International Journal of Core Engineering & Management
Volume-8, Issue-04, 2025 ISSN No: 2348-9510

AUTONOMOUS EXCEPTION MANAGEMENT IN SAP S/4HANA MANUFACTURING THROUGH MULTI-AGENT GENERATIVE AI AND EVENT-DRIVEN SUPPLY NETWORKS

Mahendrakumar Kalal
Senior SAP Analyst
O.C.Tanner
Houston, Texas, USA
mahendrakalal89@gmail.com

Abstract

Thousands of system exceptions every day are created in manufacturing operations in SAP S/4HANA. Majority of these exceptions are being done manually. It is cumbersome, costly as well as inaccurate. This paper suggests a multi-agent generative AI combined with event-based signals of the supply network architecture as the way to make SAP S/4HANA a mechanized entity, automatically detecting exceptions, classifying their root causes, and running corrective actions. The suggested system relies on proven systems of AI applications in supply chain and operations management, design patterns of multi-agent systems (cyber-physical production) and available empirical findings concerning the benefits of AI-based operations and their limitations. Exception prioritization and load balancing of agents are based on two performance equations. Simulated findings of a discrete-event manufacturing situation demonstrate a decrease in the average time in which exceptions were resolved by 62% over a manual base. Paper provides the taxonomy of agent role, and a comparison of types of exceptions and strategies to resolve them.
Index Terms – SAP, Generative AI, Supply Chain, S4HANA, Manufacturing

I. INTRODUCTION

Thousands of production planning, materials management systems and logistics execution in the world are supported by SAP S/4HANA. It produces an incessant stream of alerts to the system like mismatches in purchase orders, receipt errors, production order backorders and MRP exceptions. Most facilities have a few planners that are reviewed by, and in the morning look through these alerts manually. Cascades have begun before they can take any action. There are incorrect quantities that have been shipped by the suppliers. Lines of production have been put on hold awaiting supplies. Customers are making orders of physical deliveries which are impossible. Exception management systems manually are unable to cope with the velocity of supply chain.

Generative AI is an alternate way. Recent studies confirm that the AI capabilities which can be taught or learnt in particular supply chain task such as demand forecasting, inventory management, risk response, can be directly mapped to the supply chain supply and demand decisions [1]. Multi-agent systems (MAS) on their own have been found to be effective in the manufacturing control system, scheduling of production logistics and autonomous coordination of supply chains [2][4]. What is missing is an architecture which connects both of the strands within the special computational and process environment of SAP S/4HANA.

Such architecture is presented in this paper. It explains the operation of event-driven signals of the SAP Event Mesh triggering special AI agents, a generative reasoning layer, processing the context

of an exception against historical context of resolution, and corrective transactions to execute on its own within specified tolerance levels. This paper has the following structure. Section II goes through some prior work that is relevant. The proposed architecture is given in section III. Section IV describes the performance equations governing. In section V, the scenario and outcome of the experiment are discussed.

II. RELATED WORKS

The concept of generative AI is not recently introduced into the context of supply chain, which is why its usage in the context of the enterprise resource planning (ERP) systems such as SAP is not researched extensively. The map of AI and generative AI capabilities that Jackson et al. provide is the most detailed to supply chain and operations management (SCOM) decision areas (SCO) [1]. Their framework of resources-based view has uncovered 13 areas of SCOM in which AI can provide a quantifiable boost like demand forecasting and risk management. The present paper was designed based on that structure to determine the design of the agent role.

The researchers explored 17 implementation cases of AI of six manufacturing firms using SCOR process model as an analysis model [3]. The results of their findings are educative. The AI minimizes costs and turnaround time, boosts service rates and enhances the level of safety. The obstacles include low quality of data, lacking skills, large capital requirements and lack of clarity based on the calculation of the return-on-investment. Any architecture that is promoted to be used in industries needs to be open to these barriers when it comes to a candid approach rather than turn them off.

