Haobo Guo
Xi'an Jiaotong University



excited to be presenting at the

Asian Modelica Conference

Modelica & FMI

Hangzhou China 2026

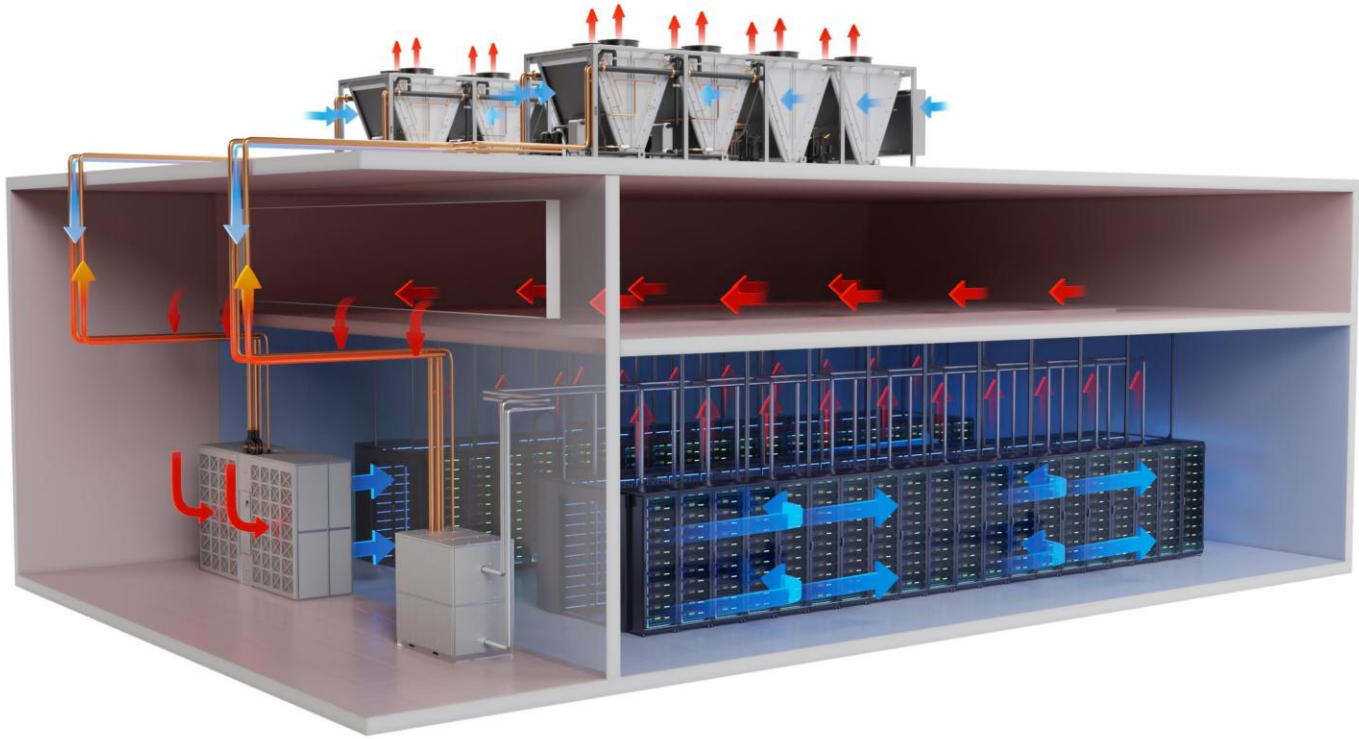
An FMI-Based Fast Thermal Prediction FMU for Real-Time Digital Twins of Data Center Server Rooms

