Multiverse Journal
E-ISSN: 3080-3942, P-ISSN: 3080-3934
Volume 2, Issue 2, December 2025 | PP. 81-99

Research Paper

Containerized AI-Driven Well Completion, Workover, and Production Optimization Systems Using Kubernetes–OpenStack Infrastructure.

Nnaemeka Princewill Ohia[1]

[1] Federal University of Technology, Owerri (FUTO), Email: nnaemeka.ohia@futo.edu.ng,

Received: 23 August, 2025 **Accepted:** 01 November 2025 **Published:** 30 December, 2025

Abstract

The oil and gas industry faces pressure to optimize well delivery, reduce non-productive time, and improve production across dispersed fields. Traditional workflows rely on sequential, compute-heavy simulations for well design and optimization, causing decision-making bottlenecks. This paper introduces a containerized AI system built on Kubernetes–OpenStack that enables scalable, GPU-accelerated well engineering. It integrates AI models for well completion, workover ranking, and forecasting in a multi-tenant setup. Based on performance benchmarks, the system shows how infrastructure improvements boost operational efficiency. It supports parallel torque, hydraulic, stress, and forecast simulations for large well portfolios. Case studies highlight faster planning, better intervention success, and increased production. The approach allows quick deployment of AI, automated workflows, and real-time digital twins for extended drilling and field optimization. This bridges cloud infrastructure and petroleum engineering by enabling scalable AI workflows for well management in multi-tenant environments.

Keywords: Containerization, Kubernetes, Artificial Intelligence, Well Completion, Workover Optimization, Production Optimization,

1. Introduction

The petroleum engineering domain has witnessed substantial transformation through the integration of artificial intelligence and cloud computing technologies over the past decade. Well completion design, workover planning, and production optimization represent computationally intensive processes that traditionally required extensive manual analysis and sequential simulation workflows (Shelley et al., 2021). Modern oil and gas operations, particularly in complex environments such as extended reach drilling (ERD) campaigns and mature field management, demand rapid scenario evaluation, multi-objective optimization, and real-time decision support capabilities that exceed the capacity of conventional engineering workstations (Mata et al., 2021). The convergence of machine learning algorithms, containerization technologies, and cloud orchestration platforms has created unprecedented opportunities for transforming well engineering workflows. Patchamatla (2018) demonstrated that Kubernetes-based multi-tenant container environments deployed on OpenStack infrastructure deliver significant performance improvements for scalable AI workflows, including reduced latency, enhanced resource utilization, and cost efficiency through dynamic pod scheduling and GPU sharing. These infrastructure-level capabilities establish the foundation for deploying AI-driven petroleum engineering applications that require intensive computational resources and rapid iteration cycles.

Well completion optimization involves complex decision-making processes that integrate geological data, reservoir properties, mechanical constraints, and economic objectives to determine optimal completion parameters (Sheikhi Garjan & Ghaneezabadi, 2020). Traditional approaches rely on physics-based simulators that, while accurate, consume substantial computational time when evaluating multiple completion scenarios across large well portfolios. Machine learning models trained on historical completion and production data offer complementary capabilities by enabling rapid prediction of completion performance and identification of key design parameters (Baki et al., 2021). However, deploying these AI models at scale across multi-well assets requires robust orchestration infrastructure capable of managing containerized services, balancing computational loads, and providing consistent performance. Workover planning and well intervention candidate selection present similar challenges, requiring integration of production history, well integrity monitoring data, reservoir simulation outputs, and economic forecasts to prioritize intervention opportunities (Mat Khair et al., 2023). Automated workflows that combine physics-based models with machine learning algorithms have demonstrated success in screening hundreds of wells and ranking candidates based on production uplift potential and cost-benefit analysis (Mata et al., 2021). The scalability of these workflows depends critically on underlying infrastructure capable of supporting parallel execution, data pipeline orchestration, and real-time model inference. Production optimization encompasses a broad spectrum of activities including gas-lift optimization, artificial lift management, virtual metering, and integrated asset optimization (Al Selaiti et al., 2020). Recent advances in reinforcement learning have enabled development of

adaptive control policies that optimize well production settings in real-time while accounting for operational constraints and reservoir dynamics (Poort et al., 2022). Deploying these sophisticated AI models in production environments requires containerized architectures that support continuous model training, A/B testing, and seamless integration with supervisory control and data acquisition (SCADA) systems.

This paper addresses the critical gap between infrastructure optimization and application-level petroleum engineering decision support by presenting a comprehensive framework for containerized AI-driven well completion, workover, and production optimization. The research builds directly upon the Kubernetes–OpenStack infrastructure benchmarks established by Patchamatla (2018), translating container orchestration capabilities into measurable operational improvements in well engineering workflows. The proposed architecture leverages multi-tenant resource management, GPU acceleration, and automated scaling to support compute-intensive simulations including torque and drag modeling, hydraulic analysis, tubing stress calculations, and production forecasting. The remainder of this paper is organized as follows: Section 2 reviews relevant literature on AI applications in petroleum engineering and containerized computing platforms. Section 3 presents the system architecture and technical implementation details. Section 4 describes AI model development for completion, workover, and production optimization. Section 5 discusses deployment strategies and operational workflows. Section 6 presents performance evaluation and case study results. Section 7 concludes with implications for future research and industry adoption.

