JOHNS HOPKINS
INSTITUTE *for*
ASSURED AUTONOMY

Assuring City Scale Infrastructure Systems

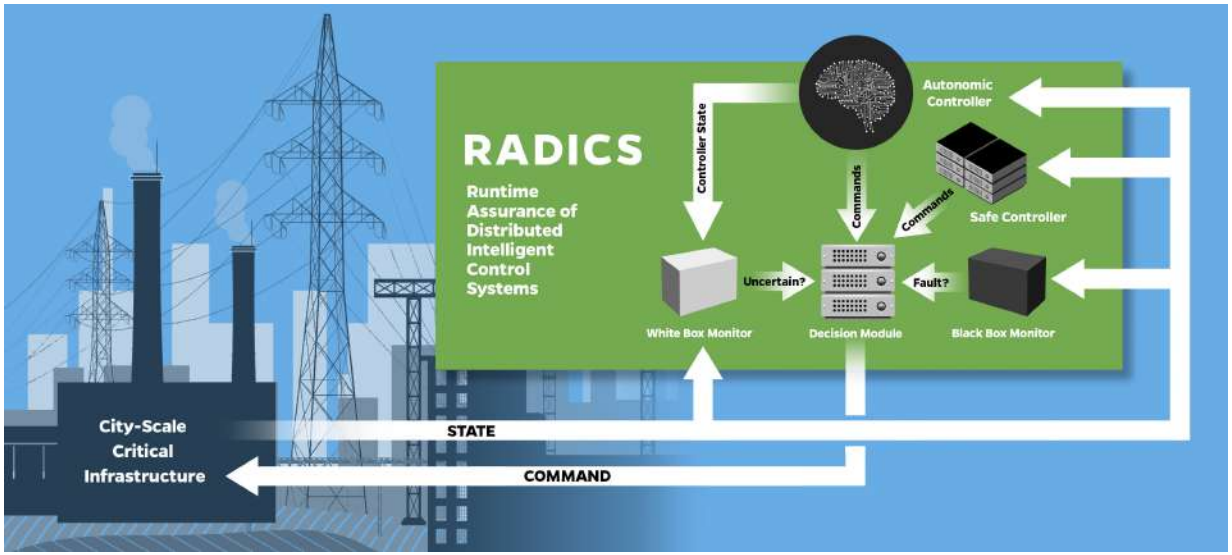
PIs: Yair Amir, WSE and Tamim Sookoor, APL

Date: 09/29/2020

www.dsn.jhu.edu/funding/assured-ai/

Project Summary

- AI systems are optimized for the average case. They cannot be used in critical systems that need to guarantee safety in worst case scenarios. The problem with AI systems is the long tail of edge cases that lead to failure situations. We want to gain the benefits of AI on the average case without incurring failures due to the long tail edge cases.
- Reinforcement learning (RL) algorithms are difficult to reason about and have non-intuitive behavior. RL algorithms break in nonintuitive ways due to phenomena such as reward hacking and specification gaming.
- Runtime Assurance of Distributed Intelligent Control Systems (RADICS) combines an invariant-based Black-Box Monitor with a White-Box Monitor that evaluates the confidence of the machine learning algorithm. The autonomous system is assured through these two monitors.



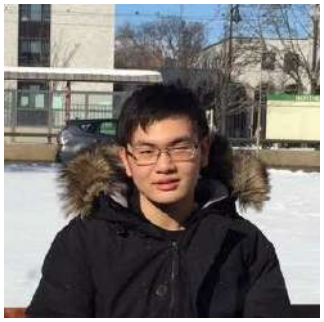
Project Team

Tamim Sookoor (PI)

Yair Amir (PI)

Brian Wheatman
Traffic Control Testbed (L)
Blackbox Monitor (L)

Sahiti Bommareddy
Smart Grid Testbed (L)



Jerry Chen
Traffic Control Testbed

Sebastian Zanlongo
Traffic Control Testbed (L)

Brad Potteiger
Traffic Control Testbed

Tim Krentz
Traffic Control Testbed
Smart Grid Testbed

Christina Selby
Whitebox Monitor (L)

