Cloud-Agnostic Container Processing Pipeline for KinaTrax: Enabling Same-Day Biomechanical Analysis Through Massively Parallel Event Streaming

Peter Winkler
Computer Science Master's Program
Full Sail University
Winter Park, Florida, USA
Email: peter.winkler@fullsail.edu

***Abstract*—The current markerless motion capture system of KinaTrax has the capability of undertaking the analysis of the biomechanical data of professional baseball through serial batch processing, resulting in a latency of over 12 hours after the end of the games before the analysis is completed. The project seeks to change the architecture of the system into a cloud-native parallel stream processing system that is capable of undertaking the analysis of data contained in 9,600 videos concurrently, ensuring that the latency of the analysis is the same day. The system also seeks to introduce "GCAS (Game-Context Auto-Scaling)," which is a predictive scaling AI that utilizes the schedules of the games of the MLB and the behavior of the data over time to ensure that the resources are scaled ahead of time, thus avoiding the lag that occurs after two to five minutes with the normal scalers. The project utilizes a mixed-methodology that entails the use of both performance studies (A/B tests, correlation tests, experiments to evaluate the scaling of throughput, and other quantitative studies) and the use of user studies (usability tests and other tests involving the interaction with n=10-12 coaches/trainers within the baseball world to ensure the satisfaction of the three hypotheses of the project that include the fact that the project would "massively parallelize and thus greatly reduce latency of critical events," "the cloud analysis has a correlation of accuracy of $\geq$95% to the pre-existing baseline of the on-premise analysis of the data through the normal batch system of the computer environment of the company," and that the GCAS model "scales cloud resources more cheaply than a naive always-on model of cloud acquisition and operation". The project also aims to contribute towards AI and the field of data science through**

***Index Terms*—cloud computing, container orchestration, biomechanical analysis, auto-scaling, machine learning, sports analytics, Kubernetes, parallel processing**

I. Introduction

A. Problem Definition

The current markerless motion capture system developed by KinaTrax takes a serial batch-processing approach that currently provides a lag of over 12 hours between the end of the game and the availability of the data. Based on a standard baseball game that produces roughly 600 biomechanical events recorded through 8-16 high-speed cameras (resulting in a total of 9,600 video files), the current system processes the video files one after the other, taking roughly 27 seconds per video file.

Research Question: Can the KinaTrax serial batch-processing architecture be adapted into a cloud-native parallel and asynchronous data stream-processing framework that provides accurate biomechanical analysis and efficiently utilizes resources?

Testable Hypothesis: By implementing a cloud-agnostic container orchestration platform with AI-driven predictive scaling (GCAS: Game-Context Auto-Scaling), we hypothesize that we can:

1) Drastically reduce critical event processing latency through massively parallel processing
2) Maintain $\geq$95% accuracy correlation with the current on-premises system
3) Reduce cloud infrastructure costs through predictive scaling compared to naive always-on provisioning
4) Enable new use cases including same-day post-game review and next-day training adjustments

B. Motivation

This study fills a valuable niche within the field of real-time sport biomechanics because the relationship between complexity and the delivery of key information is vitally important. Professional baseball organizations view sport biomechanics as highly important for analyzing and improving the performance of baseball pitchers. The data currently has a latency of over 12 hours.

Stakeholders:

- **MLB Coaching Staffs:** Same-day biomechanical feedback enables immediate technique corrections and injury prevention
- **Sports Scientists:** Validates cloud-native approaches for computationally intensive biomechanical analysis

- **KinaTrax/Sony:** Scalable infrastructure supporting 100+ MLB stadiums and expansion into other sports (post-October 2024 acquisition)
- **Cloud Computing Researchers:** Novel application of game-context aware auto-scaling in sports analytics domain

Significance: The key challenge is not the computational power but architectural issues. The fact that the computer currently performs serially is a historical relic, not a computational imperative. Mass parallel computing could result in speeds of magnitude increases through the concurrent analysis of thousands of video sequences, turning the analysis of biomechanics into real-time intelligence.