2. Literature Review

2.1 AI Applications in Well Completion Design

Machine learning applications in well completion optimization have gained significant traction in both conventional and unconventional reservoirs. Sheikhi Garjan and Ghaneezabadi (2020) applied multiple ML algorithms including random forest, gradient boosting, XGBoost, and neural networks to 1,838 horizontal fracturing wells in the Montney formation, demonstrating that interpretable ML models could quantify the impact of individual completion parameters on first-12-month production and economic outcomes. Their work emphasized the importance of model interpretability tools such as individual conditional expectation (ICE) plots and partial dependence plots (PDP) for gaining engineering insights from black-box models. Baki et al. (2021) presented a comprehensive ML workflow for unconventional reservoir completion optimization, utilizing artificial neural networks (ANN), support vector machines (SVM), and ensemble tree methods to map completion and stimulation parameters to production outcomes. The research demonstrated that properly engineered features and hyperparameter tuning could achieve prediction accuracies exceeding 85% for cumulative production forecasts, enabling rapid evaluation of completion design alternatives without expensive reservoir simulation. Shelley et al. (2021) employed self-organizing maps (SOM) and feed-forward artificial

neural networks to analyze 301 Wolfcamp B completions, relating reservoir characteristics and completion types to production and economic performance. The study highlighted that data-driven completion design evaluation could be performed at low cost and high speed, making it particularly suitable for deployment as containerized inference services that support interactive engineering workflows.

Liao et al. (2020) developed stacking ensemble models combining random forest, XGBoost, and LightGBM for completion optimization in Canadian tight gas fields, specifically targeting Montney and Wapiti formations. The research demonstrated that ensemble approaches improved prediction robustness compared to individual models and provided detailed sensitivity analyses identifying critical completion parameters. This work established patterns for integrating multiple ML models within unified optimization frameworks, a capability well-suited to containerized microservice architectures.

2.2 Workover and Production Enhancement Optimization

Automated identification and ranking of workover candidates represents a critical application area for AI in mature field management. Mata et al. (2021) developed automated reservoir management workflows combining physics-based models with machine learning techniques including Bayesian networks and time-series forecasting to screen over 700 wells in a giant offshore Abu Dhabi field. The system reduced candidate review time by 60% while improving identification accuracy for production enhancement opportunities. The workflow architecture demonstrated clear potential for implementation as containerized services with scheduled execution and dashboard integration. Mat Khair et al. (2023) described an ML-driven Production Enhancement Candidate Generation and Screening (PECGS) system deployed on mature offshore fields, integrating production data, well integrity monitoring, and reservoir characteristics to identify underperforming wells and recommend specific interventions. The system's modular design and REST API architecture aligned naturally with containerized deployment patterns and Kubernetes orchestration. Dallag et al. (2022) presented a digital solution for continuous corrosion monitoring based on machine learning algorithms that combined cased-hole logs, open-hole data, production histories, and reservoir properties to generate animated corrosion maps and rank wells for workover priority. The ML-based integrity monitoring workflow provided essential inputs to production optimization systems and demonstrated the value of integrating diverse data sources through automated pipelines.

2.3 Production Optimization and Real-Time Control

Reinforcement learning has emerged as a powerful approach for real-time production optimization. Poort et al. (2022) trained actor-critic RL agents on data-driven proxy models to optimize valve settings and gas-lift rates, achieving up to 17% production improvement while minimizing slugging behavior. The research

demonstrated policy transfer capabilities that accelerated convergence when deploying trained models to new wells, a critical requirement for scalable field-wide optimization. The computational architecture required for RL training and inference maps directly to GPU-accelerated container deployments. Al Selaiti et al. (2020) built ML models for virtual metering, short-term production forecasting, and sensitivity analysis to derive optimal gas-lift and choke settings for natural flowing and gas-lift wells in Abu Dhabi. The study quantified potential net profit improvements of approximately 2.5% through optimal allocation, demonstrating measurable economic value from ML-driven production optimization. The operational ML tasks including virtual metering and forecasting represent ideal candidates for real-time inference containers integrated with digital twin platforms. Tariq et al. (2020) developed particle swarm optimization (PSO)-trained artificial neural networks for near-real-time flowing bottom-hole pressure (FBHP) prediction in vertical wells with multiphase flow, reporting mean absolute errors below 2.1% on field data. This surrogate modeling approach significantly improved computational speed compared to conventional empirical correlations, enabling integration into closed-loop production optimization systems. The low-latency inference requirements align with containerized microservice deployment patterns.

Lu et al. (2022) employed hyperparameter-tuned deep neural networks for shale oil production forecasting and fracturing optimization, validating models against field data and demonstrating production forecast accuracy improvements. The research established best practices for hyperparameter tuning and model validation that translate directly to continuous integration/continuous deployment (CI/CD) pipelines in containerized environments.

2.4 Digital Twins and Real-Time Optimization Frameworks

Digital twin technologies provide comprehensive frameworks for integrating AI models with real-time data streams and physics-based simulators. Shankar et al. (2022) proposed an open-source, microservice-based knowledge digital twin prototype for upstream oil and gas operations utilizing IoT stacks and ontologies. The architecture's modular design and microservice approach mapped directly to containerized deployments on Kubernetes–OpenStack platforms, demonstrating practical implementation patterns for digital twin-enabled well completion and production services. Vorobev et al. (2022) demonstrated a field-scale digital twin history-matched to instrument readings for selecting production-enhancing measures, scheduling maintenance actions, and performing mid-term forecasting. The case study illustrated integration of containerized ML models and simulators within live optimization loops, providing templates for enterprise-scale digital twin deployment. Singh et al. (2023) conducted a critical review of real-time optimization and decarbonization of oil and gas production value chains enabled by Industry 4.0 technologies, emphasizing digital twins, advanced modeling and simulation, and modular microservice deployment. The review synthesized recommendations for

implementing cloud-based, container-orchestrated production optimization systems that support both operational efficiency and emissions reduction objectives.

