HYMAS: A HYBRID MULTI-AGENT FRAMEWORK FOR AUTONOMOUS STEEL PRODUCTION PLANNING

Anton Ivanov, Ralf Lenz, Heiko Wolf– PSI Software SE

Steel is a cornerstone of global infrastructure but its production faces inherent complexity due to high energy consumption, strict quality demands, and sophisticated production process. While traditional planning methods rely on rigid, centralized frameworks, they struggle to adapt to disruptions like equipment failures or material defects, leading to inefficiencies and financial losses. These complexities require autonomous, intelligent solutions that rapidly generate and propose optimized schedules, enabling swift, data-driven rescheduling in response to disruptions. To address this, we present HyMAS, a Hybrid Multi-Agent System, which enables dynamic, resilient decision-making in steel manufacturing. HyMAS employs a hierarchical network of autonomous agents: resource agents govern separate lines, process agents manage entire production stages, and process chain agents coordinate sequential operations. This modular structure allows localized adjustments to production plans without requiring full-scale re-planning, significantly reducing decision time and increasing flexibility. Integrated with Industry 4.0 frameworks, the system interfaces with Manufacturing Execution Systems, leverages digital twins to simulate disruption impacts, and applies machine learning to predict material quality and process mining to monitor performance. In a case study involving a cold rolling mill, three annealing lines, and intermediate storage, HyMAS achieved an 18% efficiency gain and an 8% improvement in on-time delivery. These results demonstrate the system's ability to improve operational performance while maintaining high levels of adaptability. By leveraging autonomous, intelligent agents, HyMAS rapidly generates and proposes optimized schedules, empowering manufacturers to mitigate financial risks, reduce decision time, and enhance operational agility. Additionally, planning adjustments can now be implemented much more swiftly than before, ensuring a more resilient and responsive production environment. By gradually integrating autonomous, intelligent rescheduling into production planning, HyMAS unlocks the potential of real-time data to systematically enhance efficiency, responsiveness, and operational resilience.

KEYWORDS: SCHEDULING – PLANNING – MULTI-AGENT SYSTEM – DIGITAL TWIN – HYMAS

INTRODUCTION: HEADING

Steel production underpins global infrastructure, yet its manufacturing processes remain among the most complex and disruption-prone in the industry. Production is highly energy-intensive — a single continuous annealing line can consume up to 400 MWh per day — and must meet stringent quality standards while coordinating tightly sequenced operations under strict delivery deadlines. Traditional planning in steel plants relies on centrally generated, deterministic schedules. These static plans are slow to adapt when confronted with unplanned events such as equipment breakdowns, material defects, or urgent order changes, often triggering manual, time-consuming re-planning. The result is efficiency loss, missed delivery targets, and increased operational risk.

The increasing complexity of supply chains and the volatility of customer demand underscore the need for planning systems that are adaptive, autonomous, and resilient. Industry 4.0 principles — digitalization, decentralized decision-making, and real-time analytics — create an opportunity to replace static scheduling with systems capable of self-adjustment. Multi-agent systems (MAS), digital twins, and AI-driven analytics have emerged as promising enablers, allowing distributed agents to negotiate, cooperate, and respond locally to disturbances without triggering full-scale re-planning.

Previous research has demonstrated that MAS architectures can enhance flexibility, fault tolerance, and responsiveness in industrial scheduling [1]. Building on these foundations, this work presents HyMAS — a Hybrid Multi-Agent System that integrates the adaptability of autonomous agents with the precision of mathematical optimization models. HyMAS autonomously generates and adjusts production schedules in near real time, guided by a digital twin that simulates the impact of disruptions and by optimization models that balance objectives such as tardiness minimization and reduction of dummy coils ("stringers").

Our industrial focus is the cold-rolling and continuous annealing stage, where efficient scheduling is crucial both for meeting downstream delivery commitments and for avoiding waste from unnecessary stringer use. The HyMAS framework employs a hierarchical agent model — resource, process, and process-chain agents — to localize decision-making while maintaining global coherence. This is coupled with a Parallel Heterogeneous Annealing Lines Scheduling (PHALS) that explicitly models multi-mode processing options and operational constraints [2].

