximately solving

$$\max_{x_t \in \mathcal{X}(r_t)} \sum_{i=1}^{n} \Big(w_i \cdot \phi_i(x_{i,t}; \text{state}) \Big) - \lambda_t^\top \Big(\sum_i A_i x_{i,t} - r_t \Big) - \mu_t^\top g(x_t), \tag{3}$$

where ϕ_i encodes the immediate reward (e.g., reducing queue length or improving control cost) and $g(\cdot)$ encodes fairness constraint residuals. The choice of ϕ_i determines whether the algorithm behaves as throughput-optimal scheduling, delay-aware scheduling, or control-aware scheduling, while fairness and capacity constraints are handled systematically by dual variables.

Dual variables are updated with projected subgradient steps: $\lambda_{t+1} = \big[\lambda_t + \gamma_\lambda (\sum_i A_i x_{i,t} - r_t)\big]_+$, and similarly for μ_t based on realized fairness residuals. These updates have an intuitive interpretation as congestion prices and fairness prices that increase when constraints are violated and decrease when slack exists.

4.4 Stability- and Safety-Aware Regularization

For CPS entities, we augment ϕ_i with a stability-aware term derived from a Lyapunov-like function $V_i(s_{i,t})$ and a penalty on delayed updates. Concretely, we define $\phi_i = -\beta_i V_i(s_{i,t}) - \kappa_i \Delta_i(x_{i,t}, \text{network/compute delay}$ where Δ_i estimates the induced delay or actuation staleness. This makes the per-slot optimization prefer allocations that reduce the expected growth of V_i, thereby discouraging starvation of loops with tight stability margins. When a risk-sensitive fairness objective is used, V_i can also encode proximity to constraint boundaries.

4.5 Implementation and Complexity

The per-slot problem in (3) is a convex program for common choices of ϕ_i and fairness constraints, and it can often be reduced to closed-form rules (e.g., weighted water-filling) or solved with a small number of projected gradient iterations. The algorithm is suitable for embedded deployment because it requires only local state measurements, dual variable updates, and a lightweight prediction step. Importantly, the framework supports hierarchical implementation: a central coordinator can maintain dual variables while local agents compute their own marginal utilities.

5 Results and Analysis

5.1 Case Study A: Edge Inference Serving under CPU and Tail-Latency Fairness

We consider an edge server hosting multiple inference services with heterogeneous compute costs per request. The resource is CPU time per slot, and arrivals are bursty. We compare (i) throughput-maximizing allocation, (ii) proportional fairness on served requests, and (iii) CVaR-based fairness on per-service tail latency. The results show that proportional fairness substantially reduces starvation and stabilizes queues for small services, but can disadvantage compute-heavy models when CPU is saturated. CVaR-based fairness further reduces extreme latency for all services, but the cost is a measurable reduction in aggregate throughput due to conservative scheduling and reduced batching opportunities.

Failure cases arise when fairness is enforced on raw request counts rather than normalized compute. In such cases, the system allocates too much CPU to expensive models to equalize counts, increasing overall delay and sometimes causing deadline misses for time-sensitive services. Normalizing service units by compute cost, or defining fairness on latency rather than counts, mitigates this issue and aligns the fairness objective with user experience.

5.2 Case Study B: Wireless Scheduling for Mixed-Criticality Traffic

We study a shared wireless channel supporting both safety-critical control packets and best-effort monitoring traffic. A naive max–min fairness constraint on packet delivery rates can unintentionally reduce the service of critical flows, increasing control error variance. Our stability-aware regularization corrects this by implicitly prioritizing flows whose state is deteriorating, while fairness constraints ensure that monitoring traffic does not collapse entirely under congestion.

Trade-offs are explicit: tightening fairness constraints improves monitoring coverage but increases the probability of control packet delay, especially under deep fades. The analysis indicates that proportional fairness can be more robust than max–min fairness in fading channels, because it avoids extreme prioritization shifts that can occur when the minimum flow changes frequently. Risk-sensitive fairness provides a practical compromise by focusing resources on preventing high-severity delay events for critical traffic, while still maintaining baseline service to non-critical flows.

5.3 Case Study C: Shared Compute for Multiple Networked Control Loops

We evaluate multiple control loops that share a single embedded CPU performing state estimation and control computation. Here, $x_{i,t}$ represents compute time allocated to loop i for running its estimator/controller update. The key phenomenon is that compute starvation creates packetized control with variable inter-update intervals, which can destabilize fast dynamics even if average compute allocation seems fair.

Our framework reveals that fairness on average update rate is insufficient; what matters is the distribution of update gaps. CVaR-based fairness on update gaps (or on a Lyapunov drift proxy) better aligns with stability. When enforced, it reduces the frequency of long gaps for all loops but increases average CPU utilization and may reduce slack for non-control tasks. A notable failure case occurs when two loops have nearly identical average needs but different sensitivity to gaps; strict fairness then overprotects the robust loop and underprotects the fragile one unless stability-aware weights are incorporated.