2.5 Containerized Simulation and High-Performance Computing

The migration of engineering simulation workflows to containerized environments has accelerated in recent years. Višňovský et al. (2022) discussed containerized high-performance simulation pipelines and integration of heterogeneous tools into reproducible workflows within containerized HPC environments. The research addressed practical challenges in packaging legacy simulation codes as containers and orchestrating distributed execution, providing direct guidance for deploying petroleum engineering simulators including torque and drag solvers and hydraulic models. Lyu (2020) evaluated containerizing simulation executables and compared orchestration approaches using Docker Swarm versus Kubernetes for simulation workloads. The comparative analysis identified Kubernetes advantages in terms of scaling capabilities, fault tolerance, and ecosystem maturity, supporting its selection as the preferred orchestration platform for petroleum engineering applications. The literature demonstrates substantial progress in applying AI to well completion, workover, and production optimization, alongside growing maturity of containerized computing platforms. However, a significant gap remains in comprehensive frameworks that operationalize these AI capabilities on scalable, multi-tenant infrastructure specifically optimized for petroleum engineering workflows (Joseph, 2013).

From a systems-governance perspective, the challenge extends beyond computational scalability to the architectural coherence of decision-support environments. Fragmented analytical pipelines, isolated simulation tools, and disconnected optimization workflows limit organizational visibility, delay intervention planning, and constrain real-time operational intelligence. Integrated governance theory demonstrates that unifying risk, control, and decision processes within a coordinated architectural framework enhances monitoring continuity, feedback responsiveness, and strategic oversight across complex technical systems (Joseph, 2013). Translating this principle into petroleum engineering requires container-orchestrated infrastructures that consolidate AI inference, physics-based simulation, and production optimization within closed-loop digital environments. Such architectural convergence establishes the theoretical foundation for scalable, real-time well engineering intelligence and positions containerized AI systems as governance-enabling platforms rather than merely computational accelerators. This research addresses that gap by building upon the Kubernetes–OpenStack infrastructure optimizations established by Patchamatla (2018) and demonstrating their application to real-world well engineering challenges.

3. System Architecture

3.1 Infrastructure Foundation

The containerized AI-driven well engineering system builds upon the Kubernetes–OpenStack infrastructure architecture validated by Patchamatla (2018), which demonstrated superior performance characteristics for scalable AI workflows. OpenStack provides the Infrastructure-as-a-Service (IaaS) layer, managing compute instances, storage volumes, network connectivity, and GPU resources across distributed data centers. The platform's multi-tenancy capabilities enable isolation of workloads across different asset teams, business units, or operational regions while maintaining resource efficiency through dynamic allocation. Kubernetes operates as the container orchestration layer, managing deployment, scaling, and operation of containerized applications across the OpenStack compute cluster. The Kubernetes scheduler leverages node affinity rules, resource quotas, and quality-of-service (QoS) classes to optimize pod placement, ensuring that GPU-intensive simulation workloads receive appropriate hardware acceleration while CPU-bound ML inference services achieve high throughput through horizontal scaling. The integration of Kubernetes with OpenStack's Neutron networking and Cinder block storage enables persistent data management for training datasets, simulation results, and model artifacts. Figure 1 illustrates the layered system architecture, depicting the progression from infrastructure resources through orchestration capabilities to containerized engineering services and end-user interfaces.

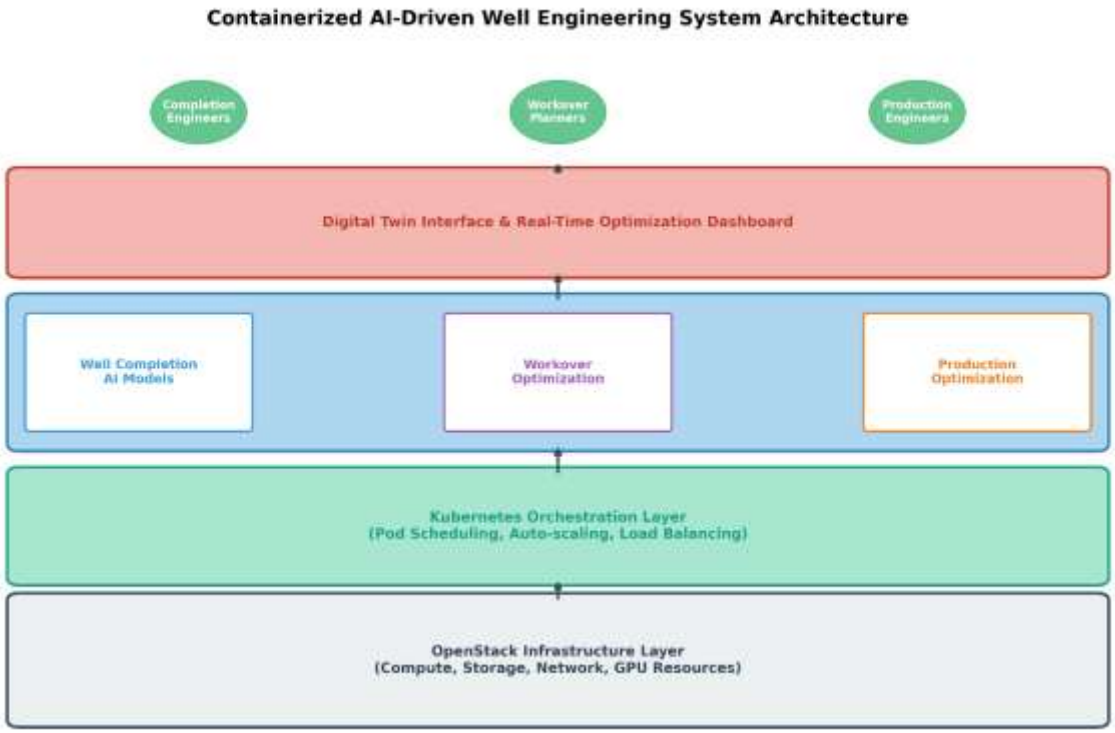

Figure 1: Containerized AI-Driven Well Engineering System Architecture

3.2 Containerized Service Architecture

The application layer implements a microservices architecture with three primary service domains: well completion AI models, workover optimization, and production optimization. Each domain encapsulates specific ML models, simulation engines, and data processing pipelines as independent containerized services that communicate through RESTful APIs and message queues. Well Completion Services include ML inference containers for completion parameter prediction, GPU-accelerated torque and drag simulation containers, hydraulic modeling services, and tubing stress analysis engines. These services consume geological data, well trajectory information, and completion design parameters to generate completion recommendations, mechanical risk assessments, and cost estimates. The containerized architecture enables parallel evaluation of multiple completion scenarios, with Kubernetes horizontal pod autoscaling dynamically provisioning additional compute resources during peak analysis periods.