The remainder of this paper first defines the steel production scheduling problem and outlines the specific operational challenges encountered in cold-rolling and annealing stages. It then describes the HyMAS architecture and its integration with a digital twin, followed by an explanation of the PHALS optimization model and how it is embedded within the agent framework. Next, the paper presents the simulation-based validation and case-study results. Finally, it concludes with a discussion of the contributions to resilient autonomous scheduling and perspectives for future research.

PROBLEM STATEMENT: COLD-ROLLING AND ANNEALING SCHEDULING CHALLENGES

The production setting considered here consists of a cold rolling mill followed by several continuous annealing lines — three in the case study. After cold rolling, coils must be annealed to achieve the specified material properties, using either batch furnaces or continuous annealing lines (CALs). In continuous annealing, coils are welded end-to-end into a single strip that passes through a high-temperature furnace at a constant speed. This arrangement enables high throughput but also imposes strict sequencing constraints: consecutive coils must be compatible in physical properties and process requirements. Large differences in annealing temperature profiles, thickness, or width can prevent direct welding.

When incompatibilities occur, a dummy coil — known as a stringer — must be inserted between the coils to preserve continuous furnace operation. Stringers are not part of any order and their use is undesirable: they consume material and energy, reduce throughput, and require extra post-processing for removal.

Scheduling cold-rolled coils onto multiple CALs is therefore a multi-criteria optimization problem. Each coil has a due date and may offer alternative processing modes — for example, different annealing temperature–speed combinations that meet the required metallurgical properties. The scheduler must determine:

- **Line assignment** – which CAL will process each coil.
- **Sequencing** – the order of coils on each line.
- **Mode selection** – which processing mode to apply when alternatives are available.

These decisions must satisfy key objectives and constraints:

- **Efficiency** – maximize continuous operation and minimize stringer insertions.
- **Tardiness minimization** – meet delivery deadlines or reduce lateness.
- **Flexibility** – accommodate real-time changes such as breakdowns or urgent orders.

The problem is NP-hard, combining features of parallel machine scheduling, sequence-dependent setups (welding compatibility), and due-date constraints [2]. Its difficulty is compounded by the frequency of disruptions in steel plants: a CAL may fail, a strip may break, or a high-priority order may arrive unexpectedly. Such events can instantly render a static plan infeasible. For example, if one line fails, its coils must be reassigned to the remaining lines; if an urgent order appears, it must be inserted without disproportionately

delaying others.

Addressing this challenge requires dynamic rescheduling — the ability to revise the plan in real time, mitigating the impact of disruptions while preserving overall efficiency and delivery performance. The combination of combinatorial complexity and operational uncertainty motivates a solution that blends optimization for high-quality plans with an agent-based architecture for decentralized, rapid responsiveness.

SYSTEM ARCHITECTURE: HYBRID MULTI-AGENT SYSTEM (HYMAS)

The HyMAS architecture is a hierarchical and modular multi-agent system designed to support real-time, resilient scheduling in steel production. Its structure mirrors the physical and organizational structure of the production process, with decision-making distributed across several layers. The system is connected to the plant's digital infrastructure through an open API and interacts directly with the Manufacturing Execution System (MES), a plant digital twin, and external services such as optimization algorithms, predictive analytics, and process-mining tools.

