

AN ADAPTIVE WIP CONTROLLER FRAMEWORK FOR DYNAMIC BUFFER MANAGEMENT IN MODULAR CONSTRUCTION

Nima Moghim¹, Sahar H. Shamaee², Haitao Yu³, Qipei Mei⁴, Vicente A. Gonzalez⁵, & Farook Hamzeh⁶

ABSTRACT

Lean construction principles emphasize maintaining production flow and managing variability to eliminate waste. However, in modular construction, high product mix and strict spatial constraints frequently disrupt this flow, rendering static buffering strategies ineffective. To address this challenge, this paper proposes an Adaptive WIP Control Framework structured in three integrated layers: a Digital Twin for stochastic simulation, an optimization layer for policy development, and a real-time execution layer. At the core of this framework is a novel hysteretic "Store-and-Release" control logic that dynamically regulates Work-in-Process (WIP) based on real-time buffer inventory, transforming the buffer from a passive queue into an active decoupling mechanism. The control logic was evaluated through a simulation-based case study of a prefabricated wall assembly line. Results demonstrate that the adaptive controller reduces downstream workload by 22% compared to unbuffered flow, effectively eliminating starvation waste. Unlike static time-buffers which sacrifice upstream speed for downstream stability, the adaptive approach optimizes total system performance without physical footprint expansion. This research contributes to Lean construction theory by establishing dynamic inventory control as a superior alternative to static scheduling for maintaining reliable flow in off-site construction.

KEYWORDS

Lean Construction, flow, work in progress/process (WIP), buffer/buffer management, modular construction.

¹ Ph.D. Student, Department of Civil and Environmental Engineering, University of Alberta, Canada, nmoghim@ualberta.ca, orcid.org/0009-0008-4733-1276

² Graduate Research Assistant, Department of Civil and Environmental Engineering, University of Alberta, Canada, hamedsha@ualberta.ca

³ R&D Lead, Landmark Homes, Edmonton, Canada, haitao@landmarkgroup.ca

⁴ Assistant Professor, Department of Civil and Environmental Engineering, University of Alberta, Canada, qipei@ualberta.ca

⁵ Hal Kvisle Professor and Tier 1 Canada Research Chair in Digital Lean Construction, Infrastructure Human Tech (IHT) Lab, Strategic Projects Insight Centre in Engineering (SPICE), Department of Civil and Environmental Engineering, Faculty of Engineering, University of Alberta, Canada, vagonzal@ualberta.ca, orcid.org/0000-0003-3408-3863

⁶ Professor, Department of Civil and Environmental Engineering, University of Alberta, Canada, hamzeh@ualberta.ca, orcid.org/0000-0002-3986-9534

INTRODUCTION

Modular construction factories operate in a unique flow-line environment where performance is heavily constrained by the intersection of high production variability and strict spatial limitations (Alsakka et al., 2024). Unlike traditional manufacturing, modular factories contend with variability from make-to-order customization (Moghadam & Al-Hussein, 2014; Ritter & Al-Hussein, 2017), dimensional deviations (Shahtaheri et al., 2017), and fluctuating resource availability (Bhatia et al., 2019; Wang et al., 2018). These factors interact with the large physical footprint of volumetric modules, which severely limits the available floor area for Work-in-Process (WIP). Consequently, factories often operate with minimal inter-station buffers, making the line highly sensitive to blocking and starvation (Weiss & Stolletz, 2015). Inadequate buffer sizing under these conditions amplifies variability, leading to frequent stoppages and reduced throughput (Salhab et al., 2023).

To mitigate these inefficiencies, the industry often turns to Lean philosophy and the Transformation-Flow-Value (TFV) theory (Koskela, 2000). While the goal of Lean is variability elimination, practitioners acknowledge that inherent fluctuations in construction must be strategically managed rather than simply removed (Hopp & Spearman, 2011; Tommelein, 2020). To prevent destabilization, Lean advocates for "Pull" production control mechanisms (e.g., Kanban or Constant Work-in-Process (CONWIP)) over traditional "Push" systems. By authorizing work release only when downstream capacity is available, Pull systems strictly limit WIP accumulation to maintain flow without consuming excessive space or capital (Arbulu, 2006; Tommelein, 2020). However, implementing these systems requires determining precise buffer levels that balance throughput against inventory costs—a complex optimization problem dependent on system configuration and variability (El Sakka et al., 2016; González et al., 2009).

Buffer management has been extensively investigated to solve this challenge, predominantly using Discrete Event Simulation (DES) to manage the trade-off between cycle time and inventory (González & Alarcón, 2009; Gupta et al., 2012). However, a fundamental limitation of current approaches is their reliance on steady-state assumptions. Studies typically derive a single, static "optimal" buffer size based on long-run averages (Alves et al., 2006; Ritter et al., 2017). While effective in stable manufacturing, this approach fails in modular construction, where high-mix, make-to-order conditions create a non-stationary production environment (Kasper et al., 2023). Furthermore, existing mathematical models or project-level tools often ignore the specific "shop floor" dynamics where a buffer is not just a number in a schedule, but a finite physical footprint (Weiss & Stolletz, 2015). In this context, static buffer approaches become unresponsive; a limit that prevents starvation for simple modules may cause severe blocking during complex, space-intensive jobs.

