Optimizing Resource Provisioning Across Diverse Computing Facilities with Virtual Kubelet Integration

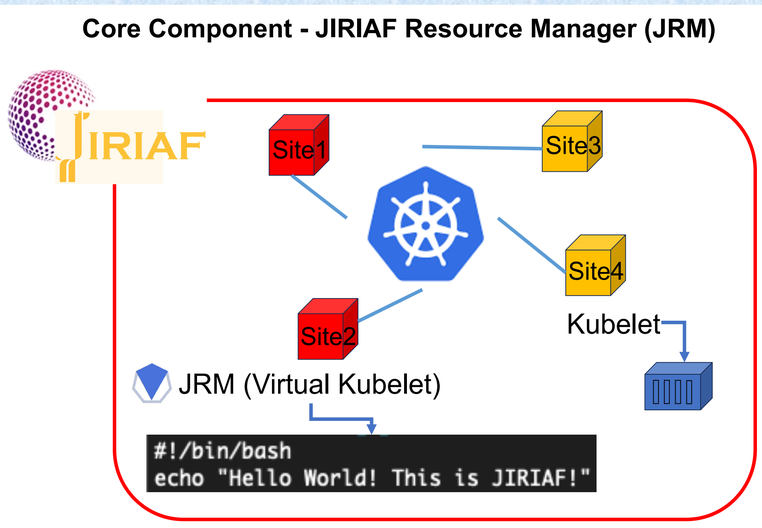
Authors: Jeng-Yuan Tsai, Vardan Gyurjyan, Graham Heyes, Christopher Larrieu, David Lawrence, Patrick Meagher
Contact: tsai@jlab.org

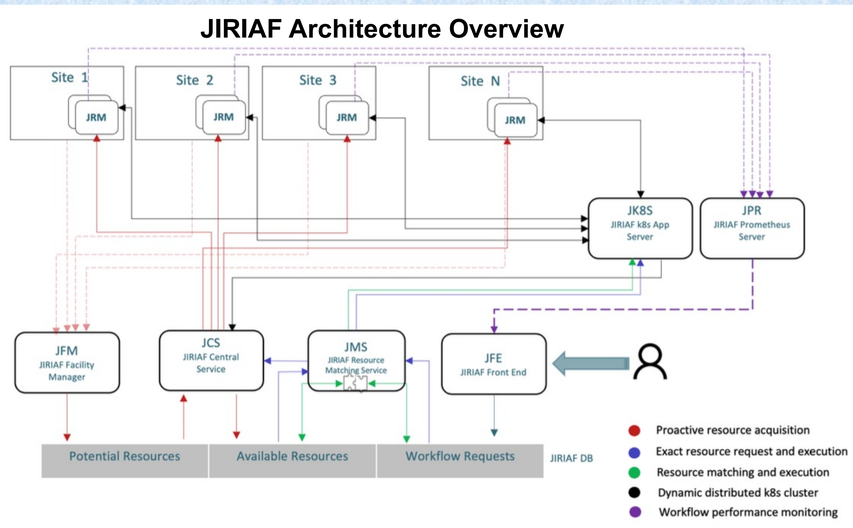
Abstract

The Jefferson Lab Integrated Research Infrastructure Across Facilities (JIRIAF) offers a scalable framework for resource management across high-performance computing (HPC) environments. Using Virtual Kubelet, JIRIAF enables Kubernetes orchestration without root access, streamlining resource allocation across diverse sites. Key components like the JIRIAF Facility Manager and Resource Manager dynamically pool and distribute resources based on real-time demands, optimizing workload management. A proof-of-concept deployment on NERSC's Perlmutter demonstrated high throughput and adaptability in remote data-stream processing. A digital twin model was also integrated, enhancing state estimation and adaptive control, which supports improved performance in stream-processing workflows. This study highlights JIRIAF's effectiveness in distributed infrastructure management with potential for expansion into more complex applications.

Core Component - JIRIAF Resource Manager (JRM)

- Kubelet Integration: JRM leverages Virtual Kubelet[1,2] to run Kubernetes in environments without root access. It translates containers into BASH scripts, enabling seamless resource management across distributed HPC environments.
- Userspace Execution: JRM can run applications as containers using BASH commands, ensuring user control without requiring root permissions.