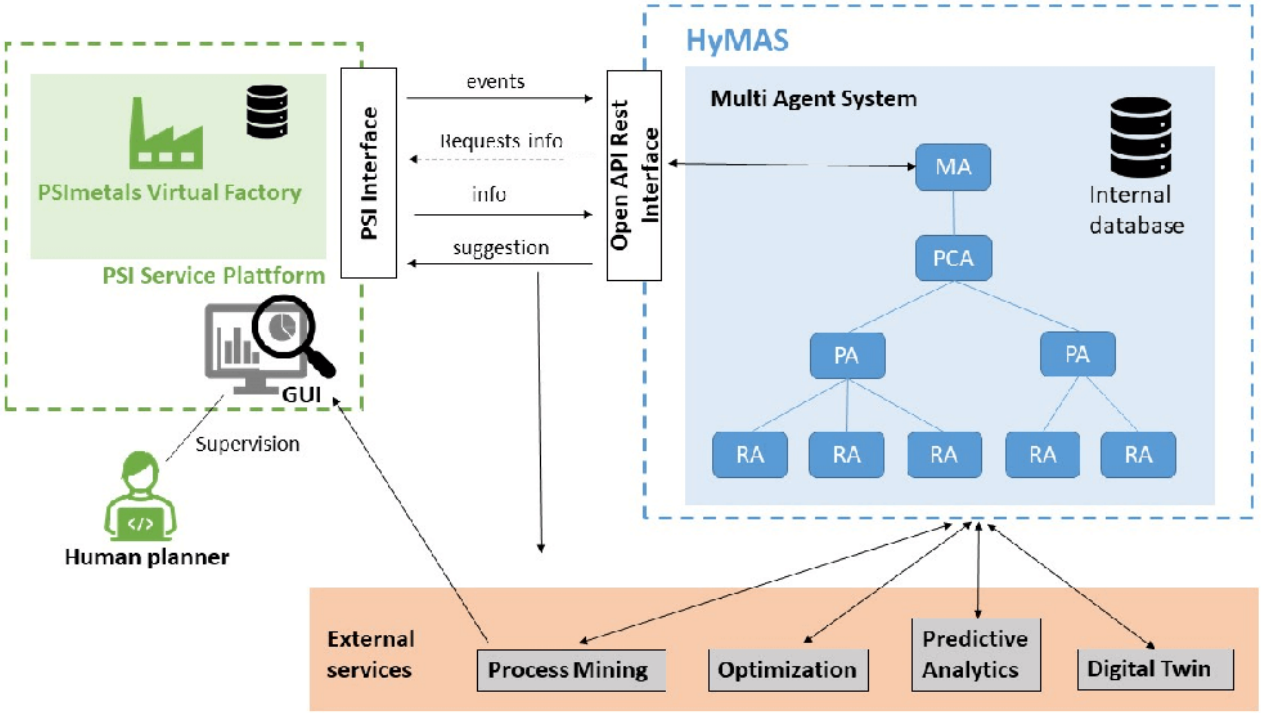

Fig. 1 - HyMAS Architecture

At its core, HyMAS is composed of the following agent types:

- **Resource Agents (RA)**: Represent individual machines or production lines, such as a specific annealing line or the rolling mill. Each RA is responsible for local scheduling and execution, proposing job sequences for its resource and monitoring progress. RAs can negotiate task handovers or sequencing changes through their supervising agents when disruptions occur.
- **Process Agents (PA)**: Manage entire production stages composed of multiple resources. For example, one PA may coordinate all annealing lines, while another manages the rolling stage. PAs aggregate constraints and schedules from their subordinate RAs, ensuring that all resources within the stage operate in synchrony and that workload is balanced.
- **Process Chain Agent (PCA)**: Coordinates the full production sequence across all stages. The PCA maintains a global view of throughput, inventory, and delivery performance, ensuring capacity alignment between upstream and downstream processes and adjusting plans to resolve bottlenecks.
- **Master Agent (MA) – optional**: Functions as a supervisory or integration layer, particularly for

external communications. In the present implementation, the PCA fulfills this role, so a separate MA is not explicitly required.

The “hybrid” nature of HyMAS refers to its combination of autonomous agent-based control with embedded optimization models. Each agent is equipped with decision-making modules that can range from lightweight heuristics to advanced mathematical programming solvers. This design enables a two-tier decision process. First, when a disruption occurs, agents respond within seconds using simple heuristics to maintain continuous operation — for instance, rerouting coils to other lines if one goes offline. Second, more computationally intensive optimization algorithms run in the background to refine the schedule over a longer horizon. When these algorithms produce an improved plan, the update is propagated across the system without interrupting ongoing execution.