C. Results Summary

[To be completed upon research conclusion. Expected outcomes include: (1) empirical validation of drastic latency reduction through parallel processing, (2) correlation coefficient $r > 0.95$ with baseline accuracy, (3) measured cost reduction through predictive scaling, and (4) user satisfaction improvements in post-deployment surveys.]

II. Background

A. Historical Context: Biomechanical Analysis in Professional Sports

Markerless motion capture technology has greatly impacted the field of sports biomechanics, allowing for accurate analysis free of the invasiveness of marker placement and offering the same level of accuracy that traditional marker-based systems afford [1], [2]. The proprietary system developed by KinaTrax, used across over 100 MLB stadiums, involves the use of 8-16 cameras that take high-speed video at a rate of over 300 FPS, analyzing each video through computer algorithms that calculate the position of joints within a 3D environment.

Since the original architectures were a result of computational necessities of the earlier model, parallel computing couldn't be considered because of the unavailability of GPU resources and the resultant data bandwidth. The cloud environment nullifies these necessities and provides a scope for re-engineering the computing pipelines.

B. Related Work and Existing Solutions

Cloud-Based Biomechanical Processing: There have been a few studies done that attempt to utilize cloud deployment for the analysis of motion captures [3], [4]. These studies are centralized around the idea of improving the computational performance of the task using enhanced hardware resources. Cloud computing is considered 'faster on-premises.'

Auto-Scaling Systems: Auto-scaling of cloud infrastructures usually depends on reactive parameters such as CPU usage, Memory Usage, and queue size with a latency of 2-5 minutes after the detection of the workload [5]. As the pattern of the workload is already observed for the scheduled sport events, a reactive mechanism of scaling is inefficient.

Sports Analytics Platforms: Commercial systems (e.g., Catapult, Vald Performance) also exist that support cloud-basedathlete tracking but are mostly used for lower-bandwidth sensor data and not high-resolution videoprocessing. Large-scale video processing of the kind performed at KinaTrax (representing 9,600videos per game) remains a difficult task.

C. Barriers to Existing Solutions

Why Hasn't This Been Solved?

1) **Legacy Architecture Inertia:** Current systems work adequately for post-hoc analysis, reducing urgency for architectural overhaul
2) **Accuracy Preservation Uncertainty:** Concern that cloud containerization might introduce processing variability affecting biomechanical measurement precision
3) **Cost Unpredictability:** Fear that cloud processing could exceed on-premises costs without sophisticated resource management
4) **Operational Complexity:** Multi-cloud orchestration, priority queuing, and real-time monitoring add engineering complexity

Why Do Naive Approaches Fail?

- **Simple Cloud Migration:** Moving serial processing to cloud doesn't solve latency; it just makes serial processing more expensive
- **Reactive Auto-Scaling:** By the time reactive systems detect load and provision resources, processing is already delayed
- **Always-On Provisioning:** Maintaining 1,600+ containers 24/7 is cost-prohibitive given sporadic game schedules

D. Key Frameworks and Theories

Amdahl's Law and Parallel Speedup: The theoretical parallelization speedup is hampered by the serial portion of the workload. The theoretical parallelization speedup for the video processing portion of KinaTrax's workload appears to be linear and near optimal (approaching the number of video streams, which is a theoretical maximum of 1,600x).

Multi-Objective Optimization: The task of minimizing latency, preserving accuracy, and minimizing costs cannot be addressed using the traditional single objective optimization approach (either minimize costs OR minimize latency), which cannot handle the trade-offs that exist in real-life situations [6].

Human-Computer Interaction in Time-Critical Domains: Studies of notification schemes and dashboard interfaces for time-critical decision-making [7] inform the HCI component of the model concerning asynchronous processing feedback and priority request interfaces.

III. Methodology

A. System Design: Cloud-Agnostic Container Orchestration Platform

The proposed architecture transforms serial batch processing into a massively parallel event streaming pipeline.