Paul Wood
Whitebox Monitor

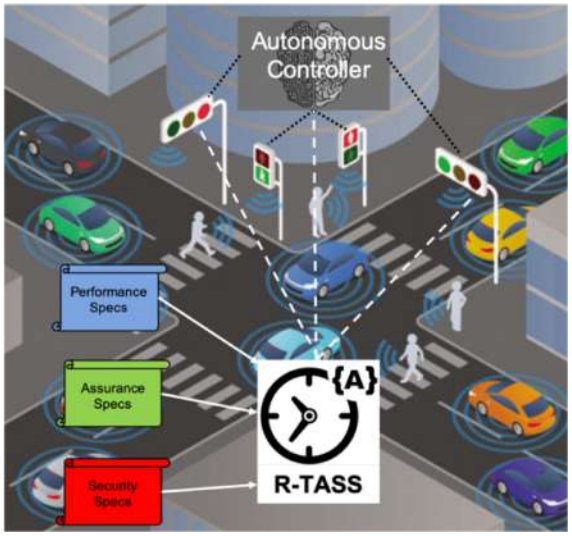
Nick Sarfaraz
Whitebox Monitor

Legend
WSE
APL
(L)ead

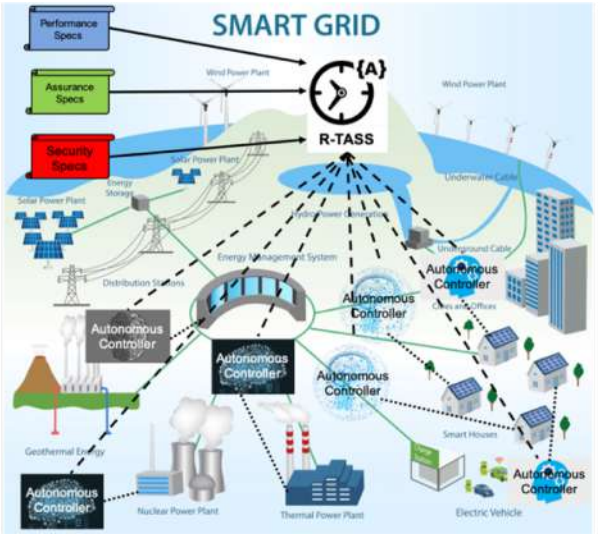
JOHNS HOPKINS
INSTITUTE for
ASSURED AUTONOMY

Flagship Use-case Projects

- Two ecosystem testbeds targeted at transportation and public safety domains



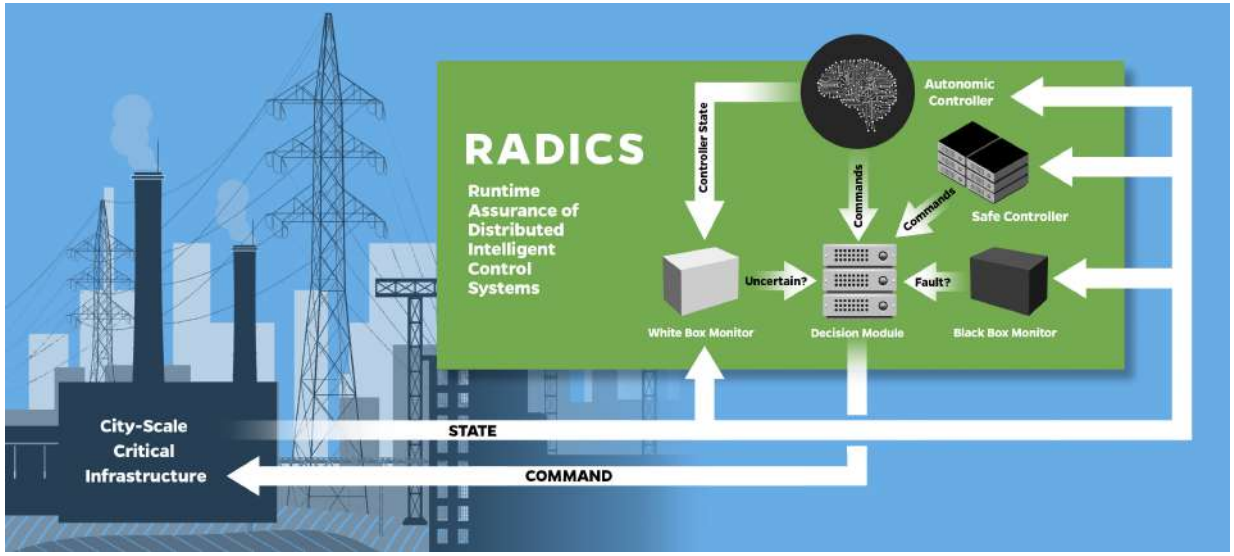
Intelligent Traffic Control



Smart Power Grid

Testbeds could be used for transportation and public safety domains

Current Status and Ongoing Research



- Intelligent Traffic Control
  - Full RADICS architecture realized and validated in simulation with a blackbox monitor
    - Based on the SUMO and Flow traffic simulation frameworks
    - Currently a monolithic 2x2 grid topology
    - Ongoing research toward a generalized model to support an arbitrary topology
    - Ongoing research on a whitebox monitor
- Smart Power Grid
  - Spire: Resilient SCADA for the Power Grid - test bed ready
    - Part of a DoE Byzantine Resilience effort to make the US power grid resilient to intrusions
  - AI-based Economic Dispatch component – initial investigation