Integration with the MES ensures that all schedules are based on up-to-date production data and can be directly executed on the shop floor. The digital twin enables the simulation of “what-if” scenarios, allowing major plan changes to be tested virtually before implementation. Predictive analytics modules, based on machine-learning models, forecast issues such as potential quality deviations or equipment failures, enabling proactive schedule adjustments. Process-mining tools continuously analyze operational data to identify inefficiencies and recurring bottlenecks; the resulting insights are used to adapt both heuristic rules and optimization models.

Through this architecture, HyMAS combines the flexibility and resilience of decentralized decision-making with the precision of advanced optimization, delivering rapid, well-coordinated responses to disruptions while continually improving plan quality.

SCHEDULING MODEL: PHALS OPTIMIZATION FOR ANNEALING LINES

To complement the agent-based architecture, an advanced scheduling optimization model was developed for the annealing stage of production. This model, referred to as PHALS (Parallel Heterogeneous Annealing Lines Scheduling) [2], captures the core decision problem of assigning and sequencing cold-rolled coils on multiple annealing lines. Its objectives are twofold: to maximize throughput by reducing both stringer usage and idle time, and to meet delivery due dates.

The problem considers a set of steel coils $\{1,2,\dots,N\}$ that have completed cold rolling and await annealing on M parallel lines with potentially different technical specifications. Each coil i has a due date D_i and may have one or more admissible processing modes, representing alternative annealing “recipes” (temperature–speed profiles) capable of meeting metallurgical requirements. Processing a coil in a given mode on a specific line determines both its processing time and its compatibility with other coils. When two consecutive coils are incompatible — for example, due to large width differences or conflicting temperature requirements — a dummy coil (stringer) must be inserted. Stringers are undesirable, as they consume time, energy, and material without contributing to production. The scheduling task therefore involves deciding for each coil: which line to assign it to, in which position to place it within that line’s sequence, and which processing mode to use. The goal is to minimize stringer insertions and tardiness while respecting processing-mode feasibility and line capacities.

The problem can be formulated as a mixed-integer linear program (MILP) [2]. However, solving it exactly becomes computationally infeasible for realistically sized instances due to the combinatorial growth of sequencing possibilities. To address this, a tailored two-phase heuristic approach was designed, combining the speed of graph-based algorithms with the precision of MILP applied to small subproblems:

- **Phase 1 – Constructive Sequencing via Shortest Path**: The sequencing problem for each line is formulated as a shortest-path search on a directed acyclic graph (DAG), where nodes represent the “last processed coil” state and arcs correspond to scheduling a specific next coil. Arc costs reflect penalties for stringers or lateness. A shortest-path algorithm is applied to generate a near-optimal sequence for each line in a simplified setting, minimizing local penalties while ensuring all coils are assigned. This step produces an initial feasible schedule quickly by greedily reducing local costs.
- **Phase 2 – Improvement via Fix-and-Optimize:** Starting from the initial solution, the schedule is refined through iterative fix-and-optimize decomposition. In each iteration, a subset of decision

variables — such as the sequence on one line or a group of coil assignments — is re-optimized using an exact MILP solver, while the rest of the schedule remains fixed. Examples include re-sequencing a single line's coils or swapping assignments between lines when this reduces stringers or delays. This systematic exploration of subproblems incrementally improves global schedule quality.

Experimental results with industry-inspired datasets show that this approach consistently outperforms a state-of-the-art commercial MILP solver in producing high-quality solutions within short computation times. In particular, PHALS achieves fewer stringers and lower tardiness under the same time limits, making it well-suited for the fast decision cycles required in HyMAS.

Within the HyMAS framework, PHALS is primarily applied at the Process Agent (PA) and Process Chain Agent (PCA) levels. The PA for the annealing stage receives coil orders, due dates, and current line status from the MES via the HyMAS interface and uses PHALS to generate an optimized annealing plan. This plan is communicated to the relevant Resource Agents (RAs) as a target sequence. RAs execute the plan locally, applying small heuristic adjustments if minor disturbances occur — for example, swapping two compatible coils when one is delayed in arrival. The PA can also re-run PHALS periodically in a rolling-horizon manner or immediately after a major disruption. The PCA ensures that the annealing schedule remains consistent with upstream and downstream operations, for instance by slowing the rolling stage when annealing becomes a bottleneck or by triggering re-optimization with updated priorities.