Consequently, there is a pressing need for an adaptive buffer management approach that can translate short-term production plans into responsive shop-floor control policies. To address this gap, this study proposes an adaptive WIP control approach for a selected space-constrained post-bottleneck buffer in modular construction. Although the studied production line contains multiple intermediate buffers and transfer points, the proposed controller is applied to the critical interface between the upstream and downstream production segments, where poor control can simultaneously trigger upstream blocking and downstream starvation. The core contribution is a hysteretic "Store-and-Release" logic that regulates downstream processing based on the real-time inventory accumulated in this buffer. Rather than relying on a fixed time delay or a single static buffer rule, the proposed controller activates downstream processing only after sufficient WIP has accumulated and continues processing until the buffer is drained, thereby reducing starvation while avoiding unnecessary upstream blocking. As a proof of concept, the controller is evaluated through offline discrete-event simulation using a representative case study from a prefabricated timber wall assembly line in Edmonton, Canada.

Sensitivity analysis is then used as a methodological proxy (Moghimi & Kashani, 2026) to identify effective WIP threshold values under realistic production variability.

METHODOLOGY

This study adopts a Design Science Research (DSR) methodology, appropriate for developing and evaluating a prescriptive solution to a practical production-control problem (Hevner et al., 2004). The study followed three main DSR activities: problem identification, controller design, and evaluation. Importantly, these activities were not strictly linear. The research began with a broader concept for an adaptive WIP control framework, but preliminary simulation-based assessment showed that the most critical and practically implementable contribution in the present context was the post-bottleneck release logic itself. Accordingly, the artifact was refined during the study to focus on a hysteretic “Store-and-Release” controller and its threshold-selection mechanism. The final evaluation therefore assesses the refined controller artifact against existing buffering practices in the case factory.

The problem identified in this study is not merely the existence of limited buffer space, but the inability of the current control method to respond to short-term production variability at a critical interface in the factory. In the case facility, the buffer between the upstream transfer-cart segment and the downstream Loading station is governed in practice by a fixed time-based separation rule. Because upstream panel completion and downstream project-level loading progress vary significantly across projects, this static policy can create two undesirable conditions: (1) upstream blocking when the physical buffer saturates, and (2) downstream starvation when insufficient work is available for Loading. The design objective of the proposed artifact is therefore to replace this static rule with a responsive inventory-based control logic that regulates release using the real-time WIP level in the post-bottleneck buffer.

In the present study, the proposed method is implemented and evaluated through offline simulation rather than through a fully deployed real-time digital control system. The simulation model is used to represent factory stochastic dynamics and test candidate WIP threshold policies under a representative production schedule. In this sense, the contribution of this paper is the controller logic and its performance evaluation, while broader real-time digital integration remains a future implementation pathway (Moghimi et al., 2026). Figure 1 illustrates the structure of the framework. The following sections describe how this methodology is operationalized through framework development, controller logic design, and simulation-based evaluation.

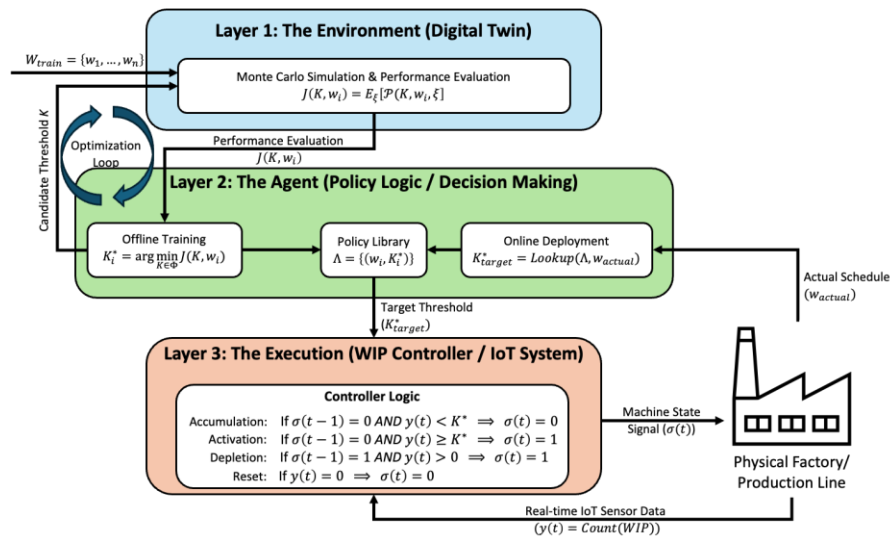


Figure 1: Structure of the Adaptive WIP Controller Framework.

FRAMEWORK STRUCTURE

The proposed method consists of three functional components: (1) an offline simulation environment used to represent stochastic factory behaviour, (2) a policy evaluation step used to assess candidate WIP thresholds, and (3) an execution logic that applies the selected threshold to regulate material release at the post-bottleneck buffer. In this paper, only the controller logic and its offline simulation-based evaluation are implemented. Real-time sensing and autonomous policy updating are discussed as future extensions rather than as implemented system components.