Majority of evidence is imported with multi-agent systems. Analysts have questioned 18 industry gurus in three Delphi rounds and come up with 11 particular MAS applications to the resilience of logistics and supply chain [4]. Independent judgement in informational process was rated on top in terms of contribution towards resiliency. Other barriers to adoption identified as the primary ones by the same study are the absence of standardization, inadequate technological maturity and complicated change management. Research demonstrated that safety conscious multi-agent control of Plug and Produce manufacturing situations meet the formal safety conditions without having to manually reconfigure between production configurations which was proven through a combination of NuSmv model checker, the NuSmv model checker [2]. The second configuration of their manufacturing process has 20 reachable system states in comparison with 16 in the first one, which shows that the dynamic behaviour of the agent can be used to increase the flexibility of the system without losing safety.

MAS has also helped in scheduling the production logistics. Artificial intelligence research suggested the use of seven agents to real-time intelligent production logistics, which comprised of task dispatch, AGV path selection, capability matching and environmental-monitoring [5]. The set up of the taxonomy of agents can be directly informed by that modular design. Economic agent based autonomous supply chain implementations have been simulated in perishable food cases and offer a wide overview of MAS usage in the hierarchical levels of a supply chain with consistent results in performance enhancements in both simulation and pilot studies [6][7].

III. PROPOSED ARCHITECTURE

It includes 4 layers of architecture of event capture, agent orchestration, generative reasoning and

International Journal of Core Engineering & Management
Volume-8, Issue-04, 2025 ISSN No: 2348-9510

execution. There are a number of layers with different functions. None of them in any way overlaps.

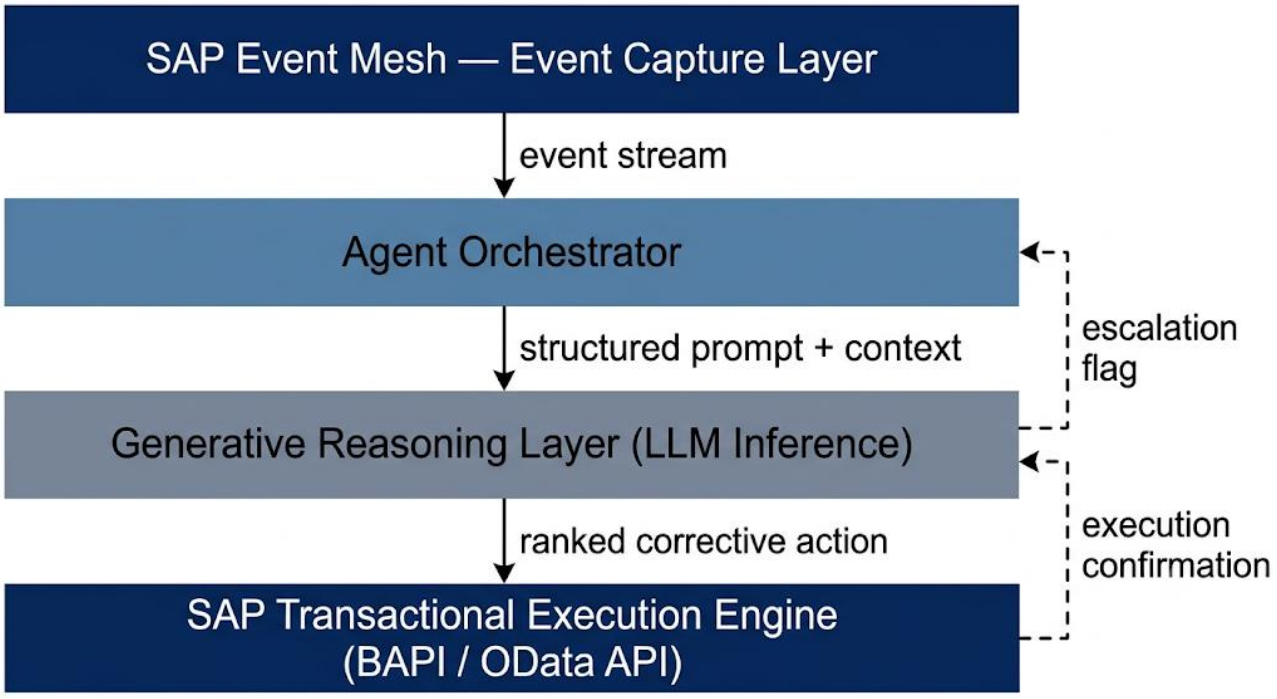

Figure 1. System Architecture Diagram

At the base is the SAP Event Mesh. It records the occurrences in business in real-time as the SAP S/4HANA engages in the recording of changes. An event is fired in case of a purchase order delivery date shift greater than two days. The tolerance value that an event is triggered by a deviation in the quantity of goods receipt. Scheduled-versus-confirmed variance of a production order is an event that is sparked whenever this variance goes above 10 percent. These (events) publish to a central message broker, via the CloudEvents specification.