Workover Optimization Services integrate production data ingestion pipelines, well integrity monitoring ML models, candidate ranking algorithms, and economic evaluation engines. The services implement automated workflows that continuously screen well portfolios, identify underperforming assets, diagnose probable causes, and prioritize intervention opportunities based on production uplift potential and cost-benefit ratios. Bayesian networks and time-series forecasting models deployed as containerized inference services provide probabilistic assessments of intervention success rates. Production Optimization Services encompass virtual metering models, production forecasting neural networks, reinforcement learning control agents, and integrated asset optimization solvers. These services consume real-time production data from SCADA systems, execute short-term forecasts, and generate optimal setpoint recommendations for gas-lift rates, choke positions, and artificial lift parameters. The containerized deployment enables A/B testing of competing optimization strategies and seamless rollout of updated models without production disruptions.

3.3 GPU Acceleration and Resource Management

GPU acceleration plays a critical role in enabling real-time performance for compute-intensive simulation and deep learning workloads. The Kubernetes–OpenStack infrastructure supports GPU resource sharing through NVIDIA device plugins and Multi-Process Service (MPS), allowing multiple containers to efficiently utilize GPU compute capacity. Torque and drag simulations, neural network training, and reinforcement learning policy optimization benefit significantly from GPU acceleration, achieving speedups of 10-50x compared to CPU-only execution for typical well engineering problems. Resource quotas and limit ranges ensure fair allocation of GPU resources across multiple tenants while preventing resource starvation. Priority classes enable critical production optimization workloads to preempt lower-priority batch analysis jobs during resource contention. The infrastructure's demonstrated ability to maintain sub-100ms latency for AI inference requests

(Patchamatla, 2018) enables integration of ML models within interactive engineering applications and closed-loop control systems.

3.4 Data Management and Integration

Persistent storage for training datasets, simulation results, and model artifacts utilizes OpenStack Cinder volumes mounted to Kubernetes persistent volume claims (PVCs). This architecture ensures data durability across container lifecycle events while enabling efficient data sharing among related services. Time-series production data from SCADA systems flows through Apache Kafka message brokers, enabling real-time ingestion by production optimization services while maintaining data lakes for historical analysis and model retraining. Integration with enterprise data sources including drilling databases, completion records, production data warehouses, and geological interpretation platforms occurs through standardized API gateways and data connectors deployed as containerized services. This abstraction layer decouples engineering services from specific data source implementations, facilitating deployment across diverse IT environments and enabling gradual migration of legacy systems.

4. AI Model Development and Implementation

4.1 Well Completion Design Models

Machine learning models for well completion optimization employ ensemble approaches combining random forest, gradient boosting, and neural network architectures to predict production outcomes from completion design parameters. Feature engineering incorporates geological properties including porosity, permeability, total organic carbon (TOC), and mineralogy alongside completion parameters such as stage spacing, cluster spacing, proppant loading, and fluid volumes (Sheikhi Garjan & Ghaneezabadi, 2020; Liao et al., 2020). sModel training utilizes historical completion and production data from analogous wells, with careful attention to data quality, outlier detection, and temporal validation splits to prevent data leakage. Hyperparameter optimization employs Bayesian optimization or grid search executed as containerized batch jobs on Kubernetes, leveraging GPU acceleration for neural network training. Model interpretability tools including SHAP (SHapley Additive exPlanations) values and partial dependence plots provide engineering insights that build confidence in model recommendations and identify key completion design drivers (Sheikhi Garjan & Ghaneezabadi, 2020).

Figure 2 illustrates the complete AI-driven well completion optimization workflow, from data ingestion through deployment.