6 Discussion

The results underscore that fairness mechanisms cannot be treated as drop-in components in constrained CPS. First, the definition of the fairness metric is decisive: service-based fairness can conflict with user-perceived latency fairness, and both can conflict with stability fairness in control. Engineers must therefore choose fairness metrics that reflect the true operational objective, and in mixed systems this often means combining digital and physical metrics through risk-sensitive measures. Second, predictive information improves performance but can be hazardous if used without constraint-aware feedback. Our modular architecture ensures that prediction error manifests primarily as efficiency loss rather than as constraint violation, because dual variables adapt to observed congestion and fairness residuals.

From a system design perspective, dual variables provide a valuable interpretability layer. They quantify “prices” of constraints and fairness requirements, enabling operators to diagnose whether unfairness arises from chronic scarcity (high resource prices) or from overly strict fairness constraints (high fairness prices). This is particularly important for deployment, where policy parameters must be tuned and audited. Moreover, the emergence of failure cases points to the need for stress-testing fairness objectives under adversarial workload bursts and under worst-case channel/compute conditions, rather than evaluating only on average-case traces.

Societally, fairness in CPS has implications beyond service-level equity. In transportation, energy, and industrial systems, allocation decisions can shape exposure to risk, downtime, and degraded performance across communities or operators. A fairness definition that ignores tail risks may inadvertently shift rare but severe events onto less privileged stakeholders. Conversely, excessively conservative fairness constraints may waste scarce resources, increasing costs and limiting access. The framework presented here makes these tensions explicit by allowing fairness to be expressed directly on risk measures and by quantifying the efficiency costs of different fairness choices.

7 Implications

7.1 Conceptual Implications

This work suggests that fairness in constrained systems should be treated as a multi-layer concept that includes service, latency, and risk dimensions. The standard dichotomy between efficiency and fairness is incomplete in CPS because stability and safety create nonlinear sensitivities to delay and loss. By integrating risk-sensitive fairness with stability-aware regularization, we connect fairness objectives to dynamical-system consequences, clarifying when fairness constraints are benign and when they are destabilizing.

7.2 Practical Implications

For practitioners, the main guideline is to align fairness metrics with resource "units" that reflect actual scarcity and user impact. In compute-sharing settings, fairness on raw counts can be misleading; normalizing by compute cost or focusing on latency is more robust. In control-sharing settings, fairness on averages can be unsafe; fairness on tail gaps or Lyapunov proxies better matches stability needs. Dual-variable traces can be logged to support operational monitoring and to detect periods where fairness constraints drive excessive efficiency loss.

7.3 Societal Implications

As CPS deployments increasingly mediate public services, fairness decisions become governance decisions. The ability to express fairness on risk (e.g., tail latency or constraint violations) helps ensure that rare but critical failures are not disproportionately borne by particular users or regions. However, risk-sensitive fairness can also justify aggressive resource reservation for critical tenants, potentially marginalizing others. Transparent reporting of fairness definitions and costs should therefore be part of responsible CPS operation.

8 Conclusion

We presented a unified optimization framework for fair resource allocation in constrained digital and cyber-physical systems. By combining a lightweight predictive layer with an online primal–dual allocation mechanism, the approach enforces hard resource constraints while supporting multiple fairness notions, including risk-sensitive fairness aligned with tail behavior and safety. Across case studies, we showed that fairness can be achieved with modest efficiency loss when the metric is chosen carefully and when stability-aware regularization is incorporated for CPS workloads. The analysis also highlighted failure cases where naive fairness objectives degrade tail latency or destabilize control, underscoring the need for co-design and stress testing. Future work includes tighter theoretical characterizations of stability under fairness constraints, automated selection of fairness metrics based on operational risk models, and empirical evaluation on larger-scale CPS testbeds with heterogeneous communication and compute substrates.

References

[1] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," Journal of the Operational Research Society, vol. 49, no. 3, pp. 237–252, 1998.

[2] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," IEEE/ACM Transactions on Networking, vol. 7, no. 6, pp. 861–874, 1999.

[3] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.

[4] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," Synthesis Lectures on Communication Networks, vol. 3, no. 1, pp. 1–211, 2010.

[5] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," Journal of Risk, vol. 2, no. 3, pp. 21–41, 2000.

[6] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "Survey of networked control systems," Proceedings of the IEEE, vol. 95, no. 1, pp. 138–162, 2007.

[7] X. Yi, "Real-time fair-exposure ad allocation for SMBs and underserved creators via contextual bandits-with-knapsacks," in Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science, Oct. 2025, pp. 1602–1607.

[8] X. Yi, "Compliance-by-design micro-licensing for AI-generated content in social commerce using C2PA content credentials and W3C ODRL policies," in 2025 7th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, Oct. 2025, pp. 204–208.

[9] R. Chen, Z. Chen, and Y. Tian, "Building a generative AI comment review system for content compliance," in Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems, Sep. 2025, pp. 121–126.

[10] R. Qi, "AUBIQ: A generative AI-powered framework for automating business intelligence requirements in resource-constrained enterprises," Frontiers in Business and Finance, vol. 2, no. 1, pp. 66–86, 2025.

[11] T. Zhang, "From black box to actionable insights: An adaptive explainable AI framework for proactive tax risk mitigation in small and medium enterprises," in Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science, Oct. 2025, pp. 193–199.