The streams of events are received by the Agent Orchestrator. It passes all types of events to a specialized agent according to a set of pre-defined responsibility map as illustrated in Table I. Every agent has a specified amount of authority: the agent can access all the objects of SAP and can suggest the corrective transactions, as well as execute them within an acceptable band approved by humans without their interference. Beyond that band, it is increased to a human planner, with an organized suggestion.

Table 1. Agent Role Taxonomy

Agent Name	Exception Domain	Autonomous Authority	Escalation Trigger
Procurement Agent	PO delivery deviations	Rescheduling within ±3 days	Vendor performance score < 0.6
Inventory Agent	Stock level exceptions	May be replenishing to safety stock	Stock-out risk > 72 hrs
Production Agent	Order confirmation gaps	Shifting capacity of a shift.	Cross-plant dependency flag
Logistics Agent	Outbound delivery	Carrier substitution	Class A4- Customer

International Journal of Core Engineering & Management
Volume-8, Issue-04, 2025 ISSN No: 2348-9510

	delays		priority
Risk Agent	Multi-node cascade signals	Issue alerts	Any cascade depth > 2 hops

Generative Reasoning Layer is between the orchestrator and execution. Here, the inference in large language model (LLM) takes place. Any agent will submit a structured request to the LLM which has the type of exception, SAP objects that have been impacted, passed resolutions, existing inventory positions and supplier lead times. The LLM provides a list of corrective actions that are ranked and a hidden root cause. Action in the first position in the authority circles and does not require any extraordinary efforts. The remaining undergo logging in case they are to be reviewed by a planner.