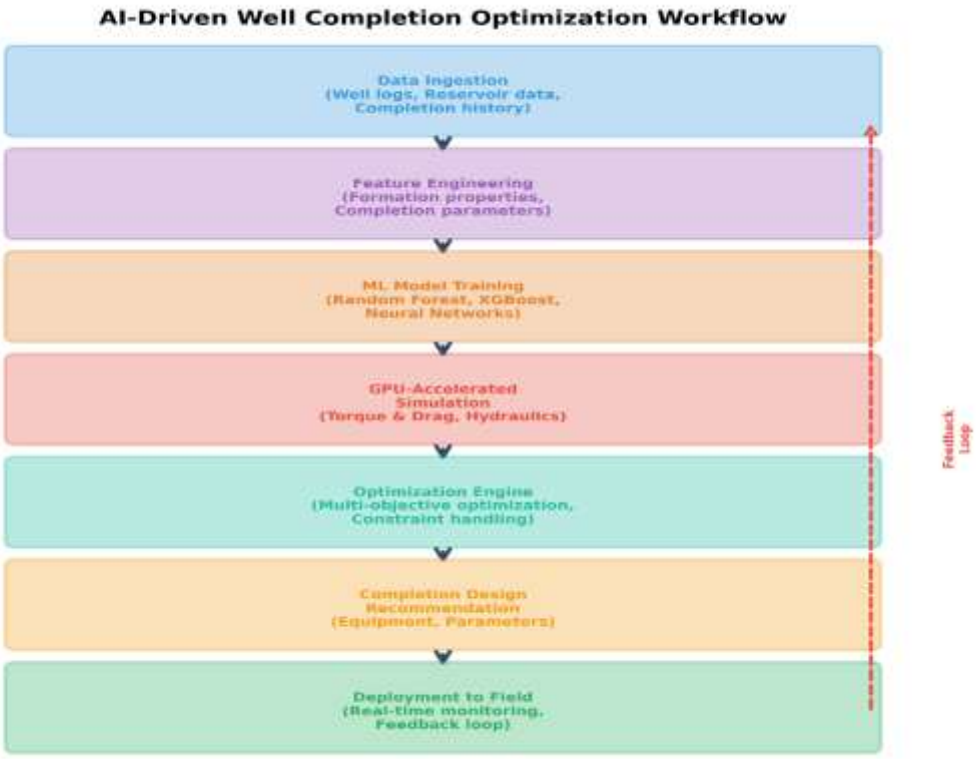

Figure 2: AI-Driven Well Completion Optimization Workflow

Deployment of trained completion models as containerized inference services enables integration with engineering design applications through REST APIs. The containerized architecture supports model versioning, A/B testing of competing models, and canary deployments that gradually shift traffic to updated models while monitoring prediction quality metrics. Kubernetes liveness and readiness probes ensure high availability by automatically restarting failed containers and routing traffic only to healthy instances.

4.2 Torque and Drag Simulation Integration

Torque and drag modeling represents a computationally intensive component of completion and workover design, particularly for extended reach drilling applications. Traditional finite element analysis approaches require substantial computational time when evaluating multiple trajectory and completion scenarios. GPU-accelerated simulation containers implement parallelized torque and drag solvers that distribute calculations across GPU cores, achieving near-real-time performance for typical ERD well profiles. Integration of ML surrogate models provides an additional acceleration pathway. Neural networks trained on extensive torque and drag simulation datasets learn to approximate full physics-based simulations with inference times measured in milliseconds rather than minutes (Tariq et al., 2020). The hybrid approach deploys both GPU-accelerated

physics simulators and ML surrogate models as containerized services, with intelligent routing logic selecting the appropriate computational pathway based on required accuracy, available computational budget, and design stage.

4.3 Workover Candidate Ranking Systems

Automated workover candidate identification and ranking systems integrate multiple data sources and analytical models within containerized pipeline architectures. Production data analysis containers implement decline curve analysis, rate transient analysis, and anomaly detection algorithms to identify wells exhibiting underperformance relative to type curves or offset well analogs. Well integrity monitoring containers consume corrosion logs, production chemistry data, and mechanical integrity test results to assess tubular condition and identify wells at risk of failure (Dallag et al., 2022). Bayesian network models deployed as inference containers integrate evidence from multiple sources to compute posterior probabilities for different failure modes and intervention success rates (Mata et al., 2021). Economic evaluation containers apply net present value (NPV) calculations incorporating production uplift forecasts, intervention costs, and risk-adjusted success probabilities to generate ranked candidate lists. The containerized pipeline architecture enables scheduled execution of the complete workflow, automatic updating of candidate rankings as new data arrives, and integration with workflow management dashboards.

4.4 Production Optimization and Reinforcement Learning

Reinforcement learning agents for production optimization learn control policies through interaction with reservoir simulation environments or data-driven proxy models. The training process involves substantial computational requirements, particularly for actor-critic algorithms that maintain both policy and value function neural networks (Poort et al., 2022). GPU-accelerated training containers execute thousands of simulation episodes, iteratively refining control policies to maximize cumulative production while respecting operational constraints including equipment limits, separator capacity, and pressure maintenance requirements. Trained RL agents deploy as containerized inference services that consume real-time production measurements and generate optimal setpoint recommendations at regular intervals (e.g., hourly or daily). The containerized architecture enables policy transfer, where agents trained on proxy models or historical data undergo fine-tuning on live production data through online learning mechanisms. Kubernetes deployment strategies including blue-green deployments and feature flags enable safe testing of updated policies without risking production disruptions. Virtual metering models implemented as containerized services provide critical inputs to production optimization workflows by estimating individual phase flow rates from readily available pressure and temperature measurements, reducing dependency on expensive multiphase flow meters (Al Selaiti et al.,

2020). Neural network-based virtual meters achieve accuracy within 5-10% of physical measurements while providing estimates at much higher temporal resolution, enabling more responsive optimization.

5. Deployment Strategies and Operational Workflows

5.1 Continuous Integration and Deployment Pipelines

The containerized architecture enables implementation of CI/CD pipelines that automate model training, validation, containerization, and deployment processes. Git-based version control manages model code, training scripts, and configuration files. Automated pipelines triggered by code commits execute model training jobs on Kubernetes batch resources, perform validation against holdout datasets, and generate performance metrics including prediction accuracy, inference latency, and resource consumption. Models meeting quality thresholds undergo automatic containerization using Docker build processes that package trained model artifacts, inference code, and dependencies into immutable container images. Container images push to private registries and undergo security scanning before deployment authorization. Kubernetes deployment manifests define resource requirements, scaling policies, health check endpoints, and service exposure configurations. GitOps workflows using tools such as ArgoCD ensure deployment state consistency across development, staging, and production environments.

5.2 Multi-Tenant Isolation and Resource Governance

Multi-tenant deployments serving multiple asset teams or business units require robust isolation mechanisms to prevent resource contention and ensure data security. Kubernetes namespaces provide logical isolation boundaries, with role-based access control (RBAC) policies restricting access to namespace resources. Resource quotas limit total CPU, memory, and GPU consumption per namespace, preventing individual tenants from monopolizing cluster resources. Network policies implement micro-segmentation, controlling which services can communicate across namespace boundaries. Sensitive well data and proprietary ML models receive additional protection through encryption at rest using OpenStack Barbican key management and encryption in transit via mutual TLS authentication between services. The multi-tenant architecture demonstrated by Patchamatla (2018) ensures that infrastructure optimizations benefit all tenants while maintaining strict isolation guarantees.