LAYER 1: THE ENVIRONMENT

The Environment component represents the production system through an offline discrete-event simulation model used to evaluate candidate WIP control policies under stochastic operating conditions. In general, a production schedule is modelled as a short-term release plan specifying the sequence and composition of jobs entering the system over a defined planning horizon. Depending on the production context, each job may differ in size, processing requirements, routing characteristics, or expected workload contribution. As a result, different schedules generate different arrival patterns and workload fluctuations at the controlled buffer.

Let $\mathcal{W}$ denote the infinite space of all theoretically possible production schedules. From this universal space, a finite training set of representative schedules $\mathcal{W}_{\text{train}} \subset \mathcal{W}$, is curated to span the diversity of actual factory operations:

$$\mathcal{W}_{\text{train}} = \{w_1, w_2, \dots, w_n\} \quad (1)$$

For any given schedule sample $w_i \in \mathcal{W}_{\text{train}}$ and a candidate buffer capacity configuration K , the Environment generates Monte Carlo realizations of stochastic variability ξ (e.g., processing time distributions, machine reliability) and evaluates the expected system performance through the cost function J :

$$J(K, w_i) = \mathbb{E}_{\xi}[\mathcal{P}(K, w_i, \xi)] \quad (2)$$

where $\mathcal{P}$ denotes a scalar system performance metric used to evaluate control policies. In this study, this metric is defined as total downstream cycle time, which directly reflects starvation-induced inefficiencies at the constrained station.

In practical terms, for a given schedule and threshold K , the simulation is run multiple times under different stochastic realizations, and the average result is used to estimate controller performance. This allows the analyst to assess how different threshold values perform under varying production conditions and uncertainty realizations. Each schedule sample therefore corresponds to a distinct short-term workload pattern imposed on the controlled buffer. For example, one schedule may concentrate high-workload jobs early in the planning horizon, while another may distribute work more evenly, resulting in different buffer accumulation and depletion dynamics. In the present study, the Environment is used offline to evaluate alternative WIP thresholds and identify effective policies under representative production conditions.

LAYER 2: THE AGENT (POLICY OPTIMIZATION AND MAPPING)

The Agent acts as a policy selection and execution mechanism that bridges offline simulation-based analysis with real-time shop-floor control. In this study, the term Agent refers to a decision-making entity responsible for selecting and enforcing pre-optimized control policies, rather than an online learning algorithm. While the overall framework is compatible with reinforcement learning-based agents, no reinforcement learning algorithm is implemented or trained in this paper. Instead, policy optimization is performed offline through simulation-based sensitivity analysis, and the Agent operates as a policy retrieval mechanism during production. Its operation unfolds through two phases:

Phase A: Offline Training (Policy Library Construction). During offline training, the Agent systematically evaluates candidate buffer policies using the Environment to construct a Policy Library (Λ). For each representative schedule w_i , the Agent solves for the buffer capacity that minimizes expected system cost:

$$K_i^* = \arg \min_{K \in \Phi} J(K, w_i) \quad (3)$$

where Φ represents the feasible set of buffer sizes constrained by physical floor space. The result is a learned policy mapping:

$$\Lambda = \{(w_1, K_1^*), (w_2, K_2^*), \dots, (w_n, K_n^*)\} \quad (4)$$

Phase B: Online Deployment (Real-Time Inference). During production, the Agent receives the actual schedule w_{actual} . It performs rapid schedule classification to identify the closest training sample $w_i \approx w_{\text{actual}}$ and retrieves the corresponding pre-optimized policy:

$$K_{\text{target}}^* = \text{Lookup}(\Lambda, w_{\text{actual}}) \approx K_i^* \quad (5)$$

This instantaneous retrieval eliminates real-time computational overhead, enabling low-latency deployment.

It should be noted that in future implementations, this policy-selection mechanism can be replaced by a fully autonomous reinforcement learning agent (e.g., policy-gradient or actor-critic methods) to enable continuous online adaptation under non-stationary conditions.

LAYER 3: THE EXECUTION (IoT-ENABLED REAL-TIME CONTROL)

The Execution layer operationalizes the selected control policy through real-time enforcement on the shop floor. During production, real-time data collection is limited to tracking entity entry and exit events at the controlled post-bottleneck buffer. In practice, this can be implemented using simple industrial sensing mechanisms such as photoelectric sensors, light curtains, proximity sensors, or RFID/barcode gates, depending on factory layout and panel geometry. Each detected entry or exit event increments or decrements the buffer inventory count $y(t)$, which serves as the sole real-time state variable required by the controller.

The IoT controller continuously compares the real-time WIP level $y(t)$ against the Agent-derived threshold K_{target}^* and generates binary control signals that govern downstream station activation and work-release authorization, as detailed in the following section. This establishes a closed-loop control structure, in which physical production events update the digital system state and digital control decisions are immediately enforced on the physical production line.

WIP CONTROLLER LOGIC

To operationalize the threshold K^* derived by the Agent, a novel "Store-and-Release" hysteresis logic governs material flow. Unlike static Kanban caps, this controller enforces a strict accumulation phase to buffer variability followed by a depletion phase to maximize utilization.