1) Core Components: **1. Event Stream Ingestion Layer**

- Kafka/Kinesis event queue for real-time video upload from stadiums
- Two-tier priority classification: Critical (pitch mechanics) vs. Routine (warmup throws)
- Multi-stadium ingestion supporting 100+ concurrent game locations

2. GCAS: Game-Context Auto-Scaling (Primary AI Innovation)

- Predictive scaling using MLB game schedules + historical load patterns
- Proactive resource provisioning 30-60 minutes before expected load spikes
- Eliminates 2-5 minute reactive lag of traditional auto-scalers
- *Features:* Game date/time, opponent, stadium, historical event volume
- *ML Model:* Linear Regression (baseline) → Random Forest (if time permits)
- *Target:* Predict processing load (events/minute) to pre-scale Kubernetes pods

3. Kubernetes Container Orchestration

- Horizontal Pod Autoscaler (HPA): Dynamic scaling from 0 to 1,600+ containers based on demand
- Docker containerization of KinaTrax's Track Joint (TJ) biomechanical analysis tools
- Cloud-agnostic deployment (AWS EKS primary; Azure AKS, GCP GKE optional)
- GPU-enabled containers for computer vision workloads

4. Processing Pipeline

- *Video Generation (VG):* 4-5 hours (parallel execution: hours → minutes)
- *Track Joint (TJ) Biomechanical Analysis:* 6-10 min/event (massively parallel)
- Priority queue ensures critical events processed before routine events
- 100% accuracy validation against on-premises baseline

5. Real-Time Monitoring Dashboard (HCI Component)

- Grafana-based visualization: queue depth, processing progress, ETA, cost tracking
- Per-event status tracking with estimated completion times
- Multi-cloud cost comparison (AWS vs. Azure vs. GCP for identical workloads)
- User-initiated priority requests for expedited processing

6. Results Delivery & Notification System

- Intelligent notification system: Email/SMS when results available
- Integration with existing KinaTrax coach/trainer workflows
- Asynchronous processing paradigm: Submit → Notification → Review

2) Technology Stack:

- **Orchestration:** Kubernetes (EKS/AKS), Docker, Helm
- **Streaming:** Apache Kafka OR AWS Kinesis
- **Queue:** Redis (priority queue management)
- **Monitoring:** Prometheus, Grafana, Jaeger
- **ML/AI:** Python (scikit-learn, TensorFlow/PyTorch), Flask API
- **Databases:** PostgreSQL (metadata), InfluxDB (time-series)
- **IaC:** Terraform (multi-cloud deployment)

B. Research Methods

This study will utilize a mixed-methods approach that integrates performance analysis, statistical validation, and user studies. Because of the baseball off-season, the validation will be performed through simulation of games using historical data rather than a real-time field study. This historical data provides a wealth of information involving ground truth data for the performance analysis of the KinaTrax system within baseball parks during the period of the 2022 and 2024 baseball seasons. The details of the research methodology are described below.

1) Measurement Instruments:

- **Performance Metrics:** Custom logging (Prometheus + InfluxDB)
- **Usability:** System Usability Scale (SUS) - validated 10-item questionnaire
- **Workload:** NASA-TLX (Task Load Index) - validated cognitive workload measure
- **Qualitative:** Semi-structured interviews with thematic coding

2) Sample Sizes and Statistical Power:

- **Quantitative Performance:** n=50-100 games from historical data (well-powered for paired t-tests with α=0.05, β=0.20, d=2.0)
- **Accuracy Validation:** n=1,000+ events across multiple stadiums and teams (robust Pearson correlation power analysis)
- **GCAS Training:** n=100-200 historical games for ML model development
- **User Research:** n=10-12 participants (qualitative saturation using dashboard prototypes with simulated real-time data)

3) Validity Threats and Mitigation:

- **Internal Validity:** Controlled experiments using identical historical game data for on-prem vs. cloud comparison; stratified sampling across game complexity (regular season vs. playoffs)
- **External Validity:** Historical data from 10+ stadiums across multiple teams and seasons (2022-2024) ensures generalizability; simulation results validated against documented on-premises performance
- **Ecological Validity:** Simulation accurately replicates production event streams and processing workloads; user studies with dashboard prototypes using realistic data playback
- **Construct Validity:** Validated instruments (SUS, NASA-TLX); ground-truth comparison to existing on-premises results