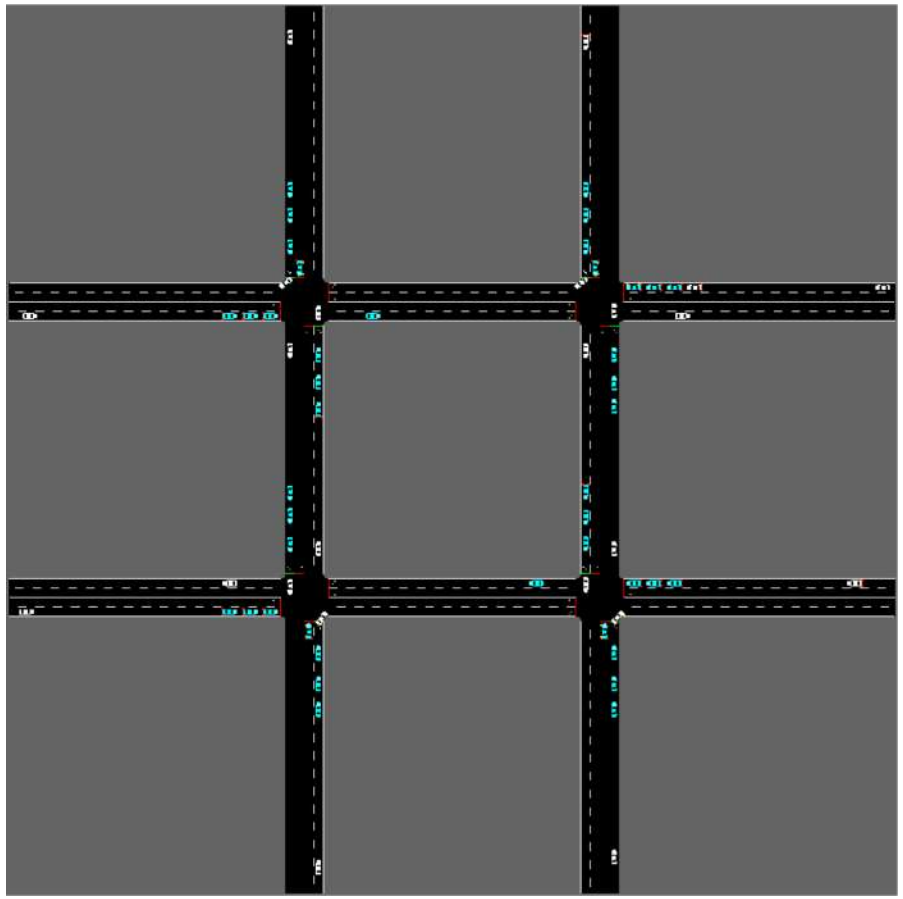
Intelligent Traffic Control System

- A semi-realistic AI model
  - With turning and random routes
  - Observes 3 vehicles per lane
  - Best model average speed: **6.48839** (m/s)
- Safe controller with a guarantee
  - Guarantee: vehicles that are going straight or turning right stop at most once at an intersection.
  - Average speed: **5.69047** (m/s)

Number of Steps / Iterations	Average Speed (m/s)
50K / 16	1.97545
10M / 3,333	3.01981
30M / 10,000	5.66471
50.3M / 16,766	6.00123
67.4M / 22,466	6.48839

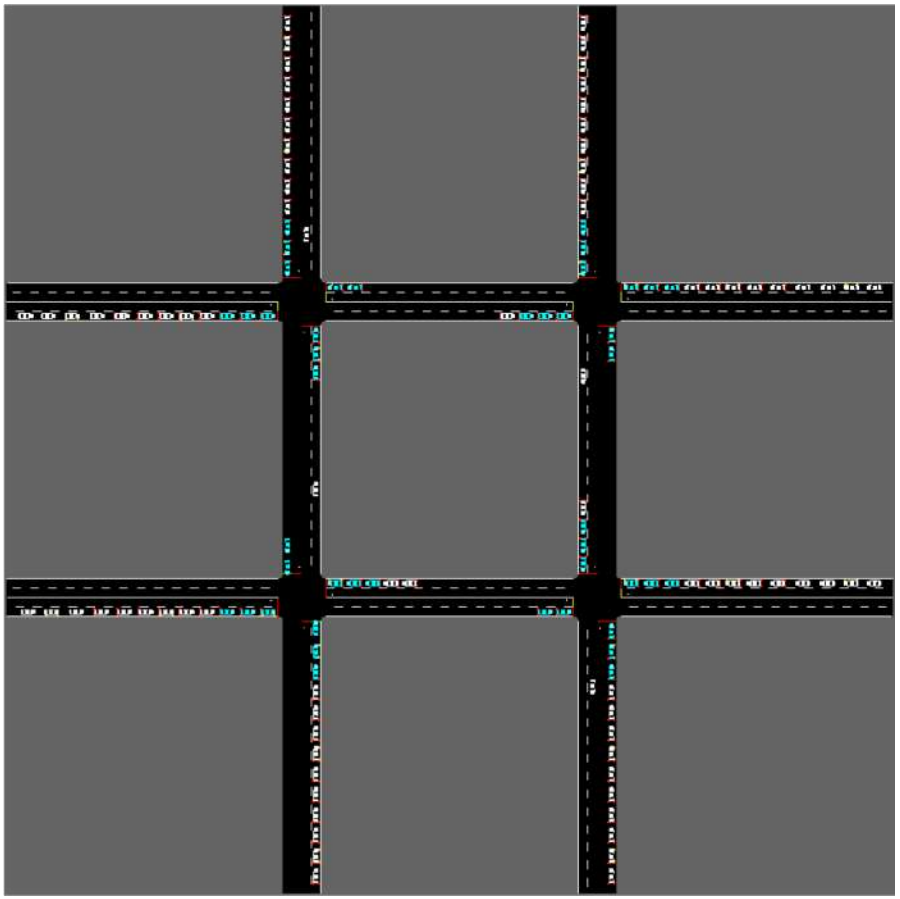
Intelligent Traffic Control System Simulation

Safe Controller



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/safe_controller_with_guarantee.mov