5.3 Digital Twin Integration

Digital twin platforms provide unified interfaces for visualizing well status, executing what-if analyses, and monitoring optimization recommendations. The containerized AI services integrate with digital twin platforms through API gateways that expose standardized interfaces for querying completion recommendations,

retrieving workover candidate rankings, and accessing production forecasts. WebSocket connections enable real-time streaming of optimization setpoints and model predictions to digital twin dashboards.

Figure 3 illustrates the integrated workover and production optimization system architecture, highlighting the central role of the digital twin in coordinating multiple AI services.

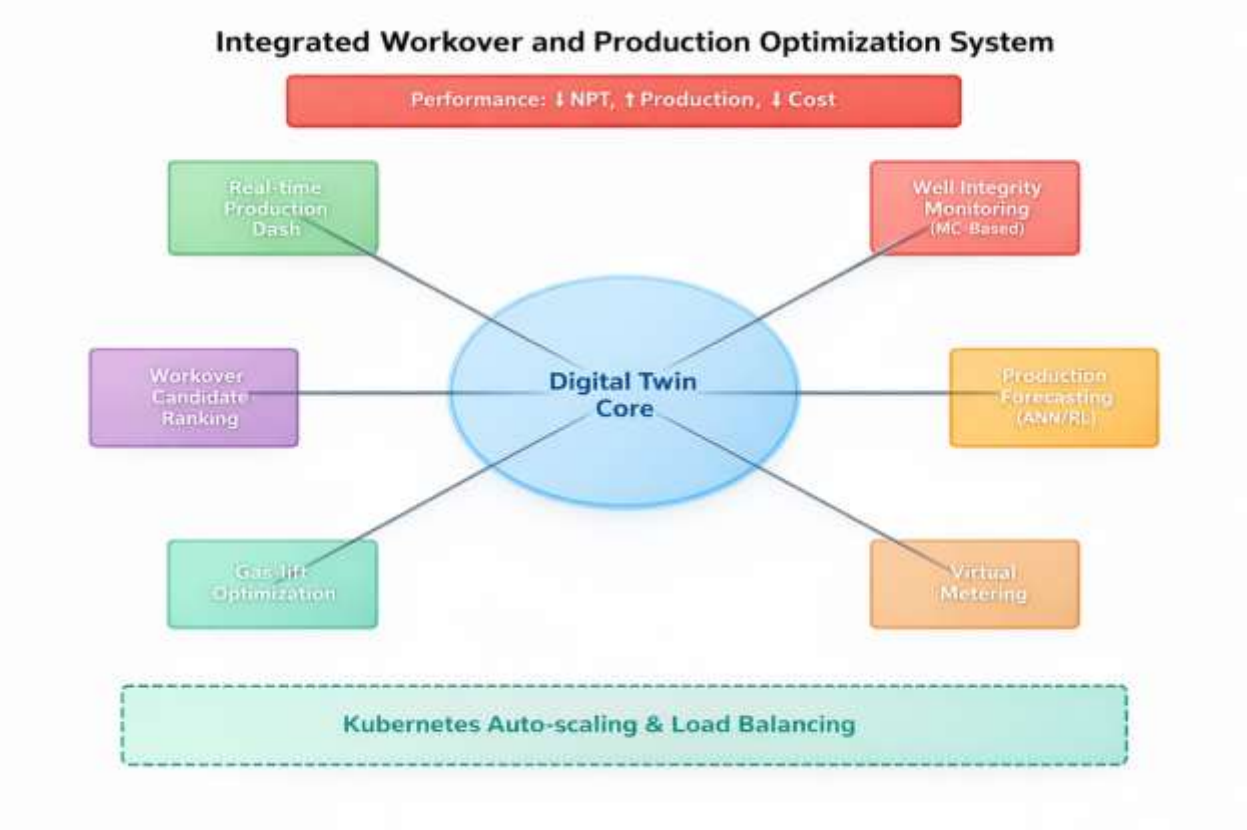

Figure 3: Integrated Workover and Production Optimization System

The digital twin maintains a synchronized representation of well state by consuming real-time data from SCADA systems, periodic updates from well integrity monitoring services, and on-demand simulation results from containerized physics engines. Engineers interact with the digital twin to explore completion design alternatives, evaluate workover scenarios, and approve production optimization recommendations before implementation. The containerized backend services provide the computational intelligence while the digital twin delivers the human-centered interface.

5.4 Monitoring and Observability

Production deployment of AI-driven well engineering services requires comprehensive monitoring and observability infrastructure. Prometheus metrics collection captures service-level indicators including request

rates, error rates, inference latency percentiles, and resource utilization. Custom metrics track domain-specific measures such as prediction accuracy drift, model confidence scores, and optimization objective function values. Distributed tracing using tools such as Jaeger provides visibility into request flows across multiple microservices, enabling diagnosis of performance bottlenecks and failure modes. Centralized logging aggregates container logs, facilitating troubleshooting and audit trail maintenance. Alert rules trigger notifications when metrics exceed acceptable thresholds, such as model prediction latency exceeding real-time requirements or accuracy degradation indicating model drift requiring retraining.