TABLE I
RESEARCH DESIGN: PROBLEMS, HYPOTHESES, METHODS, AND EXPECTED OUTCOMES

Research Problem	Testable Hypothesis	Method for Testing	Planned Analyses	Expected Outcome
Latency Reduction	Parallel processing drastically reduces event processing latency	A/B testing using historical game data: On-prem (documented baseline) vs. Cloud (simulated processing) on n=50-100 identical games	Paired t-test, Mann-Whitney U; Metrics: P50, P95, P99 latency	Significant reduction (α=0.05, power=0.80) with large effect size
Accuracy Preservation	Cloud processing maintains $\geq$95% correlation with on-prem baseline	Correlation analysis using historical data: Cloud biomechanical analysis vs. ground-truth on-prem results on n=1,000+ events across multiple stadiums	Pearson r, Bland-Altman plots, MAE	$r > 0.95$, $p < 0.001$; Bland-Altman limits within ± 2
Cost Optimization	GCAS reduces costs vs. always-on provisioning	Simulated cost tracking over 50+ historical games: GCAS predictive vs. reactive vs. always-on provisioning strategies	ANOVA, cost-per-event comparison, ROI analysis	GCAS reduces costs significantly ($p < 0.05$) compared to baseline strategies
Parallelization Efficiency	Speedup scales linearly with container count	Controlled experiments: Vary containers (10-1600); measure throughput	Amdahl's Law validation, speedup curves, strong scaling	Near-linear speedup to 1,000 containers; diminishing returns beyond
User Satisfaction	Dashboard improves decision-making speed and satisfaction	User studies with dashboard prototypes using simulated real-time data playback: Usability testing, interviews (n=10-12), pre/post surveys	SUS, NASA-TLX, thematic analysis	SUS ¿80, task time reduced ¿40%, positive qualitative feedback
GCAS Prediction Accuracy	ML model: MAE ¡30 events, R^2 ¿0.70	Train on 100-200 historical games, validate on holdout 20%; compare against reactive baseline	MAE, R^2, RMSE, feature importance, cross-validation	MAE ¡30 events/min, R^2 ¿0.70, beats reactive baseline by ¿15%

- **Reliability:** Inter-rater reliability for qualitative coding (Cohen's κ ¿0.80); reproducible simulation experiments

IV. POTENTIAL ANALYSES

A. Performance Analysis

Latency Distribution: Histogram and kernel density estimation of processing latency for control (on-prem) vs. treatment (cloud). Quantile-quantile (Q-Q) plots to assess normality. Bootstrap confidence intervals for median latency differences.

Throughput Scaling:

- *Strong scaling:* Fixed workload (1,600 videos), vary container count

- *Weak scaling:* Proportionally increase workload with containers
- *Speedup efficiency:* $S(n) = T(1)/T(n)$ where n = container count

Cost-Performance Pareto: Multi-objective optimization plotting latency vs. cost trade-off curves. Identify Pareto-optimal operating points. Sensitivity analysis for cloud pricing fluctuations.

B. Accuracy Validation

Correlation Analysis: Pearson r with 95% confidence intervals. Scatter plots: Cloud joint angles vs. on-prem (reference). Residual analysis for systematic bias.

Bland-Altman Agreement: Mean difference and limits of agreement (± 1.96 SD). Proportional bias assessment. Clinical significance: $\pm 2^\circ$ threshold.

Error Distribution: MAE, RMSE, error percentiles (50th, 95th, 99th). Outlier detection (errors ¿3 SD from mean).

C. Machine Learning Model Evaluation

GCAS Prediction Accuracy: Time series cross-validation (rolling window). Metrics: MAE, RMSE, R^2, MAPE. Feature importance analysis (Random Forest). Residual plots: homoscedasticity and normality checks.

Model Comparison: Baseline (reactive scaling) vs. GCAS v1 (Linear Regression) vs. GCAS v2 (Random Forest). Metric: Reduction in resource provisioning lag.

D. User Experience Analysis

Quantitative Usability: System Usability Scale (SUS): mean, SD, percentile ranking. Task completion time: pre vs. post deployment. Error rate during dashboard interaction. Adoption rate: percentage using priority requests.