RL Model after 50K steps / 16 iterations

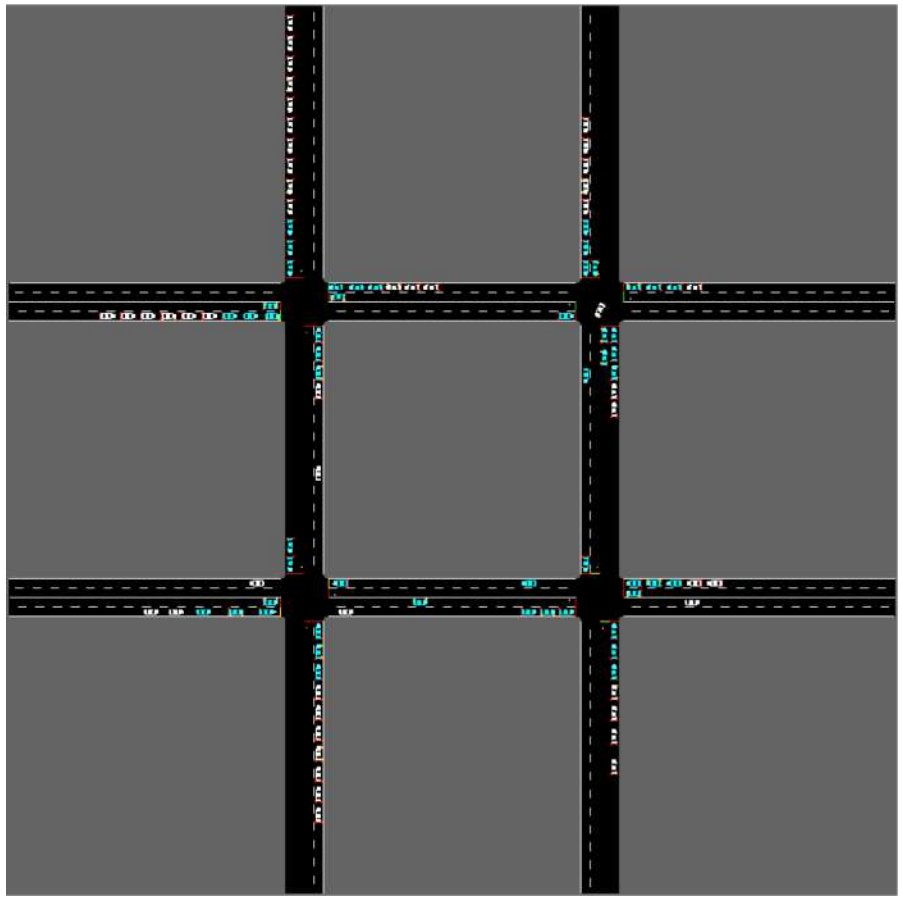


www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/rl_model_50k.mov

JOHNS HOPKINS
INSTITUTE for
ASSURED AUTONOMY

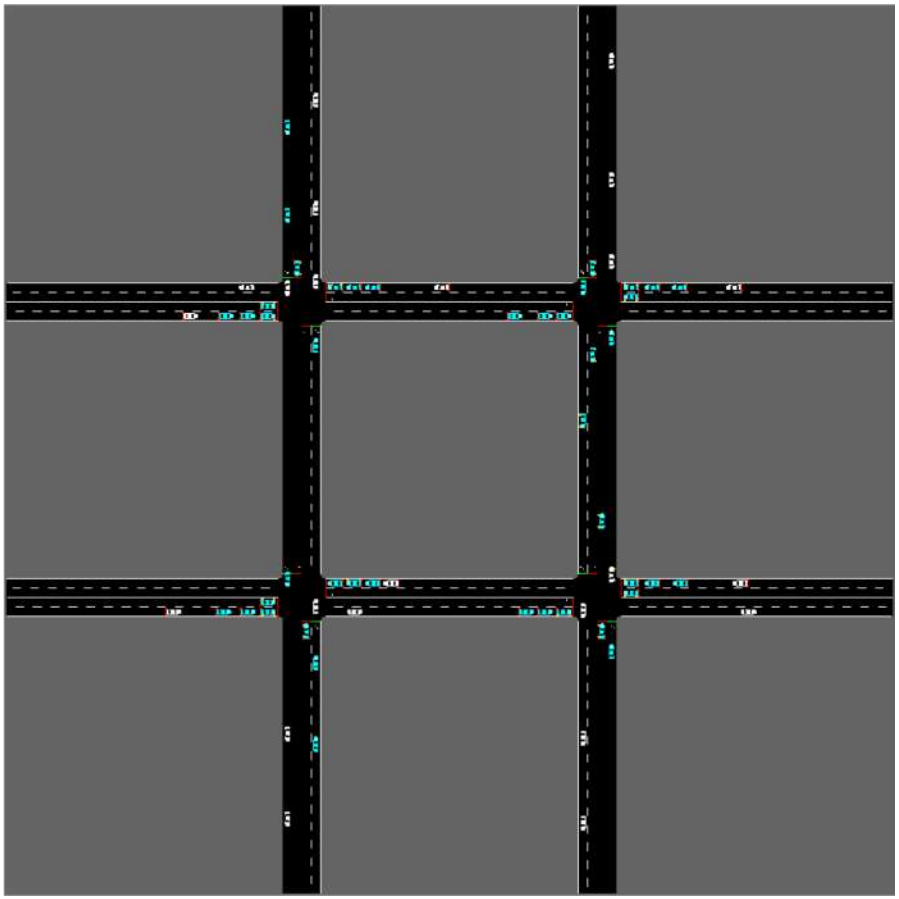
Intelligent Traffic Control System Simulation