Let $\sigma(t) \in \{0,1\}$ denote the downstream station state (0=Idle, 1=Active) and $y(t)$ the buffer inventory. The state transition rules are:

1. **Accumulation (Enforced Idle):** Processing is withheld until inventory reaches K^* .
If $\sigma(t-1) = 0$ AND $y(t) < K^* \Rightarrow \sigma(t) = 0$ (6)

2. **Activation:** The station transitions to active status only when the threshold is met.
If $\sigma(t-1) = 0$ AND $y(t) \geq K^* \Rightarrow \sigma(t) = 1$ (7)

3. **Depletion (Hysteresis Latching):** Processing continues below K^* to prevent start-stop chattering.

$$\text{If } \sigma(t-1) = 1 \text{ AND } y(t) > 0 \Rightarrow \sigma(t) = 1 \quad (8)$$

4. **Reset:** The cycle terminates only when the buffer is fully drained.

$$\text{If } y(t) = 0 \Rightarrow \sigma(t) = 0 \quad (9)$$

This logic adapts dynamically to relative station capacities. If the downstream station is the bottleneck, the threshold condition $y(t) \geq K^*$ is naturally satisfied, keeping the logic transparent for maximum utilization. Conversely, if the downstream station has excess capacity, the logic actively enforces batch accumulation, decoupling variability and preventing starvation-induced fragmentation.

Operationally, the controller behaves like a gated release rule with hysteresis. For example, suppose the selected threshold is $K=8$ panels. If only 5 panels are currently waiting in the post-bottleneck buffer, the downstream station remains idle and additional panels are allowed to accumulate. Once the buffer inventory reaches 8 panels, downstream processing is activated. Importantly, the station does not stop as soon as inventory drops below 8; instead, it continues processing until the buffer is fully drained. This hysteretic behavior avoids repeated start-stop switching and ensures that downstream work proceeds in stable bursts rather than fragmented, starvation-prone intervals. In this way, the buffer acts as an active decoupling mechanism rather than a passive storage space.

FRAMEWORK EVALUATION

While the proposed framework envisions future integration of autonomous agents and IoT technologies, rigorous evaluation of the Hysteretic WIP Controller logic in a controlled environment is required prior to full-scale deployment. In this context, IoT integration refers to simple shop-floor sensing of panel entry and exit events at the controlled buffer using photoelectric sensors, proximity sensors, or RFID/barcode gates. These event-driven signals increment or decrement the WIP count and provide the sole real-time input required by the controller. Accordingly, “real-time” in this study denotes responsiveness at the scale of production events (seconds to minutes), rather than continuous or high-frequency data acquisition.

Consistent with best practices in construction operations research, simulation-based evaluation was used to validate controller behavior before operational implementation (AbouRizk et al., 1993; Law et al., 2007). The evaluation was conducted using a discrete-event simulation of a wall assembly line at a prefabricated timber construction facility in Edmonton, Canada. The selected line is representative of modular construction environments due to (1) heterogeneous processing times arising from hybrid manual–automated operations, (2) severely constrained buffer capacities, and (3) high product variability driven by client customization.

The selected facility manufactures volumetric residential modules and fabricates wall, roof, floor, and finishing components. The wall assembly line was chosen as the study focus because it exhibits three representative characteristics typifying production environments with inherent variability: (1) hybrid automation–manual operations combining machines with labor-intensive tasks, producing heterogeneous workstations with widely varying processing times; (2) severely restricted buffer capacities between consecutive stations, constraining Work-in-Process inventory space; and (3) high product variety driven by client customization, resulting in specification-dependent process time variability. These conditions collectively create the dynamic, space-constrained production environment for which the adaptive WIP controller is designed.

SIMULATION MODEL DEVELOPMENT

The wall assembly line consists of Framing, Sheeting, CNC Nailing, Packaging, Doors & Windows (DW) installation, and Loading. Panels flow through single- and multi-capacity buffers, including four parallel DW lanes (16-panel capacity) and a six-lane Loading buffer (24-panel capacity). A discrete-event simulation model was implemented in Python using the SimPy framework to capture stochastic process dynamics, resource contention, and buffer

constraints. FIFO discipline is enforced at stations with physical sequencing constraints, while semi-FIFO behavior is implemented downstream through conditional routing logic reflecting observed shop-floor practices. Figure 2 illustrates this flow.

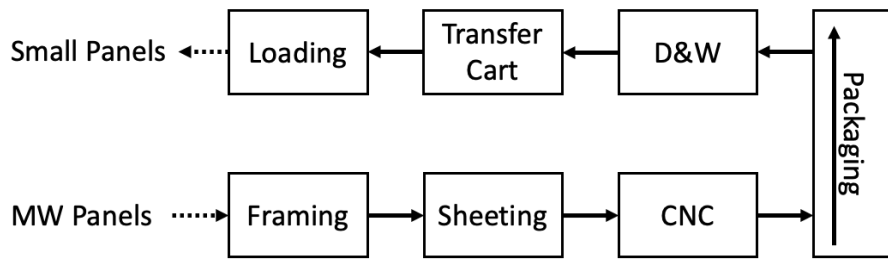


Figure 2: A schematic representation of the assembly line, illustrating the sequence of workstations and production flow.