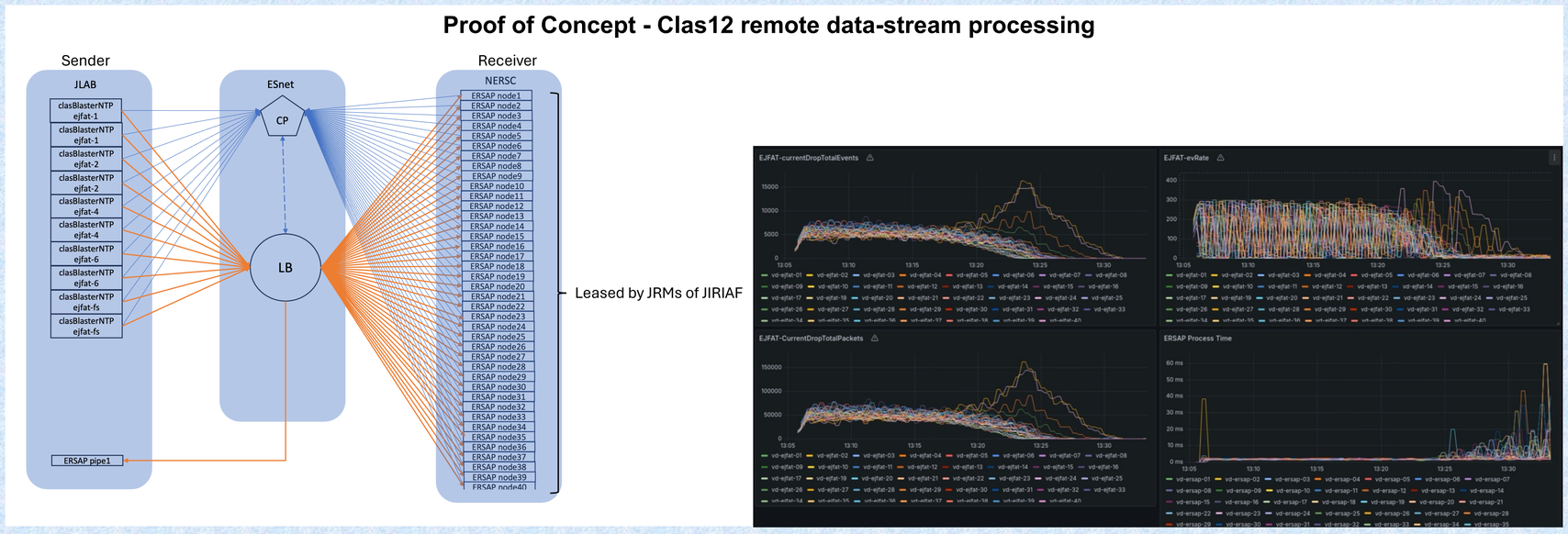




Architecture Overview

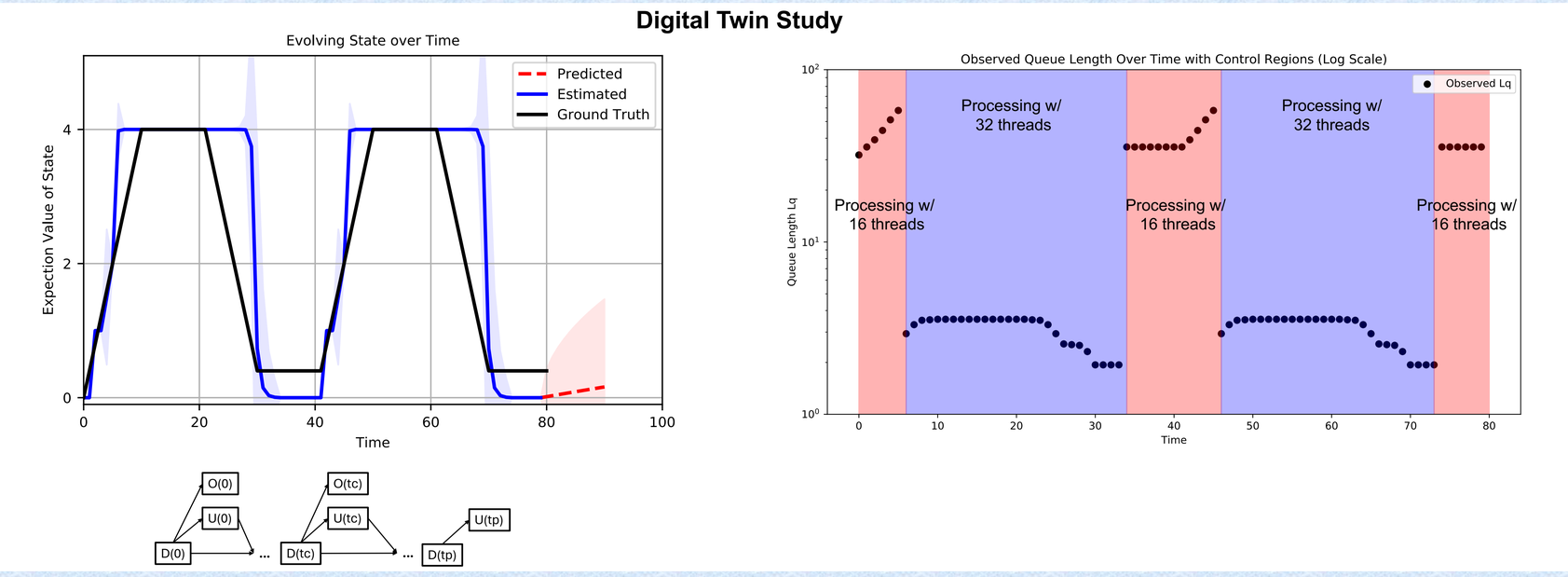
The JIRIAF architecture enables seamless integration and efficient resource management across diverse computing facilities.

- JFM (JIRIAF Facility Manager): Maintains a dynamic resource pool by periodically scraping data from each facility for up-to-date resource inventory.
- JCS (JIRIAF Central Service): Acts as the central command, initiating pilot jobs through the JRM.
- JRM (JIRIAF Resource Manager): Operates in user space to accommodate heterogeneous HPC setups, leasing resources reported by the JFM.
- JMS (JIRIAF Matching Service): Updates the resource database, aligning resources with user requests.
- JFE (JIRIAF Front End): Manages user requests and populates the workflow request table.



Proof of Concept - Clas12 remote data-stream processing

- Deployed 40-node reservation on NERSC's Perlmutter system to process CLAS12 event reconstruction.
- Each node in the Kubernetes cluster executed stream processing applications.
- The system demonstrated high data throughput and efficient resource allocation.
- The ESnet/JLAB FPGA accelerated transport system[3] was utilized for this proof of concept.



Digital Twin for Stream Processing Study

- Objective: Leverage a digital twin with a Dynamic Bayesian Network (DBN)[4] to optimize stream processing in a queue system through real-time state estimation and adaptive control.