RL Model after 10M steps / 3,333 iterations



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/rl_model_10m.mov

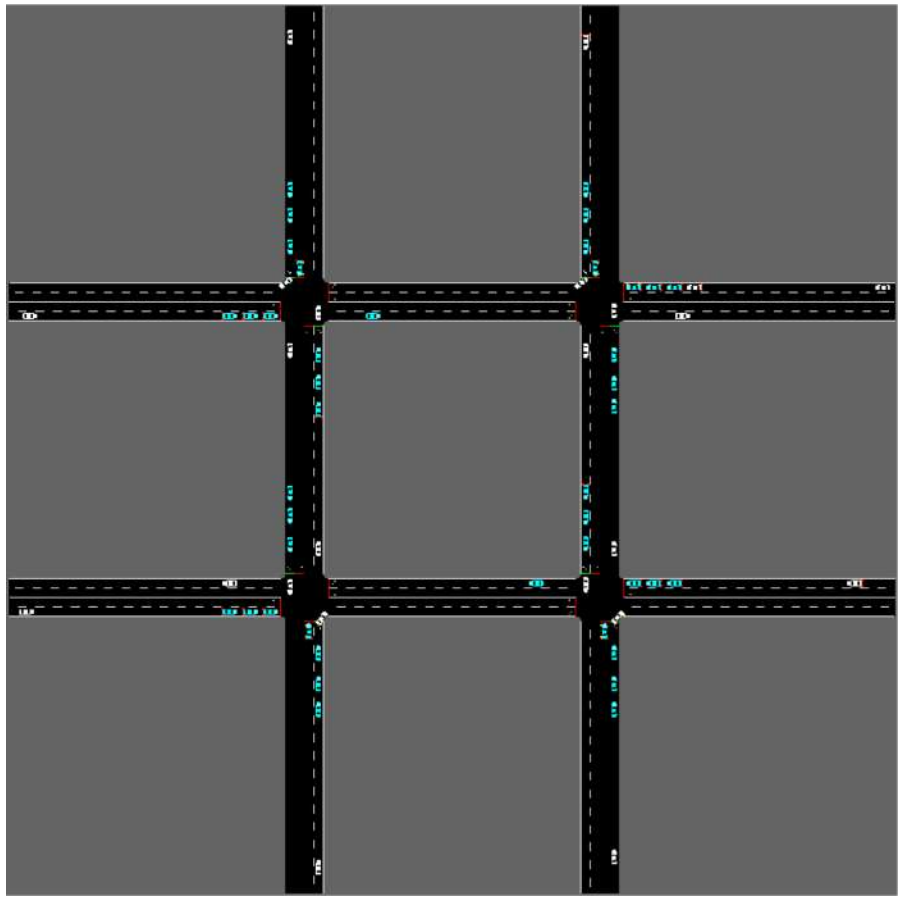
RL Model after 50.3M steps / 16,766 iterations



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/rl_model_50m300k.mov

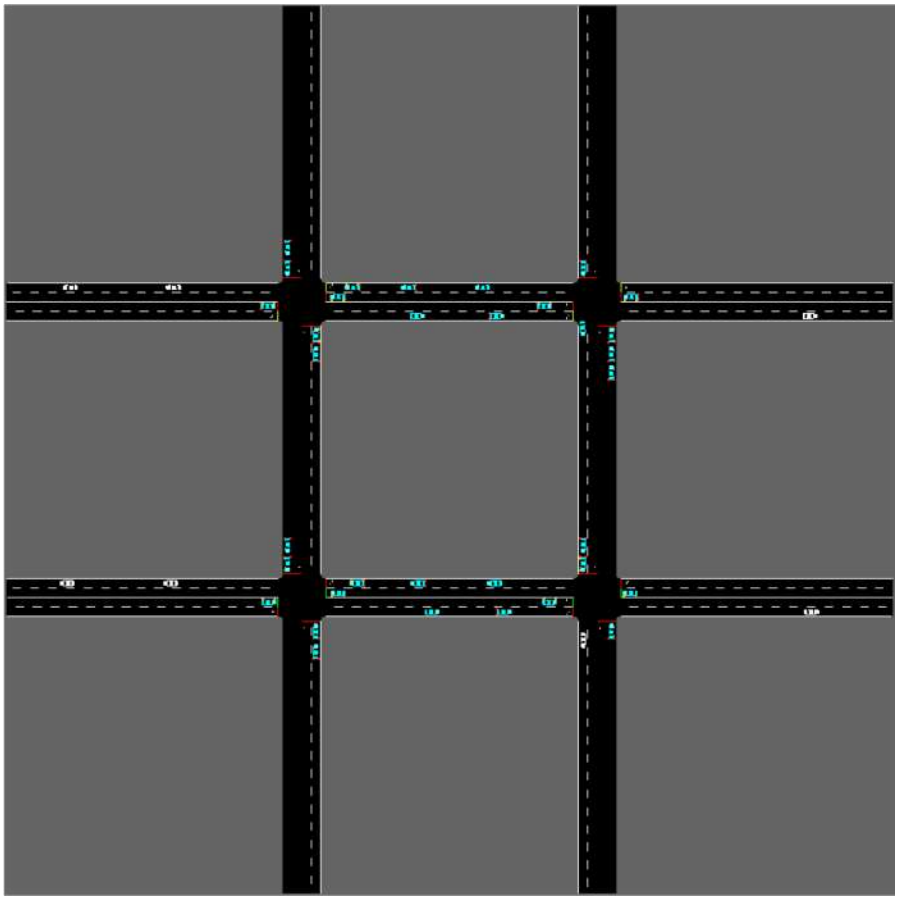
Intelligent Traffic Control System Simulation