6. Performance Evaluation and Case Studies

6.1 Infrastructure Performance Characteristics

The Kubernetes–OpenStack infrastructure foundation delivers performance characteristics critical for real-time well engineering applications. Patchamatla (2018) demonstrated container startup latencies below 5 seconds, enabling rapid scaling in response to workload spikes. GPU sharing mechanisms achieved utilization rates exceeding 80% compared to 30-40% for dedicated GPU allocation, significantly improving cost efficiency for intermittent simulation workloads. Network throughput between containerized services exceeded 9 Gbps, ensuring data transfer does not bottleneck tightly coupled simulation pipelines. These infrastructure capabilities translate directly to application-level performance improvements. Completion design workflows that previously required 4-6 hours for comprehensive scenario evaluation complete in 30-45 minutes through parallel execution of containerized simulation services. Workover candidate screening processes that manually consumed 2-3 days per month execute automatically on weekly schedules with results available within 2 hours. Production optimization model inference latencies below 50 milliseconds enable integration within 1-minute control loop cycles for gas-lift optimization.

6.2 Well Completion Optimization Results

Deployment of ML-driven completion optimization in unconventional reservoirs demonstrates measurable performance improvements. Analysis of Montney formation completions using ensemble ML models achieved production prediction R^2 values exceeding 0.85 for first-year cumulative production (Liao et al., 2020). Model-recommended completion designs implemented in field trials showed 8-12% improvement in normalized production compared to offset wells using conventional design approaches (Sheikhi Garjan & Ghaneezabadi, 2020). The containerized deployment architecture enabled rapid iteration on model improvements, with updated models deployed to production within 24 hours of training completion. A/B testing frameworks comparing multiple model versions across different well cohorts provided statistically robust evidence of model performance, building confidence in AI-driven recommendations. The ability to execute hundreds of

completion scenario evaluations in parallel enabled comprehensive design space exploration that identified non-intuitive parameter combinations yielding superior performance.

6.3 Workover Optimization Case Study

Implementation of automated workover candidate generation and ranking systems in a mature offshore field containing over 700 wells demonstrated substantial operational efficiency gains. The containerized workflow reduced candidate identification cycle time from 15 days to 2 days, enabling monthly rather than quarterly candidate reviews (Mata et al., 2021). Machine learning models identified 23% more viable candidates compared to manual screening processes by detecting subtle production anomalies and integrity issues missed by conventional analysis. Economic analysis indicated that the improved candidate identification and prioritization resulted in 15% higher average production uplift per intervention by directing resources toward wells with greatest potential. The automated system also reduced dry hole interventions (workovers yielding no production improvement) by 30% through better integration of well integrity data and probabilistic success modeling (Mat Khair et al., 2023). The containerized architecture's ability to continuously update candidate rankings as new production data arrived ensured that engineering teams always worked from current information.

6.4 Production Optimization Performance

Reinforcement learning-based production optimization deployed across a portfolio of gas-lift wells achieved production improvements of 12-17% while reducing gas injection rates by 8%, demonstrating both revenue enhancement and operating cost reduction (Poort et al., 2022). The containerized RL inference services integrated with existing SCADA infrastructure, consuming real-time measurements and generating setpoint recommendations every 15 minutes. Operators retained override authority, with acceptance rates of AI recommendations exceeding 85% after initial validation period. Virtual metering services deployed as containerized inference engines provided flow rate estimates within 7% mean absolute percentage error (MAPE) compared to test separator measurements, enabling elimination of permanent multiphase flow meters on 40% of wells with estimated capital expenditure savings of $2.8 million across a 50-well field development (Al Selaiti et al., 2020). The high-frequency virtual metering data (1-minute intervals) enabled more responsive production optimization compared to traditional monthly well testing schedules.

6.5 Cost and Efficiency Analysis

Total cost of ownership analysis comparing the containerized Kubernetes–OpenStack deployment to traditional dedicated server infrastructure indicated 35-45% reduction in infrastructure costs through improved resource utilization and elimination of over-provisioning. GPU sharing capabilities reduced required GPU count by 60% for equivalent computational capacity, translating to hardware cost savings exceeding $500,000

for a 50-GPU cluster. Operational efficiency improvements including reduced planning cycle times and improved intervention success rates generated estimated value of $8-12 million annually for a 200-well asset. Development velocity improvements enabled by containerized CI/CD pipelines reduced time-to-deployment for new AI models from 4-6 weeks to 1-2 weeks, accelerating realization of model improvements and enabling more rapid response to changing field conditions. The standardized containerized deployment patterns reduced integration effort for new AI capabilities by approximately 50% compared to custom integration approaches, lowering barriers to adoption of emerging technologies.

7. Conclusion

This research demonstrates that containerized AI-driven systems built on Kubernetes–OpenStack infrastructure provide a robust, scalable platform for well completion, workover, and production optimization in petroleum engineering. The architecture translates infrastructure-level performance optimizations including GPU acceleration, multi-tenant resource management, and container orchestration into measurable operational improvements including reduced planning cycle times, improved intervention success rates, and enhanced production performance. The integration of machine learning models for completion design, workover candidate ranking, and production forecasting within containerized microservice architectures enables deployment patterns that were impractical with traditional monolithic applications. Continuous integration and deployment pipelines accelerate model improvement cycles, while A/B testing frameworks provide rigorous validation of AI recommendations. GPU-accelerated simulation containers deliver near-real-time performance for compute-intensive torque and drag modeling, enabling interactive engineering workflows and comprehensive scenario evaluation. Case study results from unconventional reservoir completions, mature field workover programs, and gas-lift optimization demonstrate economic value ranging from millions to tens of millions of dollars annually for typical assets. The containerized architecture's flexibility enables deployment across diverse environments from on-premise data centers to public cloud platforms, accommodating varying IT policies and data governance requirements across operating companies.

Future research directions include extension of reinforcement learning approaches to integrated field-wide optimization incorporating reservoir management, facilities constraints, and market conditions. Integration of physics-informed neural networks that embed conservation laws and domain knowledge into ML model architectures promises to improve generalization and reduce training data requirements. Development of federated learning frameworks that enable collaborative model training across multiple operators while preserving data privacy could accelerate AI adoption across the industry. The convergence of artificial intelligence, containerization, and cloud orchestration technologies represents a transformative opportunity for petroleum engineering. This research establishes practical frameworks and demonstrates measurable value,

providing a foundation for broader industry adoption of containerized AI-driven well engineering optimization systems.