Qualitative Analysis: Thematic coding of interviews (2 independent coders, Cohen's κ). Emergent themes: workflow changes, decision-making impacts, feature requests. Case studies: documented same-day coaching adjustments.

NASA-TLX Workload: Six subscales: mental demand, physical demand, temporal demand, performance, effort, frustration. Pre/post comparison using paired t-tests. Hypothesis: Reduced temporal demand and frustration.

E. Cost-Benefit Analysis

Total Cost of Ownership (TCO): Cloud infrastructure (compute, storage, networking). Operational costs (engineering maintenance). Comparison: Cloud TCO vs. on-prem depreciation.

ROI: Quantifiable benefits (reduced injury rates from faster feedback). Qualitative benefits (competitive advantage, satisfaction). Break-even analysis: games processed before cost-effectiveness.

Sensitivity Analysis: Impact of cloud pricing fluctuations ($\pm 20\%$). Scaling scenarios: 200+ stadiums or other sports.

V. Expected Contributions

This research will contribute to three distinct domains:

A. AI/Machine Learning

- Novel application of game-context aware predictive scaling in sports analytics
- Empirical validation of domain-specific features for resource prediction
- Comparison of model complexity in time-critical deployment contexts

B. Human-Computer Interaction

- Design patterns for asynchronous processing dashboards in sports domains
- User research on priority request interfaces for biomechanical workflows
- Notification system design balancing immediacy with interruption costs

C. Data Science/Systems

- Empirical analysis of parallelization efficiency at massive scale (1,600+ containers)
- Multi-objective optimization framework for latency-accuracy-cost trade-offs
- Validation methodology for cloud-native biomechanical analysis accuracy

D. Practical Impact

- Production-ready cloud-agnostic platform deployable to 100+ MLB stadiums
- Open-source reference architecture for massively parallel video processing
- Cost-benefit analysis informing KinaTrax/Sony cloud infrastructure investment

VI. Timeline and Milestones

Month 1 (Foundation): AWS EKS deployment, Docker containerization, historical data ingestion pipeline, 100% accuracy baseline validation using ground-truth on-premises results.

Month 2 (GCAS Development): GCAS v1 (Linear Regression) training on 100-200 historical games, Kubernetes integration, Grafana dashboard implementation with simulated real-time data playback, priority queue development.

Month 3 (Simulation Validation): Large-scale simulation experiments (50-100 games), parallelization scaling tests (10-1,600 containers), cost comparison analysis, user studies with dashboard prototypes (n=10-12), usability testing.

Month 4 (Analysis & Documentation): Statistical analysis across all research questions, performance validation, cost-benefit assessment, thesis completion (progressive writing across all months: literature review → methods → results → discussion).

VII. Limitations and Future Work

A. Scope Limitations

MVP Focus: Single cloud provider (AWS), Linear Regression GCAS, Grafana dashboard, simulation-based validation using historical game data.

Stretch Goals (Deferred): Multi-cloud (Azure/GCP), Random Forest GCAS, MLB Statscast API automation, custom React dashboard, live stadium deployment.

B. Methodological Limitations

Simulation-Based Validation: Using historical game data provides controlled experimental conditions and larger sample sizes, but does not capture real-time operational challenges (e.g., network variability, live camera failures). However, simulation enables reproducible experiments and eliminates dependency on MLB season scheduling.

Advantages of Simulation Approach:

- Larger scale testing (50-100+ games vs. 15-20 in live pilot)
- Perfect reproducibility for comparative experiments
- Immediate start (no waiting for baseball season)
- Ground-truth validation against documented on-premises results
- Controlled variable manipulation for causal analysis

C. Known Risks

- **Month 2 Timeline:** GCAS development + dashboard + integration = tight schedule
- **Historical Data Access:** Depends on KinaTrax/Sony data sharing approval
- **Simulation Fidelity:** Must accurately replicate production event timing and workload patterns