In summary, PHALS provides the analytical engine within HyMAS for optimizing the annealing stage. By explicitly modeling multi-mode, parallel-machine scheduling with domain-specific objectives, it enables the agent system to produce high-quality, disruption-resilient production schedules that are both computationally efficient and operationally realistic.

INTEGRATION AND APPICATION

The HyMAS framework was implemented as a prototype and evaluated through simulation in a realistic steel production environment. This required both software and process integration between the academic prototype and industrial systems.

The core HyMAS agents were developed using the JIAC multi-agent framework — a Java-based agent middleware from TU Berlin — and extended with custom HyMAS optimization modules. On the industrial side, the plant's Manufacturing Execution System (MES) provided the simulation environment and a graphical user interface for monitoring and interaction. Communication between HyMAS and the MES was handled via a RESTful API. This interface allowed the agents to subscribe to production events, request information, and transmit schedule updates or recommendations. The two-way communication ensured that the simulation state and agent decisions remained synchronized at all times.

For evaluation, a case study was configured using realistic data from a cold rolling and annealing operation. The simulated plant setup included one cold rolling mill producing coils for an intermediate buffer, a finite-capacity storage area where coils awaited annealing, and three parallel continuous annealing lines. The lines differed in process speeds and coil-size capabilities, reflecting the heterogeneous nature of the real facility. Approximately 200 coil orders of varying grades, dimensions, and due dates were scheduled over a multi-day planning horizon. Some orders required strict prioritization due to tight deadlines, while others had more flexible delivery windows.

The digital twin was loaded with this configuration, and realistic operational variability was introduced to test system adaptability. This included stochastic annealing line stoppages, random minor delays, one simulated rush order arriving mid-horizon, and occasional quality deviations requiring reallocation or reprocessing of coils.

During the simulation, HyMAS was responsible for monitoring events, adapting schedules, and proposing

updated plans. Human planners could observe these proposals in the GUI, review their associated key performance indicators (KPIs), and decide whether to accept them.

To assess performance, HyMAS was benchmarked against a conventional static planning approach. The baseline plan was generated in advance for all orders, without dynamic rescheduling. When disruptions occurred, the baseline approach simply paused and resumed execution as closely as possible to the predefined sequence, without re-optimization. In contrast, HyMAS continuously updated schedules in response to events, allowing for dynamic reallocation and sequencing adjustments.

Three KPIs were used to compare performance:

- **Operational efficiency**: productive time of annealing lines and the number of stringers used.
- **On-time delivery rate**: percentage of coils delivered by their due dates.
- **Rescheduling responsiveness**: time required to produce a meaningful updated schedule after a disruption.

This setup provided a controlled yet realistic environment for validating HyMAS's ability to deliver more efficient, timely, and disruption-resilient scheduling compared to conventional static methods.

RESULTS AND DISCUSSION

The simulation experiments demonstrated clear performance benefits of the integrated HyMAS approach compared to a conventional static plan. Table 1 summarizes the results. The dynamic, agent-driven scheduling of HyMAS yielded notable gains in both operational efficiency and delivery timeliness.

HyMAS improved annealing line efficiency by approximately 18 %, increasing average utilization from 75 % under the static plan to 88–90 %. This improvement was largely driven by the system's ability to intelligently group compatible coils, thereby minimizing stringer insertions and reducing idle periods. The average number of stringers required per 100 coils fell from 5.2 to 1.1 — an almost fivefold reduction. By inserting stringers only when absolutely necessary, HyMAS not only improved throughput but also lowered material waste and energy consumption.