- Real-time Assimilation: DBN-based digital twin effectively assimilates real-time data for state estimation and control.
- Accurate Estimation: States closely align with ground truth, especially during increasing queue lengths.
- Adaptive Control: Smooth transitions from 16 to 32 threads, reducing system pressure and optimizing performance.
- Decision-Making: Digital twin recommends actions based on real-time data, enabling automated control.
- Future Work: Apply model to more complex, realistic scenarios with additional variables.

Motivation

JIRIAF aims to streamline management of distributed infrastructures by:

- Efficiently migrating and scaling workloads across multiple computing sites.
- Utilizing opportunistic resources to improve efficiency.
- Maintaining system integrity while operating in user space.

JIRIAF provides a robust framework for resource management across high-performance computing systems.

Acknowledgement

This project is funded through the Thomas Jefferson National Accelerator Facility LDRD program. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under contract DE-AC05-06OR23177.

References

1. Virtual Kubelet Github Repository: https://github.com/virtual-kubelet/virtual-kubelet
2. JIRIAF Github Repository: https://github.com/JeffersonLab/jiriaf-virtual-kubelet-cmd
3. ESnet/JLAB FPGA Accelerated Transport: https://ieeexplore.ieee.org/document/10046405
4. Kapteyn, M.G., et al. "A probabilistic graphical model foundation for predictive digital twins." *Nat. Comput. Sci. 1, 337–347 (2021)*.