Hangzhou, China
September 20 - 22

Data-center thermal-management architecture

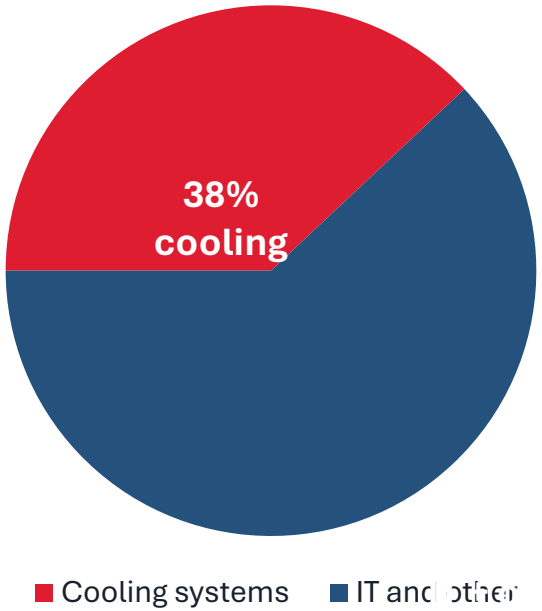
The server room is a coupled thermal system: IT racks generate heat, while CRAH units move and remove that heat through the air path.

Airflow path: CRAH units deliver cool air to the cold aisle; the air passes through the rack servers, picks up IT heat in the hot aisle, returns to the CRAH units, and transfers that heat to the building cooling loop, which rejects it outdoors.



Cooling-energy share and dynamic temperature response

The cited cooling literature reports that cooling systems can account for up to 38% of total data-center power use.



Why dynamic prediction matters

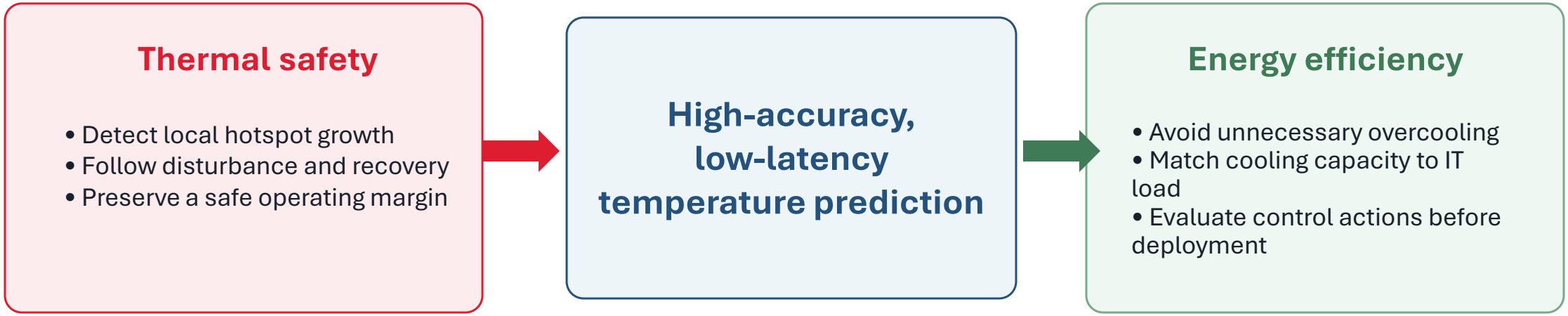


High-efficiency thermal management needs an accurate dynamic temperature state prediction, not only a static average.

The control-relevant quantity is the predicted temperature field at the next step.

Two main tasks of thermal management

A useful prediction component should support decisions on both sides of this trade-off.