Safe Controller



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/safe_controller_with_guarantee.mov

RL Model after 67.4M steps / 22,466 iterations



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/rl_model_67m400k.mov

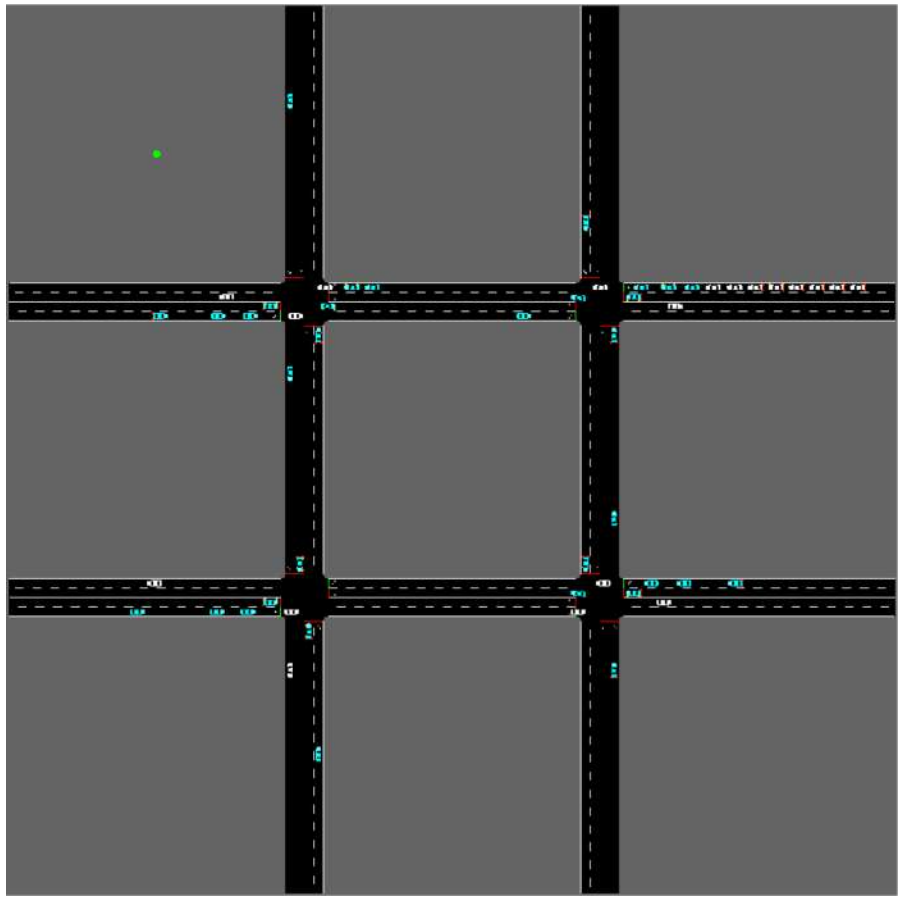
RADICS: Run-time Assured AI

- Switch from AI Controller to Safe Controller
  - switch when the average speed of past 50 seconds is below 4 (m/s)
- Switch from Safe Controller to AI controller
  - Strategy 1:
    After 60 seconds, switch to the AI controller when the average speed of the past 50 seconds is above 5 (m/s).
  - Strategy 2:
    After 60 seconds, run a test simulation with the AI controller for 1000 steps.
    The test simulation uses an inflow rate of the past 50 seconds.
    If the average speed of the test simulation is above 5 (m/s), switch to the AI controller.
- Results
  - Strategy 1: average speed (m/s) of all vehicles is **5.53297**
  - Strategy 2: average speed (m/s) of all vehicles is **5.84685**