An event-driven discrete-event simulation model was developed in Python using the SimPy framework to capture the stochastic dynamics of the wall assembly line (Brown et al., 2019). Process flow, resource contention, and buffer constraints are modeled using single- and multi-capacity SimPy Store objects, enforcing FIFO discipline where sequencing is physically constrained and semi-FIFO behavior downstream via conditional routing logic. Single-capacity stores represent transfer elements, while multi-capacity stores represent larger buffers (e.g., DW: 16 panels; Loading: 24 panels). Blocking and starvation emerge endogenously from these constraints, and dynamic routing at the Packaging station assigns panels to DW lanes based on panel type and queue availability, consistent with observed shop-floor behavior. Model inputs were derived from field observations of 84 multiwall and 135 small panels, with task durations modeled using triangular distributions. For each scenario, 50 independent simulation replications were conducted to ensure stable estimation of performance measures at a 95% confidence level (Moghimi & Kashani, 2025). Validation against observed production data yielded Mean Absolute Percentage Errors (MAPE) of 16–18% across Framing, DW, and Loading stations, which is acceptable for comparative policy evaluation.

WIP CONTROLLER IMPLEMENTATION

The wall assembly line contains multiple intermediate buffers and transfer points across Framing, Sheeting, CNC Nailing, Packaging, Doors & Windows (DW) installation, and Loading. However, the proposed WIP controller is applied only to the interface between upstream production (Framing through DW) and the Loading station. This location was selected because it constitutes the critical decoupling point in the studied system. Loading is constrained by external logistics (trailer availability and dock capacity), limiting throughput to three projects per day and establishing it as the project-level bottleneck. In contrast, the upstream transfer cart acts as an internal bottleneck at the panel level. The facility currently employs a static one-day time buffer to decouple these segments. Because this buffer is temporal rather than inventory-based, it does not adapt to project-specific variability, leading to WIP accumulation and upstream blocking when production exceeds buffer capacity, or Loading starvation when production falls short. This behaviour provides direct evidence of variability propagation across the buffer when time-based separation is used (Hopp & Spearman, 2011). The buffer between the transfer cart and Loading therefore constitutes the critical control point for implementing adaptive WIP regulation.

EXPERIMENTAL DESIGN AND SENSITIVITY ANALYSIS

The evaluation utilizes a representative one-week production schedule, comprising 14 projects with heterogeneous specifications. The one-week horizon is deliberately selected to align with standard lookahead planning practices used in the case study factory, where production

commitments and release decisions are typically made on a weekly basis. Although the evaluation window spans five working days, individual projects extend across multiple days, ensuring that the system experiences overlapping work, queue accumulation, and dynamic release decisions rather than daily resets.

The 14 projects were selected to reflect the range of production variability encountered in practice. Project size varied across the weekly schedule, with projects ranging from 7 to 15 multiwall panels. Projects also differed in their exterior (EXT) and interior (INT) panel composition, which is operationally significant because these panel types impose different workloads on different stations. EXT panels require sheeting after framing, whereas INT panels do not; therefore, projects with a higher share of EXT panels generate more work at the Sheeting and CNC stations. EXT panels also tend to impose higher demand at the Doors and Windows (DW) station, since many require door or window installation. In contrast, INT panels primarily increase workload at the Packaging station. At the Loading stage, processing time is influenced mainly by the number of panels in each project, since each project is loaded onto a separate trailer and larger projects increase loading complexity under limited-space conditions. As a result, projects with similar total panel counts can still generate different workload patterns across the line, producing a non-uniform arrival pattern at the controlled buffer.

To quantify the controller's effectiveness at decoupling the segments, two complementary performance metrics were defined:

- **Upstream Segment Performance (J_{upstream}):** Total Cycle Time from system entry (Framing) to the target buffer entry (before Loading), measuring the duration required for all projects in the weekly schedule to complete upstream operations. This metric reveals whether the controller successfully prevents upstream blocking and maintains upstream productivity.
- **Downstream Segment Performance ($J_{\text{downstream}}$):** Total Cycle Time from buffer exit to Loading completion, measuring throughput efficiency and starvation frequency at the constrained Loading operation. This metric reveals whether the controller maintains adequate WIP feed to prevent Loading underutilization.

These dual perspectives enable assessment of whether the proposed controller achieves the dual objective of protecting upstream operations from downstream constraints while ensuring sufficient work feed to maintain downstream utilization.

The feasible set of buffer thresholds is bounded below the physical buffer capacity (i.e., 24 panels): $\Phi = \{1, 2, 3, \dots, 20\}$ panels making sure upstream doesn't face blocking; Sensitivity analysis explores performance across this entire domain. To contextualize the WIP controller's performance gains, simulation results were compared against three control policies:

- **Baseline 1 – Continuous Production (No Decoupling, $K = 0$):** Models the scenario where no buffer or decoupling mechanism exists; upstream and downstream operations are tightly coupled. This baseline establishes theoretical lower bounds and highlights the necessity for buffering.
- **Baseline 2 – Current Practice (Time-Based Buffering):** Replicates the plant's existing one-day time buffer approach, where downstream work release is governed by a fixed temporal schedule rather than inventory status. This baseline quantifies performance under current operations.
- **Proposed Controller (Optimal K^*):** The Hysteretic WIP Controller with the threshold identified through sensitivity analysis as delivering superior performance, demonstrating the performance achievable through dynamic inventory-based control.

RESULTS AND DISCUSSION

This section reports the results of the simulation-based experimental design described previously. Three control scenarios—continuous flow, current time-based buffering, and the proposed adaptive WIP controller—are evaluated under identical production conditions using a representative one-week schedule. Performance is assessed using upstream and downstream cycle time metrics across multiple stochastic replications.