Tab. 1 - Performance of static planning vs. HyMAS approach

Performance Results		
Performance Indicators	**Conventional Static Plan**	**HyMAS**
Annealing Lines Efficiency	75%	88-90%
On-Time Delivery Rate	85%	92%
Average number of stringers per 100 coils	5.2	1.1
Average Responce Time to Disruption	~30 min	< 1 min

Delivery performance also improved substantially. With HyMAS, 92 % of orders were completed on or before their due dates, compared to 85 % for the static plan. The gain was a direct result of dynamic rescheduling: when disruptions threatened to delay specific orders, HyMAS agents reprioritized affected coils and adjusted sequences to preserve on-time performance wherever possible. In contrast, the static plan, once disrupted, often led to cascading delays that impacted multiple orders downstream.

Importantly, HyMAS had to balance two potentially competing objectives: maximizing efficiency and minimizing tardiness. While pushing for absolute efficiency might delay some urgent orders, and eliminating all tardiness could require suboptimal line utilization, the multi-objective optimization within HyMAS effectively managed

this trade-off.

From an industrial perspective, these improvements are significant. An 18 % increase in line efficiency translates into substantial annual savings through higher throughput and lower energy usage, while an 8 % increase in on-time delivery directly enhances customer satisfaction and supply chain reliability. The ability to generate a viable reschedule in less than one minute — compared to around 30 minutes for the static approach — reduces work-in-process accumulation and stabilizes operations, giving planners confidence that routine disturbances can be resolved autonomously.

Overall, the results demonstrate that HyMAS delivers resilience, responsiveness, and efficiency improvements that align with the Industry 4.0 vision of self-regulating, smart manufacturing systems. For complex production environments such as steel plants, this approach enables a transition from static, manually managed schedules to adaptive, autonomous planning.

CONCLUSION

This work presented a Hybrid Multi-Agent System (HyMAS) integrated with an advanced scheduling optimization model to address the complexity of steel production planning. The framework enables resilient and autonomous scheduling that adapts dynamically to disruptions while optimizing key operational objectives. HyMAS combines a distributed, hierarchical decision structure — with Resource, Process, and Process Chain Agents — and embedded optimization modules, achieving a balance between the agility of local decisions and the rigor of global optimization.

In a realistic case study involving a cold rolling mill and multiple continuous annealing lines, HyMAS achieved substantial performance gains: up to an 18 % increase in throughput and an 8 % improvement in on-time delivery compared with a conventional static planning approach. These gains were accompanied by significant reductions in stringer usage and scheduling response times, underscoring the system's ability to deliver both efficiency and reliability.

From an academic standpoint, the results illustrate how coupling multi-agent system architectures with operations research techniques can yield robust, scalable solutions for complex industrial scheduling problems. The hierarchical MAS structure provided modular, stage-level control, while the PHALS optimization model served as a computational engine capable of handling the combinatorial complexity of multi-mode, parallel-machine scheduling. The hybrid design leveraged the strengths of both approaches: the responsiveness and local autonomy of agents, and the optimality-seeking capabilities of mathematical programming. This synergy represents a promising pathway for future smart factory implementations.

For industrial practitioners, the findings demonstrate the potential for tangible gains in operational performance without requiring major infrastructure changes. By dynamically re-optimizing production plans in response to events, HyMAS reduces downtime, improves delivery performance, and minimizes waste from inefficient sequencing. Its deployment can be incremental: initially functioning as a decision-support tool that provides real-time, optimized scheduling proposals to human planners, and later evolving toward full autonomy as confidence in the system grows. In such a setup, planners shift their focus from routine rescheduling to managing exceptions and aligning operations with strategic objectives, thereby reducing scheduling workload and increasing overall system resilience.

REFERENCES

[1] Iannino V., Vannocci M., Vannucci M., Colla V. A Multi-Agent Approach for the Self-Optimization of Steel Production. Int J of Simulation: Systems, Science & Technology. 2018;19(5):20 (pp. 1-8)

[2] Wegel S., Ivanov A., Lenz R., Volling T. Scheduling of parallel continuous annealing lines with alternative processing modes to optimize efficiency under tardiness constraints. European Journal of Operational Research. 2024;316(1):282-294