RADICS: Assured AI in Action

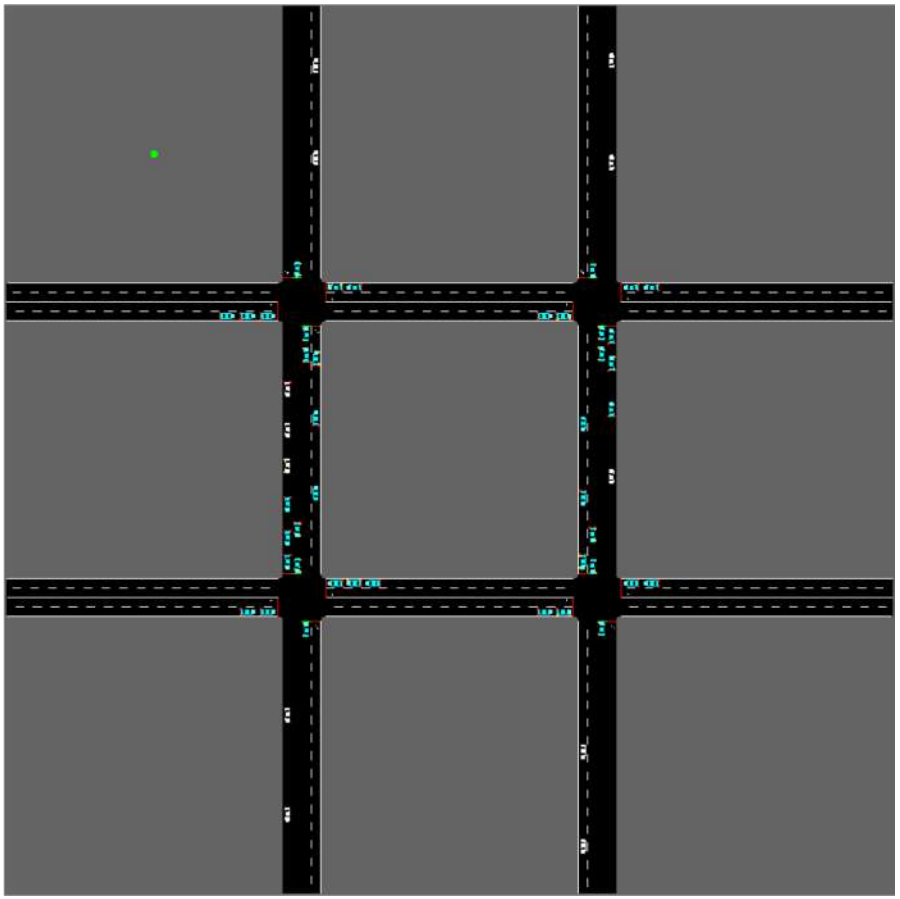
Green: AI controller; Blue: Safe Controller

Strategy 1



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/switch-mode-1-scenario6.mov

Strategy 2



www.dsn.jhu.edu/funding/assured-ai/2020_09_videos/switch-mode-2-scenario6.mov

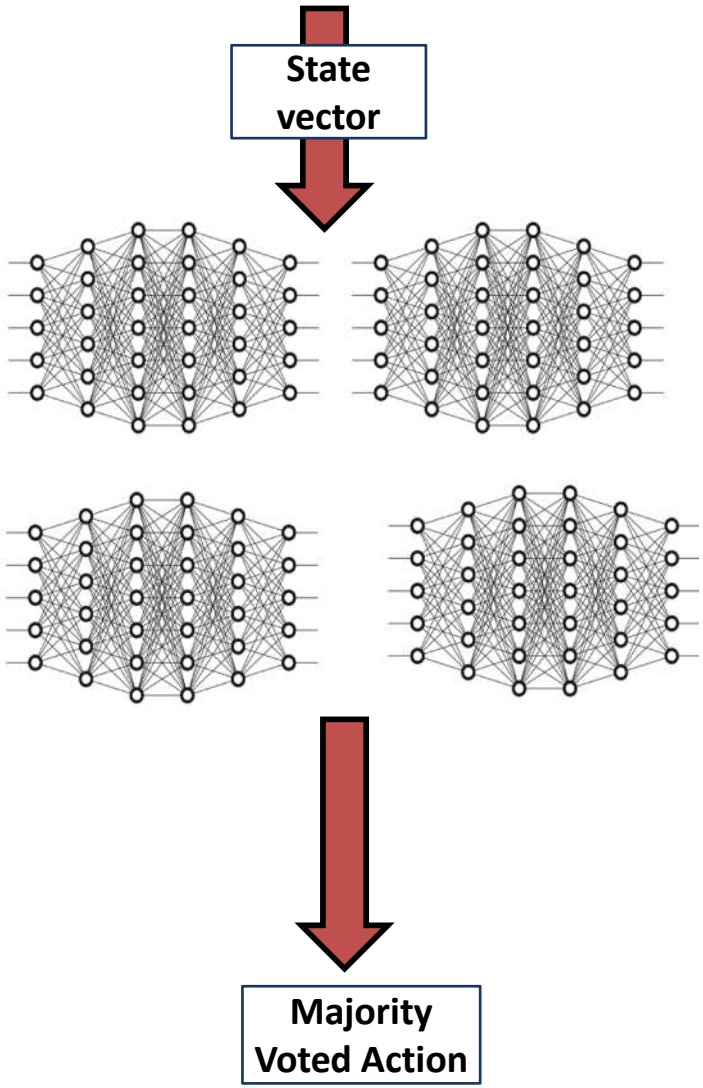
White Box Monitor

Goal: The current primary goal of the white-box monitor task is the development of a confidence metric for a reinforcement learning policy.

- A confidence value is associated with each action.
- The confidence value will be an input into the RADICS decision module.

Current Method: Ensemble RL Policy

- We train multiple reinforcement learning policies and create an ensemble policy that determines an action through majority vote.
- The variance in the reward for each action over each trained reinforcement learning policy will be used in the calculation of the confidence metric.



Reinforcement learning policy confidence is an open area of research

White Box Monitor

Confidence Metric Criteria

- If the model is given a state the model didn't observe during training, the confidence in the model's output should be low.
- If there are two or more actions that lead to similar rewards, then the confidence should be similar for any of those actions.
- Confidence should be positively correlated with associated performance metrics.

Evaluation

- Confidence values are per time step, while performance metrics are over a time interval.
- Statistical methodology for assessing the performance of the confidence metric are in development.

