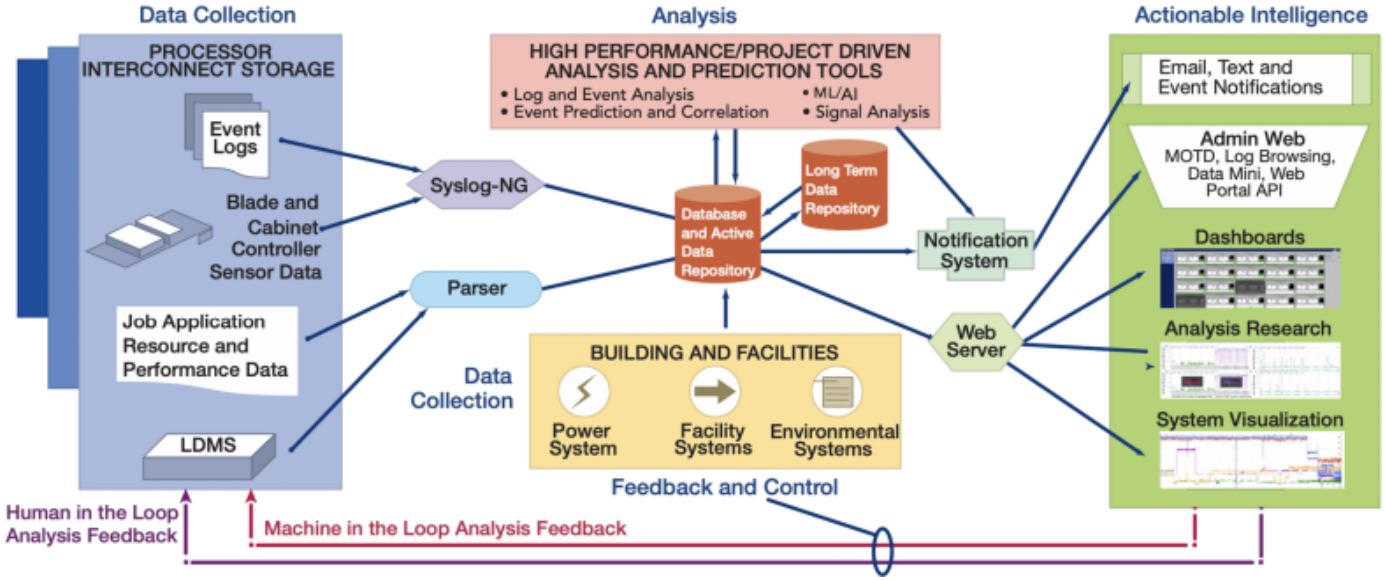
SAND2022-12094C

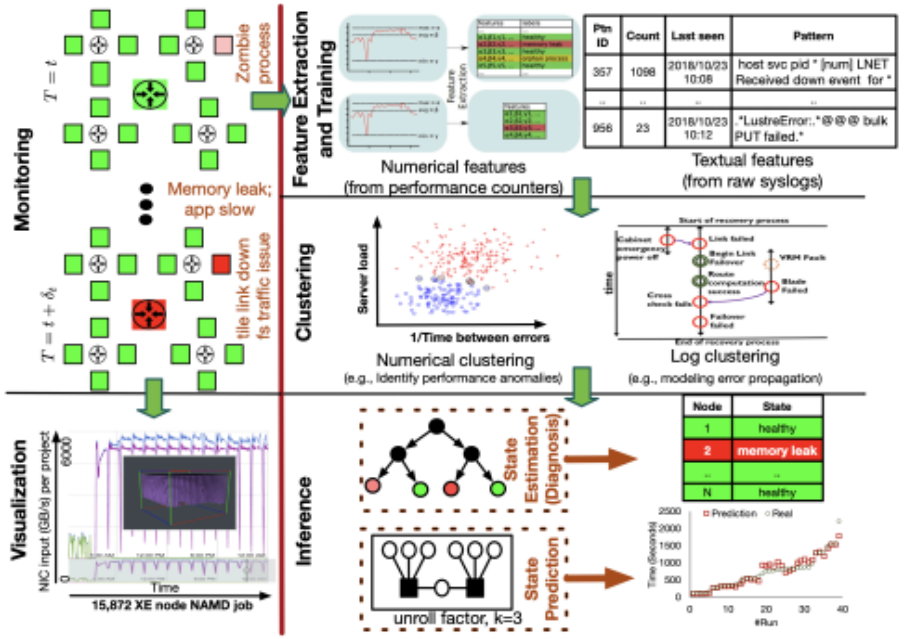
Moving Towards Autonomous HPC Facilities

Currently: HPC facility operation and rule-based orchestration is based on human experience, inference, and reaction times.