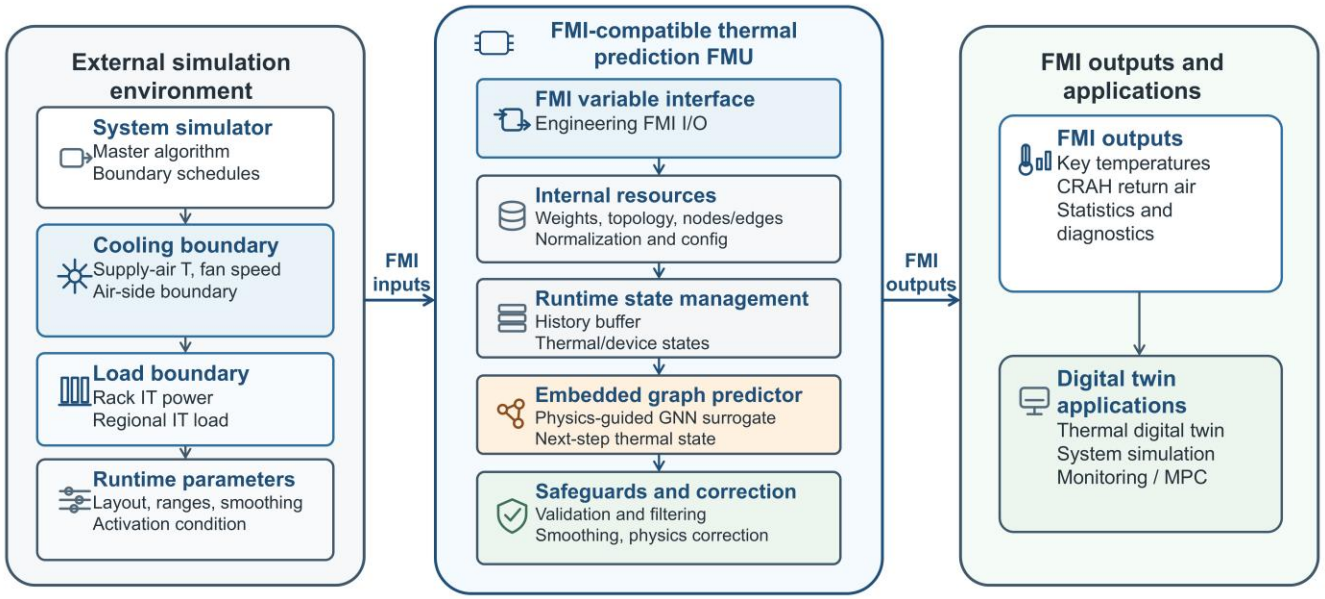


The predictor supplies future temperature trajectories to model-based control methods.

MPC: optimize cooling actions under thermal and actuator limits

Model-based RL: evaluate policy actions with model rollouts

FMU input, output and internal framework



The external system exchanges physical variables. The FMU maintains history, topology and model tensors internally.

FMI input • interaction variables

Supply-air temperature
CRAH fan speed / frequency
Rack or regional IT power

Monitoring points • state and validation

Room temperature sensors
CRAH supply / return-air points
Device status and replay data

FMI output • thermal states and diagnostics

Predicted key temperatures
CRAH return-air temperature
Validity / correction status

Key requirements for the temperature-prediction model

The prediction model is the main technical component inside the FMU. Its requirements come from continuous engineering use.