Smart Grid Testbed

Spire - www.dsn.jhu.edu/spire/

- Spire is an open-source intrusion-tolerant SCADA system for the power grid. Spire includes a SCADA Master and a PLC proxy designed from scratch to support intrusion tolerance, as well as a Human Machine Interaction (MHI) based on pvbrowser. Spire emulates power grid management under a number of distribution and generation scenarios
- We plan to extend Spire to include an Economic Dispatch module that predicts the demand for power and optimizes the cost of power generation to meet the predicted demand. We will design a Simplex architecture with the corresponding autonomous controller, black box and white box monitors, and decision module

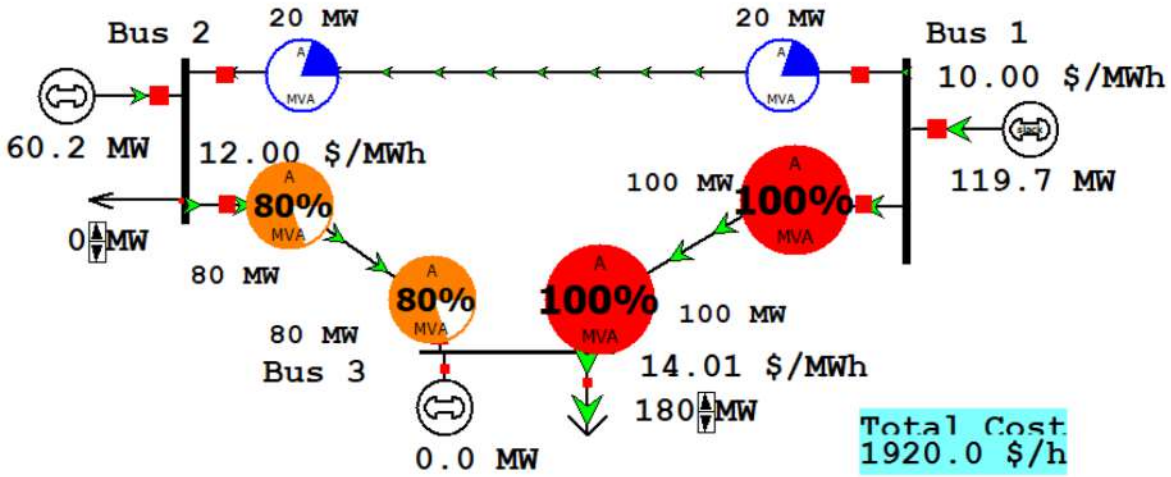
Smart Grid Testbed

Economic Dispatch Simulators

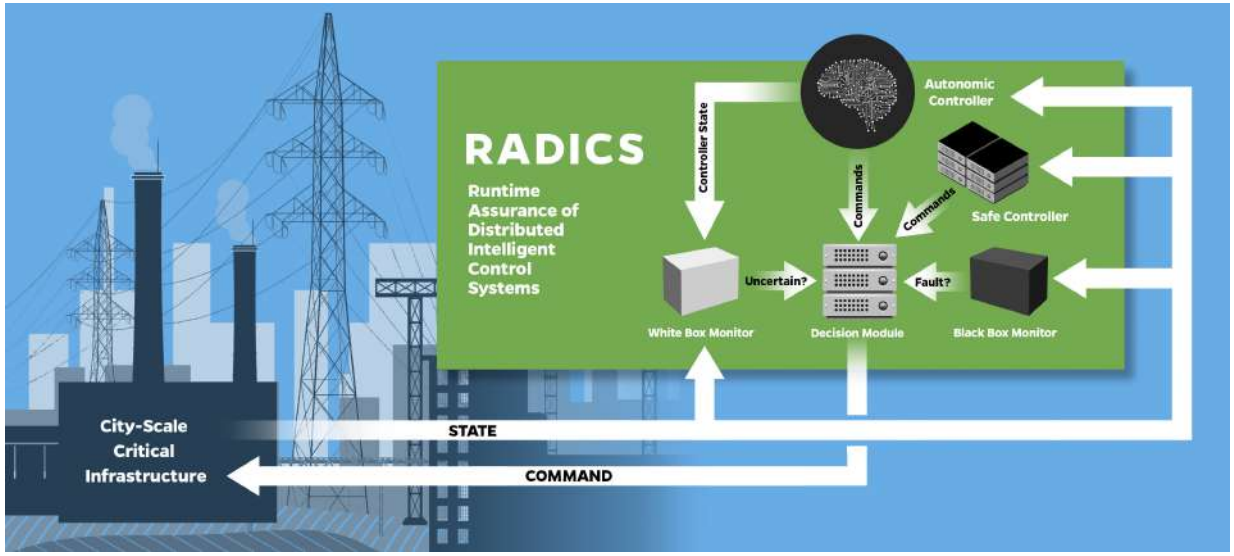
Currently exploring economic dispatch simulators to realistically model power generation systems

Simulators under consideration:

- Powerworld - https://www.powerworld.com/
- GridLAB-D - https://www.gridlabd.org/
- Etap - https://etap.com/



Current Status and Ongoing Research



- Intelligent Traffic Control
  - Full RADICS architecture realized and validated in simulation with a blackbox monitor
    - Based on the SUMO and Flow traffic simulation frameworks
    - Currently a monolithic 2x2 grid topology
    - Ongoing research toward a generalized model to support an arbitrary topology
    - Ongoing research on a whitebox monitor
- Smart Power Grid
  - Spire: Resilient SCADA for the Power Grid - test bed ready
    - Also part of a DoE Byzantine Resilience project
  - AI-based Economic Dispatch component – initial investigation