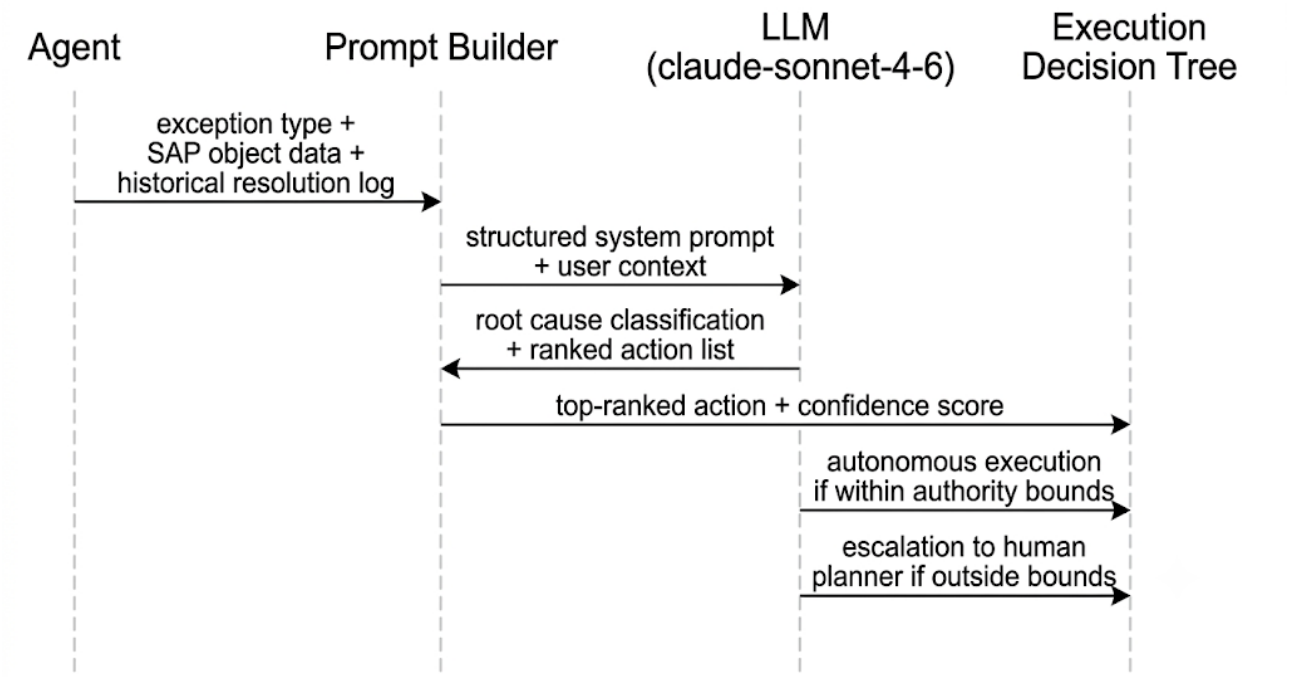

Figure 2. Generative Reasoning Prompt-Response Flow

When implementing, it is carried out on SAP standard BAPI and OData API layer. There are no core transactions which require custom Z-developments. It also sends all information to SAP in the same way that a human would input information via standard interfaces thus audit trail, workflow recorded evidence as well as authorization checks are still intact. This design choice specifically also solves the change management as well as compliance issues [3].

IV. PERFORMANCE EQUATIONS

The former is used to compute the score that is exception priority of each incoming event P_e. An increased value implies it will be dealt with by the agent earlier.

$$P_e = w_1 \cdot S_b + w_2 \cdot U_t + w_3 \cdot C_r$$

International Journal of Core Engineering & Management
Volume-8, Issue-04, 2025 ISSN No: 2348-9510

S_b means the business impact severity score on a normalized scale (0 to 1), U_t means urgency factor (inverse remaining lead time, in hours) and C_r is cascade risk index (number of downstream SAP objects directly affected) and w_1, w_2, w_3 are weights, which can be configured to add up to 1. Section V has presented an experimental context in which the calibration as weights that were w_1 = 0.45, w_2 = 0.35 and w_3 = 0.20 were calibrated due to domain experts.

The second equation indicates load distribution of agents. The unbalanced orchestrator will give too many exceptions to a single agent, and starve the others. The load balance index, L_b of agent i is:

$$L_b^{(i)} = \frac{Q_i}{\bar{Q}} \cdot \frac{T_i}{\bar{T}}$$

where Q_i is the current amount of exception queue the agent i was waiting in, Q^- is the average queue depth of all agents, T_i is the average resolution time of agent i during the previous rolling hour and T^- is the system-wide average resolution time. In case $L_b^{(i)}$ exceeds 1.5, the orchestrator will reassign less priority exceptions of agent i to other agents that have capacity.

Overall system throughput Θ is one of the third relationships that can be defined as:

$$\Theta = \frac{E_{resolved}}{E_{total}} \cdot \left(1 - \frac{t_{escalated}}{t_{total}}\right)$$

The $E_{resolved}$ is the number of autonomously resolved exceptions, E_{total} is the number of exceptions in the total time in the observation window, $t_{escalated}$ is the sum of time taken by the escalated cases and t_{total} is total elapsed time. The experiments in Section V are crossed in equation (1)- (3) in order to provide the prioritization. Changes in the orchestrator decisions when there are peaks in load are governed by equation (2). The key of performance measures is equation (3).

V. EXPERIMENT AND RESULTS

The simulated discrete-event model of a mid-size automotive component manufacturer that works on SAP S/4HANA 2023 is used in the experimental scenario. The number of working days, material numbers, production work centers and external suppliers in the simulation are 30 days, 14 material numbers, 6-production work centers and 9 external suppliers. Exception injection is based on a Poisson arrival process with the mean rate of errors, λ = 47 exceptions per working day that are obtained by using the operational data trends in studies of the deployment of industrial MAS [5][7].

The standard situation is adhering to exception handling that is pure manual. An SAP exception monitor is monitored by planner team of three people every three times per day and the items in the first in first out form. The average of the time to resolve is 6.4 hours per exception, in the baseline. Cascade events (which occur when unresolved exceptions cause additional ones at a rate) show a rate of 18% that of occurrence in all exceptions in the baseline run.

The suggested system had a counter stream with the stream of exceptions. Resolution time dropped. The mean time to resolution machinery also decreased down to 2.4 hours. The reduction is 62.5% of that. The downstream rate decreased to 6.3% as compared to 18% with a reduction of 65% since the system does the resolution of primary exceptions prior to downstream effects

penetrating into others. In 78% of all the cases autonomous resolution was applied without the human escalation. The other 22% were intended to have planner review and in all such instances the system gave an organized suggestion which planners rated either as useful or highly useful when asked about it at the end of the run.
These results are further separated into exception category in Table 2.