Why prediction is difficult in a real server room

High accuracy

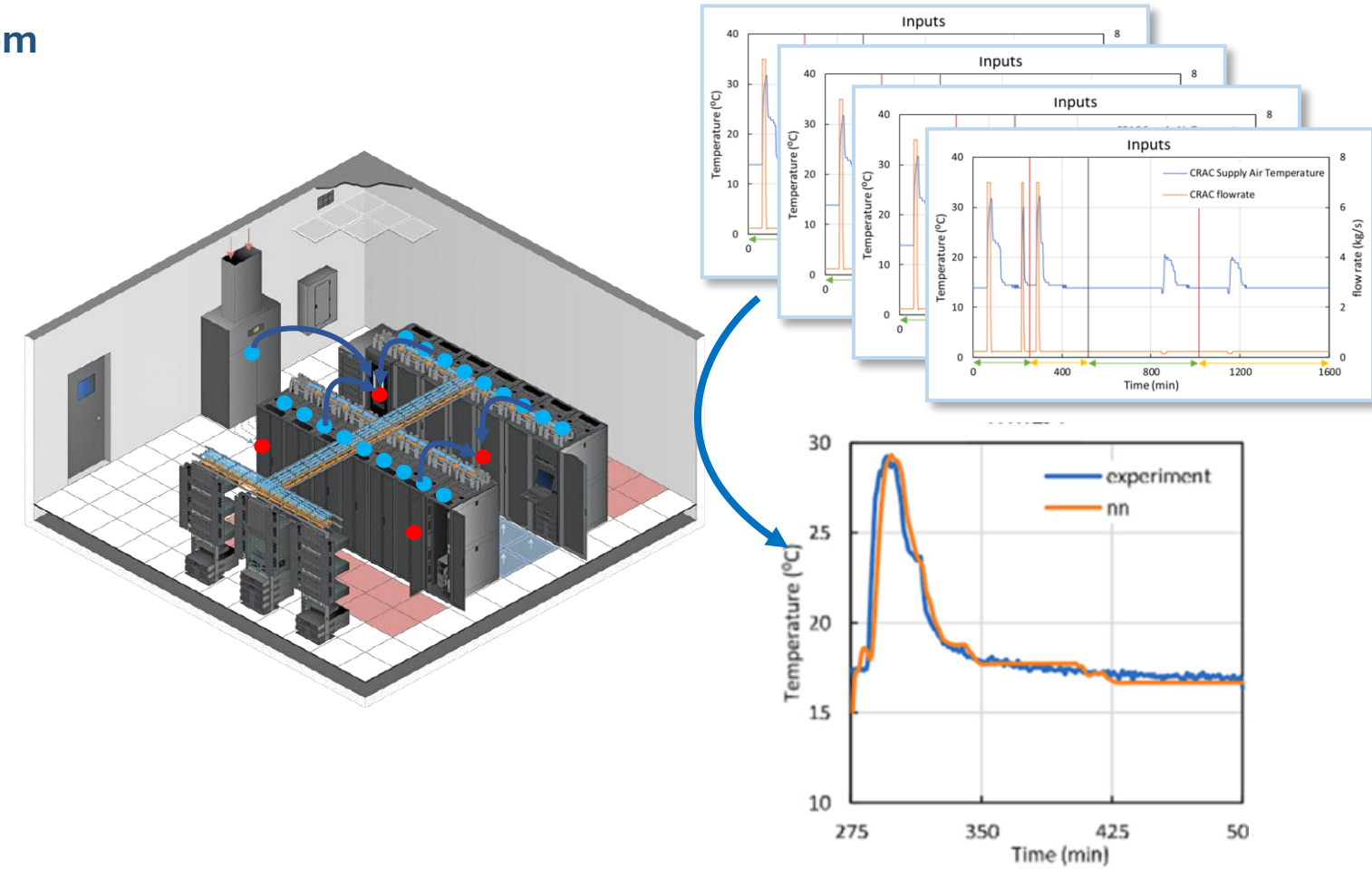
Local hotspots and return-air points respond differently after the same cooling action.

Generalization

The model must tolerate changing IT loads, cooling boundaries and room operating states.