Goal (short term): AI provides data-driven **Decision Support** for **humans** to support better resource utilization and higher resource availability (equivalent to level 0 in auto (blind spot warning))

Goal (long term): AI used to **Orchestrate and Optimize resources** and **operations** subject to physical constraints such as power and priority with humans providing support in terms of repair and iteration on next generation capabilities (equivalent to level 5 in auto (fully autonomous))

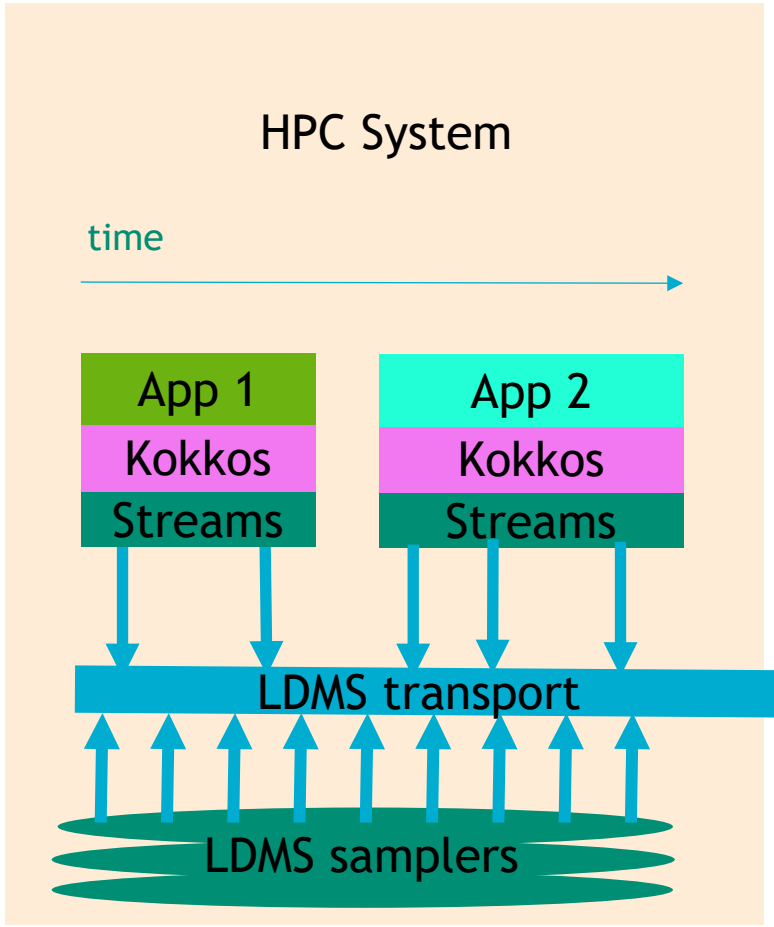


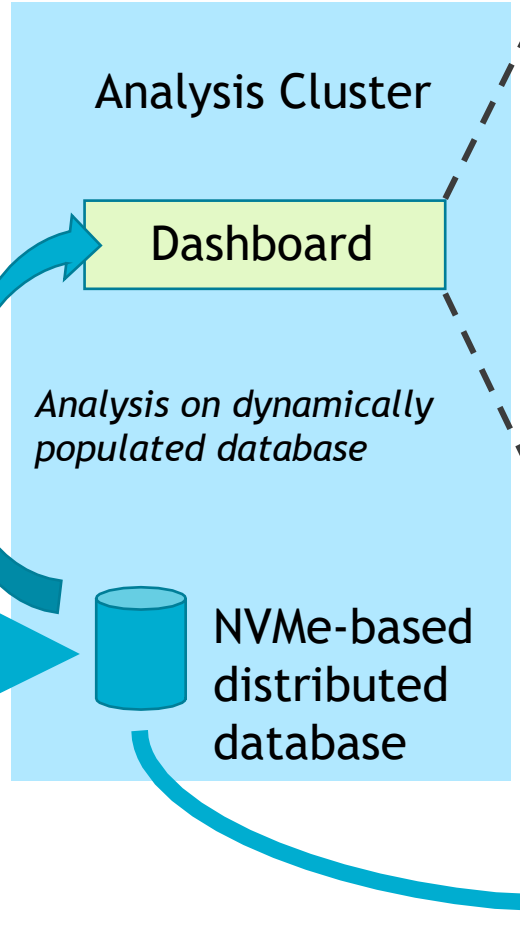


AppSysFusion: Characterization of Application and System State/Performance

Applications dynamically and irregularly inject data into the LDMS transport

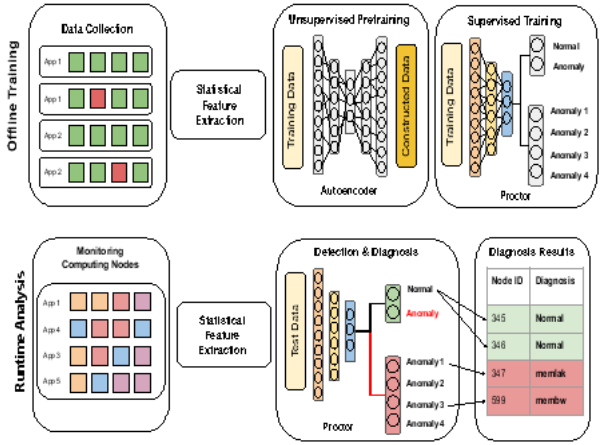
LDMS continuously and regularly collects and transports full system data





Application Progress

System Metric



ML-based Decision Support

Challenges:

- Application labeling (including input parameters)
- Hardware instrumentation
- Instrumentation meaning in context of architecture

Towards Autonomous HPC Facilities

Working with university partners to explore a variety of approaches to understanding application behavioral characteristics using ML

Exploring edge computing in conjunction with data collection for applying models to anomaly detection and locally tracking resource utilization

Re-imagining HPC resources as autonomous peer components that can negotiate among themselves and with applications to optimize utilization of resources and performance while honoring constraints such as power and priority