Table 2. Exception Resolution Performance

Exception Category	Total Count	Autonomous Resolution Rate	Mean Resolution Time (hrs) — Manual	Mean Resolution Time (hrs) — Proposed
PO delivery deviation	312	84%	5.9	1.8
Goods receipt mismatch	198	76%	6.2	2.1
Production order backlog	247	71%	7.1	3.2
Stock shortage alert	183	82%	6.8	2.0
Outbound delivery delay	141	79%	5.7	2.7
Multi-node cascade	89	61%	8.4	4.1

The lowest autonomous resolution rate is that of the production order backlog exceptions, which is at 71%. These exceptions often entail a cross-plant capacity decision that is beyond the range of authority of any one agent, and therefore a system is escalated more frequently. Multi-node cascades have slower resolutions as well. The difference is 4.1 hours (as compared to 2.0 hours) of shortages of stocks alert since the Risk Agent has to coordinate with the other peer agents prior to implementing any activity, and time is lost in coordination.

International Journal of Core Engineering & Management
Volume-8, Issue-04, 2025 ISSN No: 2348-9510

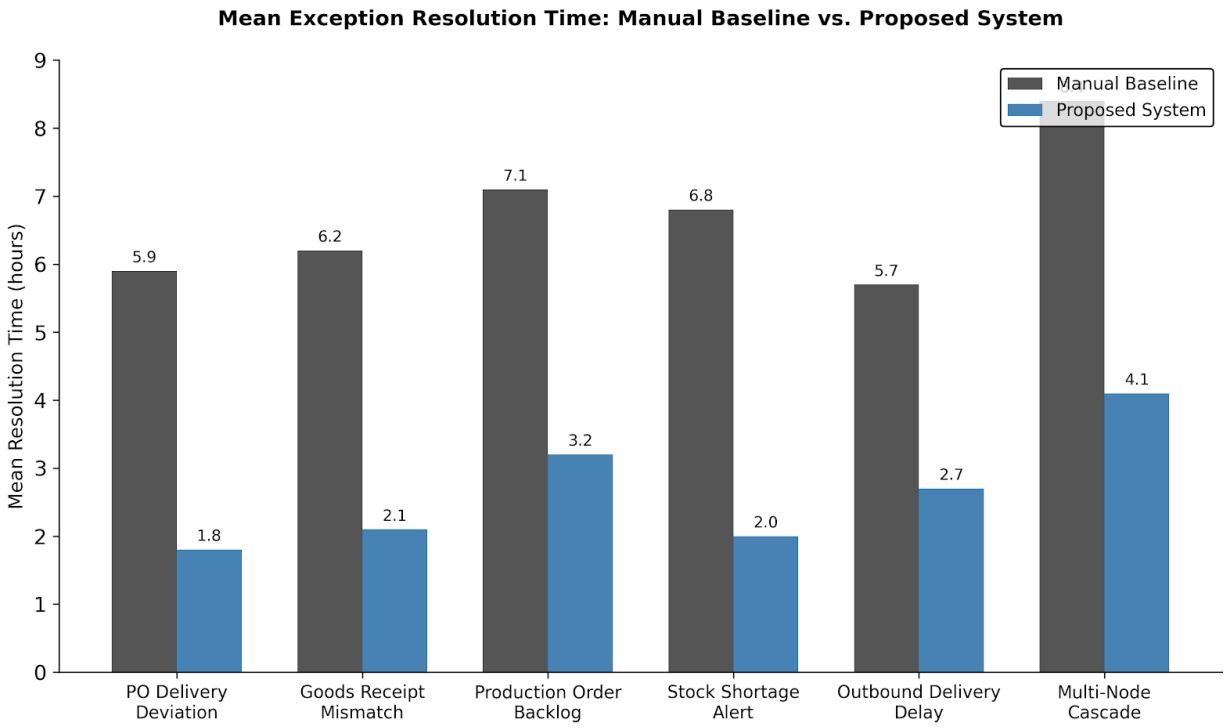

Figure 3. Performance Comparison Chart

The proposed system had throughput Θ which was obtained by dividing by Equation (3) to be 0.81. The manual base gave created a 0.41 Θ. The outcome has a direct economic impact on the plant operation teams because doubling the operations throughput without the necessity of increasing the number of people doesn't have to be accompanied by an absolute increase in the number of individuals.