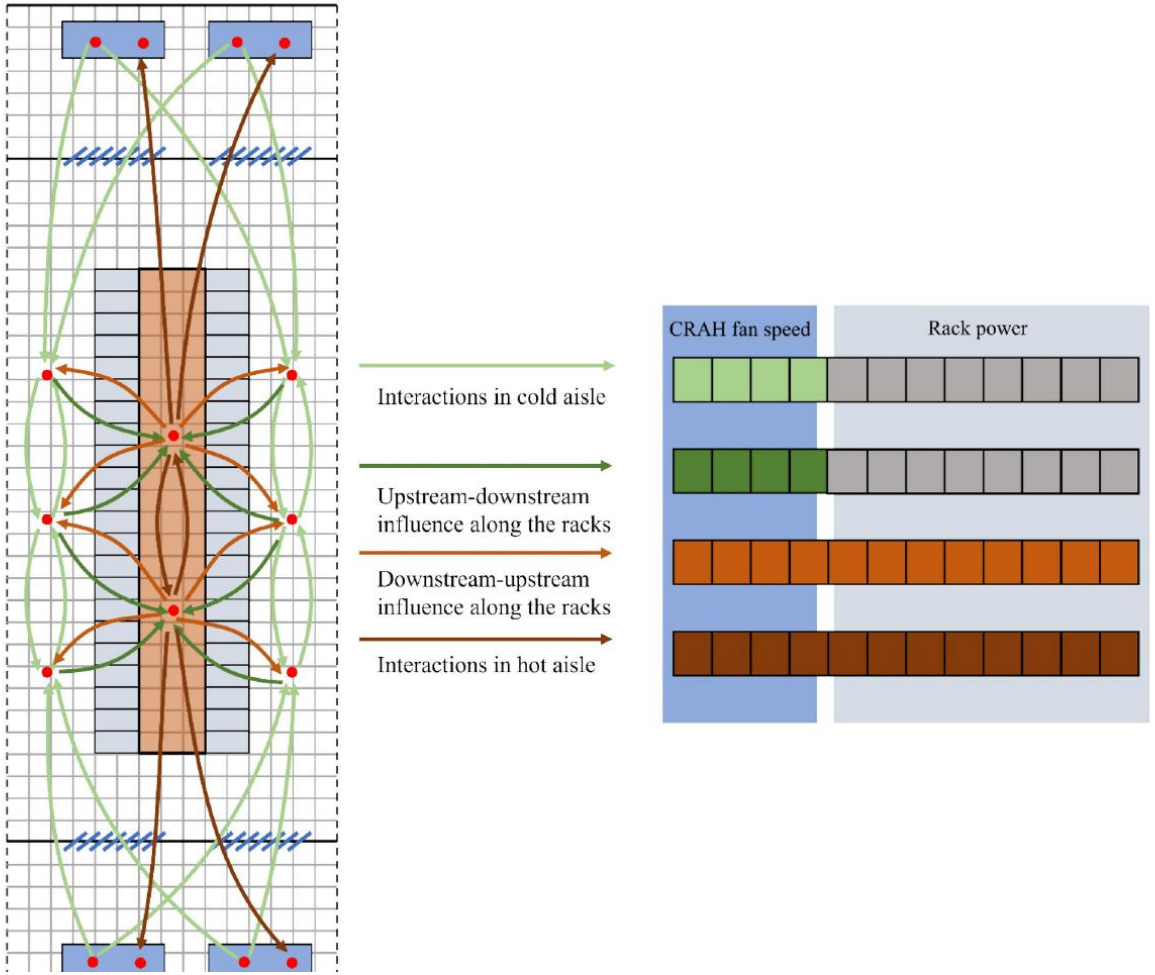
Real-time execution

Each step must update history, predict the next state and return outputs within the simulation interval.



Sparse sensors + changing boundaries + rolling state

Graph structure encodes where heat can travel



Node features

Temperature at a measured or modeled location, together with its spatial position and history.