A sensitivity analysis evaluated the impact of the WIP gate threshold $K \in \Phi = \{1, 2, \dots, 20\}$ on system performance. For each threshold value, 50 independent Monte Carlo simulation replications were executed, sampling stochastic processing times from calibrated triangular distributions. This replication level was selected based on variance stabilization tests, which showed no material change in mean estimates or relative policy rankings beyond this point (Law et al., 2007). Two performance indicators were recorded: upstream cycle time (J_{upstream}) and Downstream Cycle Time ($J_{\text{downstream}}$).

SENSITIVITY ANALYSIS AND OPTIMAL POLICY SELECTION

Figure 3 presents the sensitivity analysis results, showing mean cycle time and one-standard-deviation variability for both upstream and downstream operations. A distinct global optimum is observed at $K^* = 15$, where downstream cycle time reaches a minimum of approximately 17.7 hours with controlled variability. Notably, upstream cycle time remains nearly constant at approximately 32 hours across the entire threshold domain. This indicates that, because the adaptive thresholds remain below the physical buffer capacity (24 panels), downstream stabilization is achieved without inducing upstream blocking. As a result, the controller simultaneously maximizes downstream utilization while preserving upstream productivity.

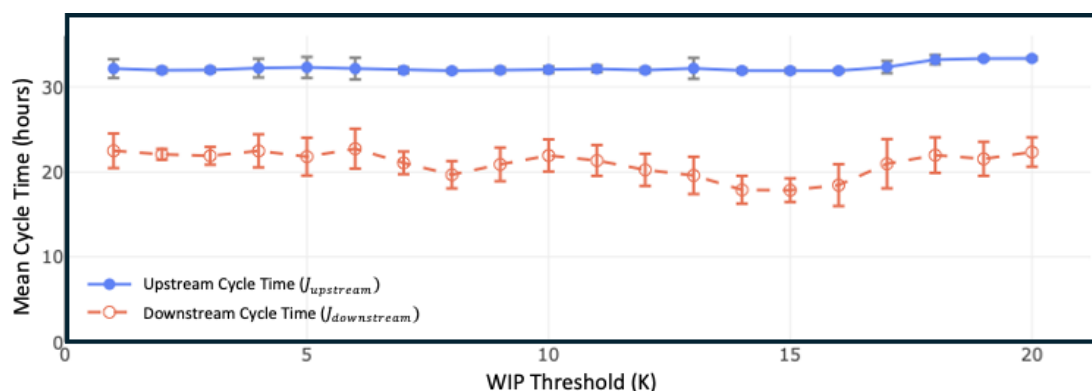


Figure 3: Result of sensitivity analysis illustrating the impact of the WIP gate threshold on the mean and variability of upstream and downstream cycle times.

PERFORMANCE EVALUATION AGAINST OPERATIONAL BASELINES

The operational effectiveness of the optimal threshold $K^* = 15$ was validated against two distinct control paradigms (Figure 4), using the mean upstream and downstream execution timelines obtained from 50 independent Monte Carlo simulation replications for each policy: Under immediate release (no decoupling), upstream completion is rapid (~ 32 hours), but downstream operations suffer from severe starvation, increasing active Loading time to approximately 22 hours. In contrast, the current practice of static time-based buffering stabilizes Loading (~ 17 hours) but saturates the physical buffer, triggering upstream blocking and increasing upstream cycle time to 37 hours—exceeding the facility’s four-day production target.

The proposed adaptive WIP controller achieves dual optimization by matching the downstream efficiency of the static buffer (17.2 hours) while maintaining the upstream speed

of the unbuffered scenario (31.8 hours). This demonstrates that dynamic, inventory-based control can effectively decouple production segments without the performance penalties inherent to static scheduling.