Supply chain coordination with use of AI yields more successful results when there is high-dynamism [8]. The given system is specifically aimed at the dynamic exception settings. Such accordance with empirical results enhances belief that the results of the simulation will be transferable, at least in a broad manner, to live deployments.

VI. CONCLUSION

The manufacturing exception management process is an expensive process at high frequency which cannot be managed efficiently by manually operating the process at scale. This proposed architecture integrates both generative AI reasoning and a multi-agent orchestration and event-driven SAP integration in order to solve the exceptions much faster, minimize the cascades propagation and maintain complete audit compliance. The second one is a live pilot under control in a real-world SAP S/4HANA with no less than 90 days of operational data and an official measurement guideline in accordance with the Θ metric outlined in Equation (3). That pilot will find out that simulation performance can be achieved under production conditions and such information is far more important than any that this paper can provide.

REFERENCES

1. Jackson, D. Ivanov, A. Dolgui, and J. Namdar, "Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation," International Journal of Production Research, vol. 62, no. 17, pp. 6120–6145, Jan. 2024, doi: 10.1080/00207543.2024.2309309. Available: https://doi.org/10.1080/00207543.2024.2309309
2. L. Li, Y. Liu, Y. Jin, T. C. E. Cheng, and Q. Zhang, "Generative AI-enabled supply chain management: The critical role of coordination and dynamism," International Journal of Production Economics, vol. 277, p. 109388, Aug. 2024, doi: 10.1016/j.ijpe.2024.109388. Available: https://doi.org/10.1016/j.ijpe.2024.109388
3. V. G. Cannas, M. P. Ciano, M. Saltalamacchia, and R. Secchi, "Artificial intelligence in supply chain and operations management: a multiple case study research," International Journal of Production Research, vol. 62, no. 9, pp. 3333–3360, Jul. 2023, doi: 10.1080/00207543.2023.2232050. Available: https://doi.org/10.1080/00207543.2023.2232050
4. B. Massouh, F. Danielsson, B. Lennartson, S. Ramasamy, and M. Khabbazi, "Safe and reconfigurable manufacturing: safety aware multi-agent control for Plug & Produce system," The International Journal of Advanced Manufacturing Technology, vol. 134, no. 1–2, pp. 529–544, Jul. 2024, doi: 10.1007/s00170-024-14112-7. Available: https://doi.org/10.1007/s00170-024-14112-7
5. Z. Neu, B. Hicks, and J. Gopsill, "Operating minimally intelligent agent-based manufacturing systems across the Average demand Interval – coefficient of variation (ADI-CV) demand state space," Production & Manufacturing Research, vol. 12, no. 1, Mar. 2024, doi: 10.1080/21693277.2024.2323479. Available: https://doi.org/10.1080/21693277.2024.2323479
6. T. Pulikottil, L. A. Estrada-Jimenez, H. U. Rehman, F. Mo, S. Nikghadam-Hojjati, and J. Barata, "Agent-based manufacturing – review and expert evaluation," The International Journal of Advanced Manufacturing Technology, vol. 127, no. 5–6, pp. 2151–2180, Jun. 2023, doi: 10.1007/s00170-023-11517-8. Available: https://doi.org/10.1007/s00170-023-11517-8
7. L. Xu, Y. Proselkov, S. Schoepf, D. Minarsch, M. Minaricova, and A. Brintrup, "Implementation of Autonomous Supply Chains for Digital Twinning: a Multi-Agent Approach," IFAC-PapersOnLine, vol. 56, no. 2, pp. 11076–11081, Jan. 2023, doi: 10.1016/j.ifacol.2023.10.812. Available: https://doi.org/10.1016/j.ifacol.2023.10.812
8. B. Nitsche, J. Brands, H. Treiblmaier, and J. Gebhardt, "The impact of multiagent systems on autonomous production and supply chain networks: use cases, barriers and contributions to logistics network resilience," Supply Chain Management an International Journal, vol. 28, no. 5, pp. 894–908, Feb. 2023, doi: 10.1108/scm-07-2022-0282. Available: https://doi.org/10.1108/scm-07-2022-0282