D. Future Research Directions

- **Live Stadium Deployment:** Validate simulation findings with real-time production deployment during MLB season
- Real-time in-game processing (vs. post-game batch processing)
- Multi-sport generalization (NFL, NBA, soccer biomechanics)
- Injury prediction models using longitudinal biomechanical patterns
- Edge computing hybrid (stadium-local + cloud elasticity for bandwidth optimization)

VIII. Conclusion

This work offers a revolutionary paradigm shift for sports biomechanical analysis through the re-conception of the KinaTrax data processing architecture. Contrary to the perspective of cloud computing merely offering "faster hardware resources," the project takes a cloud native approach to parallel computation to enable drastic latency reduction and a paradigm shift for coach decision-support. The addition of the GCAS (Game Context Auto Scaling) component offers a new addition to the field of predictive resource scaling within sports analytics, and a complete and accurate methodology provides for the preservation of accuracy and feasibility. Successful work will result in a production-level platform that is deployable within the kinatrax networks of venues.

Acknowledgments

This research is conducted in partnership with KinaTrax (acquired by Sony Hawk-Eye, October 2024) with mentorship from Derek Smith (KinaTrax CTO) and Rob Neyer (VP of Baseball Operations). The project leverages KinaTrax's existing biomechanical analysis infrastructure deployed across 100+ MLB stadiums.

References

[1] J. A. Turner, C. R. Chaaban, and D. A. Padua, "Validation of OpenCap: A low-cost markerless motion capture system for lower-extremity kinematics during return-to-sport tasks," *Journal of Biomechanics*, vol. 171, p. 112200, Jun. 2024, doi: 10.1016/j.jbiomech.2024.112200.

[2] G. S. Fleisig et al., "Comparison of marker-less and marker-based motion capture for baseball pitching kinematics," *Sports Biomechanics*, vol. 23, no. 12, pp. 2950–2959, Dec. 2024, doi: 10.1080/14763141.2022.2076608.

[3] C. Martínez et al., "Cloud-Native GPU-Enabled Architecture for Parallel Video Encoding," in *Euro-Par 2024: Parallel Processing*, Cham: Springer Nature Switzerland, 2024, pp. 319–333, doi: 10.1007/978-3-031-69583-4_23.

[4] M. C. Luizelli et al., "Online architecture for predicting live video transcoding resources," *Journal of Cloud Computing*, vol. 8, no. 1, p. 13, Aug. 2019, doi: 10.1186/s13677-019-0132-0.

[5] L. Toka, G. Dobreff, B. Fodor, and B. Sonkoly, "Machine Learning-Based Scaling Management for Kubernetes Edge Clusters," *IEEE Trans. Network and Service Management*, vol. 18, no. 1, pp. 958–972, Mar. 2021, doi: 10.1109/TNSM.2021.3052837.

[6] M. Zhao et al., "Multi-objective workflow scheduling in cloud computing: trade-off between makespan and cost," *Cluster Computing*, vol. 25, no. 1, pp. 579–597, Feb. 2022, doi: 10.1007/s10586-021-03432-y.

[7] H. Alsulmi et al., "Optimizing Clinical Decision Support System Functionality by Leveraging Specific Human-Computer Interaction Elements: Insights From a Systematic Review," *JMIR Human Factors*, vol. 12, p. e69333, Jan. 2025, doi: 10.2196/69333.

[8] R. M. Kanko et al., "Comparison of markerless and marker-based motion capture systems using 95% functional limits of agreement in a linear mixed-effects modelling framework," *Scientific Reports*, vol. 13, no. 1, p. 22880, Dec. 2023, doi: 10.1038/s41598-023-49360-2.

[9] A. Chavan et al., "Predictive Container Auto-Scaling for Cloud-Native Applications," in *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, Sydney, Australia: IEEE, Dec. 2019, pp. 143–148, doi: 10.1109/CloudCom.2019.00033.

[10] M. Usman et al., "Normative In-Game Data for Collegiate Baseball Pitchers Using Markerless Tracking Technology," *Journal of Athletic Training*, vol. 59, no. 11, pp. 1069–1076, Nov. 2024, doi: 10.4085/1062-6050-0134.24.