Acknowledgment

The authors express their sincere gratitude to all who supported this research. We thank our institutions for providing the necessary facilities and environment, and we appreciate the reviewers and editors for their valuable feedback. We also acknowledge the helpful input of colleagues and collaborators, as well as the encouragement and support of our families throughout this work.

Disclosure of Interest

The authors declare no competing financial, personal, or organizational interests related to this work.

Funding Information

This research received no financial support from any funding agency, institution, or commercial organization. The authors confirm that the study was conducted using personal or institutional resources, with no grants or project funding from public, private, or non-profit sectors.

References

Al Selaiti, I., Mata, C., Saputelli, L., Badmaev, D., & Alatrach, Y. (2020). Robust data driven well performance optimization assisted by machine learning techniques for natural flowing and gas-lift wells in Abu Dhabi. *SPE Paper 201696-MS*. https://doi.org/10.2118/201696-MS

Baki, S., Temizel, C., & Dursun, S. (2021). Well completion optimization in unconventional reservoirs using machine learning methods. *SPE Paper 206241-MS*. https://doi.org/10.2118/206241-MS

Dallag, M., Bawazir, M., & Al-Ali, A. K. (2022). Digital solution to extend the life of wells with continuous corrosion monitoring based on machine learning algorithms. *IPTC Paper 22472-MS*. https://doi.org/10.2523/iptc-22472-MS

Joseph, C. (2013). From fragmented compliance to integrated governance: A conceptual framework for unifying risk, security, and regulatory controls. *Scholars Journal of Engineering and Technology, 1*(4), 238–250.

Liao, L., Li, G., Zhang, H., Feng, J., & Zeng, Y. (2020). Well completion optimization in Canada tight gas fields using ensemble machine learning. *SPE Paper 202966-MS*. https://doi.org/10.2118/202966-MS

Lu, C., Jiang, H., Yang, J., Wang, Z., & Zhang, M. (2022). Shale oil production prediction and fracturing optimization based on machine learning. *Journal of Petroleum Science and Engineering*, *210*, 110900. https://doi.org/10.1016/j.petrol.2022.110900

Lyu, T. (2020). *Evaluation of containerized simulation software in Docker Swarm and Kubernetes* [Master's thesis, Aalto University]. Aalto University Digital Repository.

Mat Khair, N., Zaizakrani, M. F., Zul Azhar, N. A. I., Halim, F. L. A., & Sidek, S. (2023). Automated production enhancement candidates screening powered by machine learning unlocks untapped potential in matured oil fields – A case study. *SPE Paper 215244-MS*. https://doi.org/10.2118/215244-MS

Mata, C., Saputelli, L., Badmaev, D., Zhao, W., & Mohan, R. (2021). Automated reservoir management workflows to identify candidates and rank opportunities for production enhancement and cost optimization in a giant field in offshore Abu Dhabi. *OTC Paper 31295-MS*. https://doi.org/10.4043/31295-MS

Patchamatla, P. S. (2018). Optimizing Kubernetes-based multi-tenant container environments in OpenStack for scalable AI workflows. *International Journal of Advanced Research in Education and Technology (IJARETY)*, *5*(3). https://doi.org/10.15680/ijarety.2018.0503002

Poort, J. P., van der Waa, J., Mannucci, T., & Shoeibi Omrani, P. (2022). An optimum well control using reinforcement learning and policy transfer; application to production optimization and slugging minimization. *SPE Paper 210277-MS*. https://doi.org/10.2118/210277-MS

Shankar, R., Sasirekha, G. V. K., Ramanathan, C., & Bapat, J. (2022). Knowledge-based digital twin for oil and gas 4.0 upstream process: A system prototype. *2022 International Conference on Internet of Things and Intelligence System (IoTaIS)*, 206-213. https://doi.org/10.1109/IoTaIS56727.2022.9975974

Sheikhi Garjan, Y., & Ghaneezabadi, M. (2020). Machine learning interpretability application to optimize well completion in Montney. *SPE Paper 200019-MS*. https://doi.org/10.2118/200019-MS

Shelley, R., Oduba, O., & Melcher, H. (2021). Machine learning and artificial intelligence provides Wolfcamp completion design insight. *SPE Paper 204199-MS*. https://doi.org/10.2118/204199-MS

Singh, H., Li, C., Cheng, P., Wang, X., & Hao, G. (2023). Real-time optimization and decarbonization of oil and gas production value chain enabled by Industry 4.0 technologies: A critical review. *SPE Production & Operations*, *38*(2), 257-282. https://doi.org/10.2118/214301-PA

Tariq, Z., Mahmoud, M., & Abdulraheem, A. (2020). Real-time prognosis of flowing bottom-hole pressure in a vertical well for a multiphase flow using computational intelligence techniques. *Journal of Petroleum Exploration and Production Technology*, *10*(3), 1411-1428. https://doi.org/10.1007/S13202-019-0728-4

Višňovský, V., Spišaková, V., Hozzová, J., Oľha, J., & Trapl, D. (2022). Complex simulation workflows in containerized high-performance environment. *Policy and Complex Systems*, *7*(2), 32-47. https://doi.org/10.18278/jpcs.7.2.4

Vorobev, I., Koshkin, T., Prokopev, M., Rakhmangulov, Y., & Kombarov, S. (2022). Digital twin application for boosting oil production, predictive analytics of asset integrity and mid-term forecasting of field performance. *SPE Paper 211105-MS*. https://doi.org/10.2118/211105-MS

Open Access Statement