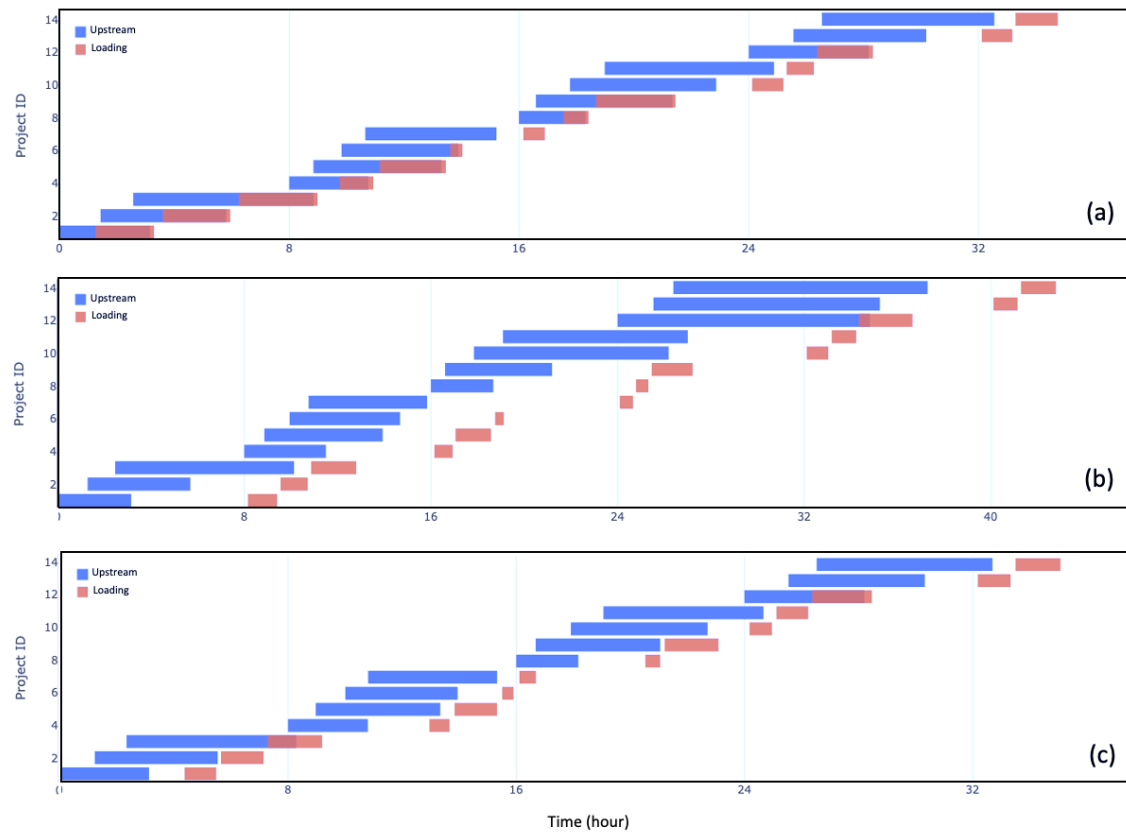


Figure 4: Mean Project execution timelines for upstream and downstream of the targeted buffer, aggregated across 50 Monte Carlo simulation replications, comparing three buffering approaches: (a) no decoupling, (b) static 8-hour buffer, and (c) optimal WIP threshold control ($K^* = 15$).

DISCUSSION AND PRACTICAL IMPLICATION

The comparative analysis indicates that the Hysteretic WIP Controller transforms the buffer from a passive queue into an active decoupling mechanism, reducing downstream starvation while avoiding upstream blocking. The observed 22% reduction in active Loading workload (from 22 hours to 17.2 hours) creates operational benefits beyond throughput, enabling more predictable idle windows that support crew multi-skilling and ancillary activities such as maintenance or quality checks. Simultaneously, reducing upstream cycle time from 37 hours under static buffering to 31.8 hours mitigates chronic overtime and associated fatigue. These results confirm that buffer performance in modular construction is highly sensitive to short-term workload variability and cannot be effectively managed using fixed time-based buffers. By relying solely on real-time WIP counts obtained through simple entry–exit sensing, the proposed approach offers a practical and implementable alternative to static buffering, aligning with and extending prior work based on steady-state assumptions (e.g., Hopp and Spearman, 2011; Alves et al., 2006).

This finding can be interpreted in relation to earlier lean-construction research on buffer management. Prior IGLC-related studies have shown that buffers play an important role in protecting flow and absorbing variability and have mainly examined how static buffer sizing affects project or production performance (e.g., Alves et al., 2006; González & Alarcón, 2009;

Gupta et al., 2012). The present study is consistent with that body of work in recognizing the value of buffers as flow-stabilizing mechanisms but extends it by treating the buffer not as a fixed design parameter, but as an actively controlled decoupling point governed by an inventory-based release rule. In this sense, the contribution is not simply identifying an appropriate amount of buffering, but showing that under short-term, non-stationary modular production conditions, dynamic threshold-based control can outperform static time-based buffering while preserving upstream flow.

CONCLUSION AND FUTURE WORK

This study addressed the challenge of managing production variability in modular construction environments, where high product mix and spatial constraints limit the effectiveness of static buffering strategies. An adaptive WIP control approach was proposed, centered on a hysteretic store-and-release logic for regulating a selected post-bottleneck buffer under variable production conditions. Evaluation through a simulation-based case study of a prefabricated timber wall assembly line showed that the proposed controller outperforms both continuous production and the facility's current time-based buffering practice. The adaptive approach reduced active workload at the bottleneck Loading station by approximately 22% while preserving upstream production speed, thereby avoiding the zero-sum trade-off inherent in static time buffers. These findings indicate that dynamic inventory-based control offers a promising alternative to static scheduling rules for space-constrained off-site construction systems. From a practical perspective, the proposed logic enables reliable decoupling of production stages using only real-time WIP information, supporting more predictable logistics integration and improved flow stability.

This study has several limitations. The evaluation is based on a single industrial case study, focuses on one selected control point rather than coordinated control across multiple buffers, and assesses the proposed logic through offline simulation rather than field implementation. The findings should therefore be interpreted as a proof of concept in a representative modular production environment. Future work will focus on extending the approach to broader production settings, evaluating coordinated multi-buffer control, and exploring more autonomous policy-selection mechanisms for non-stationary production conditions.

ACKNOWLEDGEMENT

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance grant ALLRP 581074–22, Discovery Grant RGPIN-2020-04420, and NSERC Alliance—Alberta Innovates Advance Program 232403286. All findings and conclusions expressed in this paper are those of the authors and do not reflect those of the contributors.