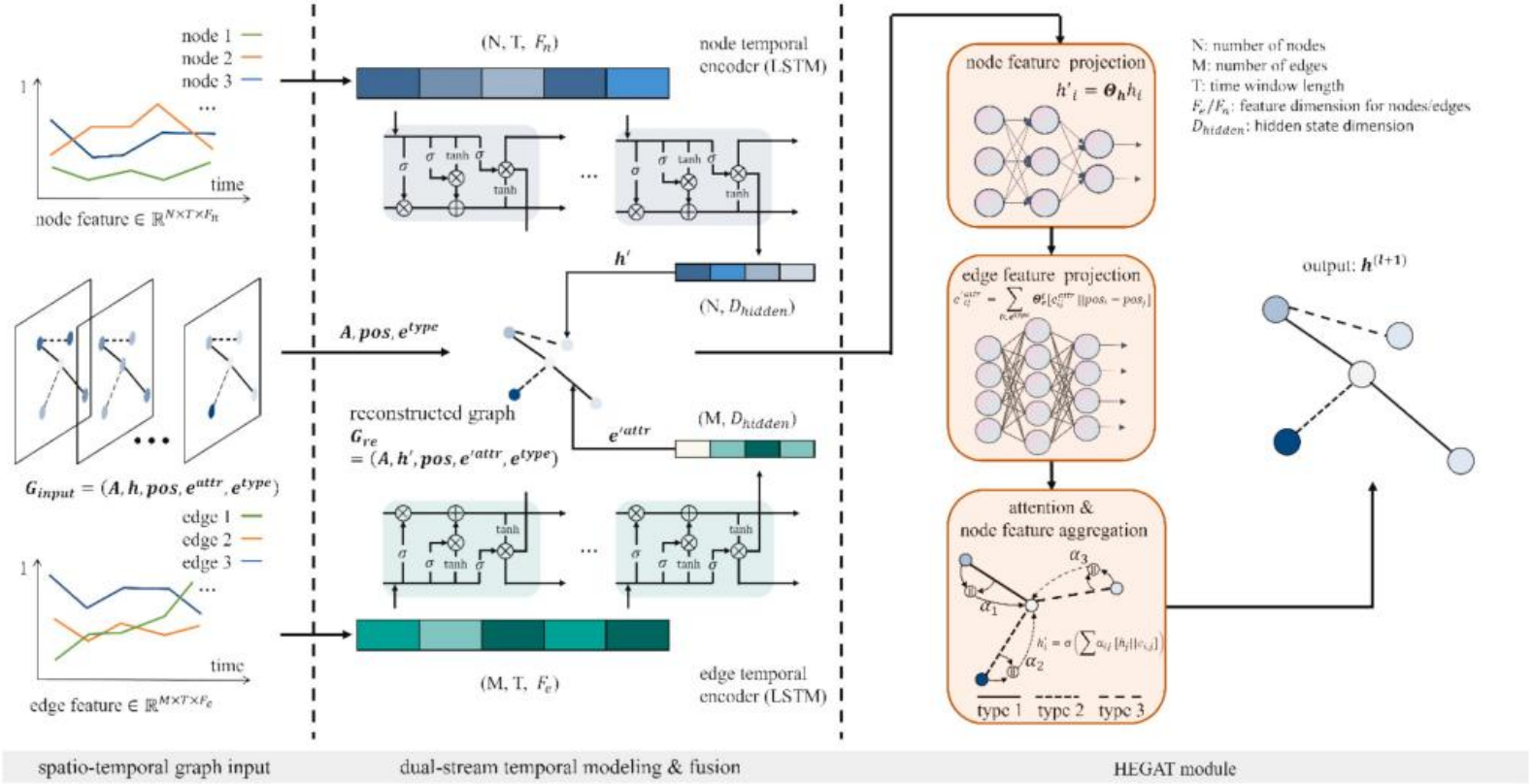
Edge attributes

Nearest CRAH fan speed and nearby rack power describe the operating condition along an influence path.

Edge types

Cold-aisle, along-rack and hot-aisle interactions distinguish different physical pathways.

ST-HEGAT combines temporal history with edge-aware attention



1 **Graph sequence**

Historical node temperatures and dynamic edge attributes form the input window.

2 **Temporal encoding**

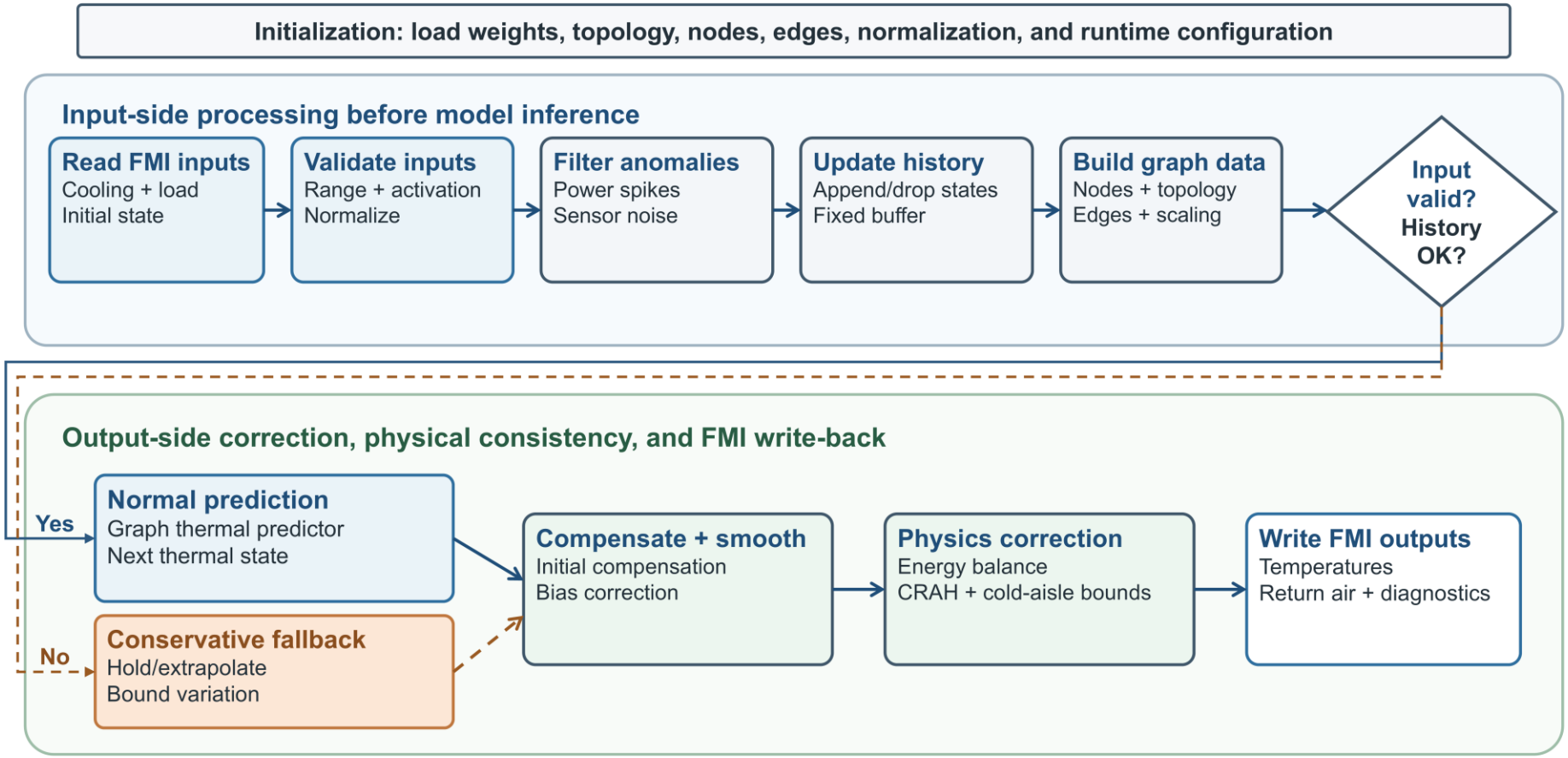
The recurrent state retains lag, propagation and recovery information.