REFERENCES

- Alsakka, F., Yu, H., El-Chami, I., Hamzeh, F., & Al-Hussein, M. (2024). Digital twin for production estimation, scheduling and real-time monitoring in offsite construction. *Computers & Industrial Engineering*, 191, 110173. <https://doi.org/10.1016/j.cie.2024.110173>
- Arbulu, R. (2006). Application of pull and CONWIP in construction production systems. *Proceedings of the 14th Annual Conference of the International Group for Lean Construction, Santiago, Chile*, 25–27.
- Bhatia, A. P. S., Han, S., Moselhi, O., Lei, Z., & Raimondi, C. (2019). Data Analytics of Production Cycle Time for Offsite Construction Projects. *Modular and Offsite Construction (MOC) Summit Proceedings*, 25–32.

- Brown, R., Barkokebas, B., Ritter, C., & Al-Hussein, M. (2019). Predicting performance indicators using BIM and simulation for a wall assembly line. *Proceedings of the 27th Annual Conference of the International Group for Lean Construction (IGLC 2019)*, Dublin, Ireland, 3–5.
- da CL Alves, T., Tommelein, I. D., & Ballard, G. (2006). *SIMULATION AS A TOOL FOR PRODUCTION SYSTEM DESIGN IN CONSTRUCTION*.
- El Sakka, F., Eid, K., Narciss, T., & Hamzeh, F. (2016). Integrating lean into modular construction: A detailed case study of company X. *Proceedings of the 24th Annual Conference of the International Group for Lean Construction*, 23–33.
- González, V., & Alarcón, L. F. (2009). A methodology for integrated buffer design and management in repetitive construction projects. *Proceedings of the 17th Annual Conference of the International Group for Lean Construction*, 393–406.
- González, V., Alarcón, L. F., & Molenaar, K. (2009). Multiobjective design of work-in-process buffer for scheduling repetitive building projects. *Automation in Construction*, 18(2), 95–108.
- Gupta, A., Gonzalez, V., & Miller, G. (2012). Understanding the relationship between productivity and buffers in construction: A simulation-based case. *Proceeding IGLC-20: Twentieth Annual Conference of the International Group for Lean Construction*.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75–105.
- Hopp, W. J., & Spearman, M. L. (2011). *Factory physics*. Waveland Press.
- Kasper, T. A. A., Land, M. J., & Teunter, R. H. (2023). Non-hierarchical work-in-progress control in manufacturing. *International Journal of Production Economics*, 257, 108768.
- Koskela, L. (2000). *An exploration towards a production theory and its application to construction*. VTT Technical Research Centre of Finland.
- Law, A. M., Kelton, W. D., & Kelton, W. D. (2007). *Simulation modeling and analysis* (Vol. 3). McGraw-hill New York.
- Moghadam, M., & Al-Hussein, M. (2014). An enhanced scheduling technique for modular construction manufacturing. *Proceedings of the International Group for Lean Construction (IGLC-22)*.
- Moghim, N., Arshang, O., & Hamzeh, F. (2026). Digital twins in offsite construction: Current implementations, challenges, and future pathways. *Automation in Construction*, 183, 106820.
- Moghim, N., & Kashani, H. (2025). A probabilistic framework to design the financing scheme for resilience enhancement of infrastructure systems. *Reliability Engineering & System Safety*, 262, 111179. <https://doi.org/10.1016/j.ress.2025.111179>
- Moghim, N., & Kashani, H. (2026). Probabilistic Modeling to Develop the Optimal Seismic Resilience Enhancement Strategy for Industrial Facilities. *Journal of Infrastructure Systems*, 32(1), 04025041. <https://doi.org/10.1061/JITSE4.ISENG-2754>
- Ritter, C., & Al-Hussein, M. (2017). Planning of modular construction manufacturing plant layouts using non-linear optimization. *Modular and Offsite Construction (MOC) Summit Proceedings*.
- Ritter, C., Zhang, Y., Dupuis, R., & Al-Hussein, M. (2017). Simulation of production line improvement in modular home manufacturing. *Proceedings of the 25th Annual Conference of the International Group for Lean Construction*, 737–744.
- Salhab, D., Alsakka, F., & Hamzeh, F. (2023). A narrative review of workspace planning in construction: Challenges and insights. *Proc., 31st Annual Conf. Int. Group for Lean Construction (IGLC31)*, 1208–1219.
- Sargent, R. G. (2010). Verification and validation of simulation models. *Proceedings of the 2010 Winter Simulation Conference*, 166–183.

- Shahtaheri, Y., Rausch, C., West, J., Haas, C., & Nahangi, M. (2017). Managing risk in modular construction using dimensional and geometric tolerance strategies. *Automation in Construction*, 83, 303–315.
- Tommelein, I. D. (2020). Takt: The parade of trades: Use of capacity buffers to gain work flow reliability. *Annual Conference of the International Group for Lean Construction*.
- Wang, Z., Hu, H., & Gong, J. (2018). Framework for modeling operational uncertainty to optimize offsite production scheduling of precast components. *Automation in Construction*, 86, 69–80.
- Weiss, S., & Stollitz, R. (2015). Buffer allocation in stochastic flow lines via sample-based optimization with initial bounds. *OR Spectrum*, 37(4), 869–902.