3 **Edge-aware attention**

Heterogeneous message passing weights neighboring nodes using edge attributes and edge types.

Output: next-step room and return-air temperatures

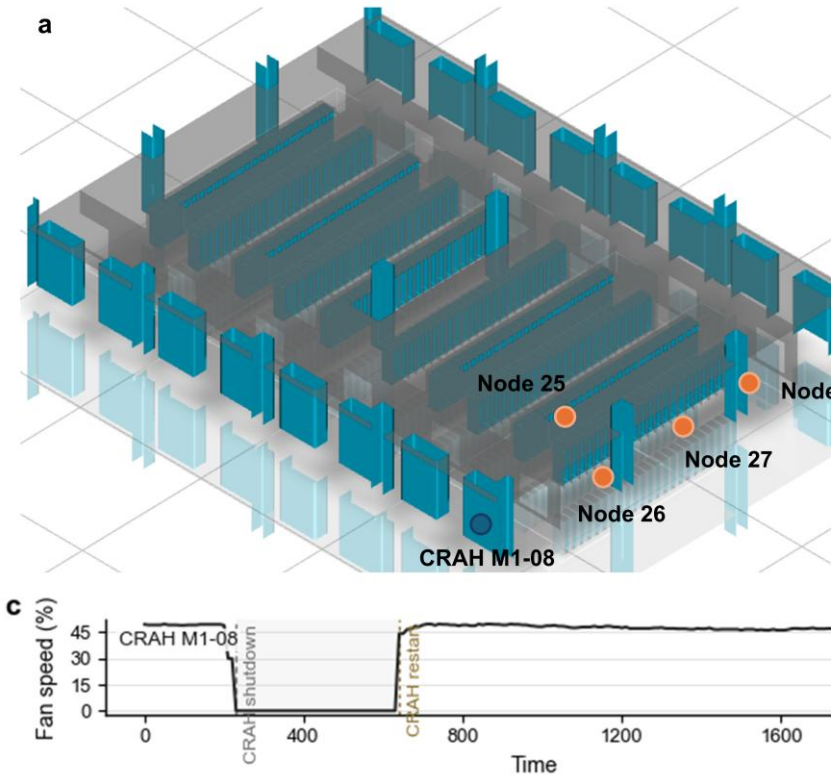
FMU step-wise execution framework



One do_step call

- Read cooling and load inputs
- Validate and normalize
- Update the history window
- Run graph prediction
- Correct and write FMI outputs

Application case: CRAH M1-08 shutdown and restart



Why this case?

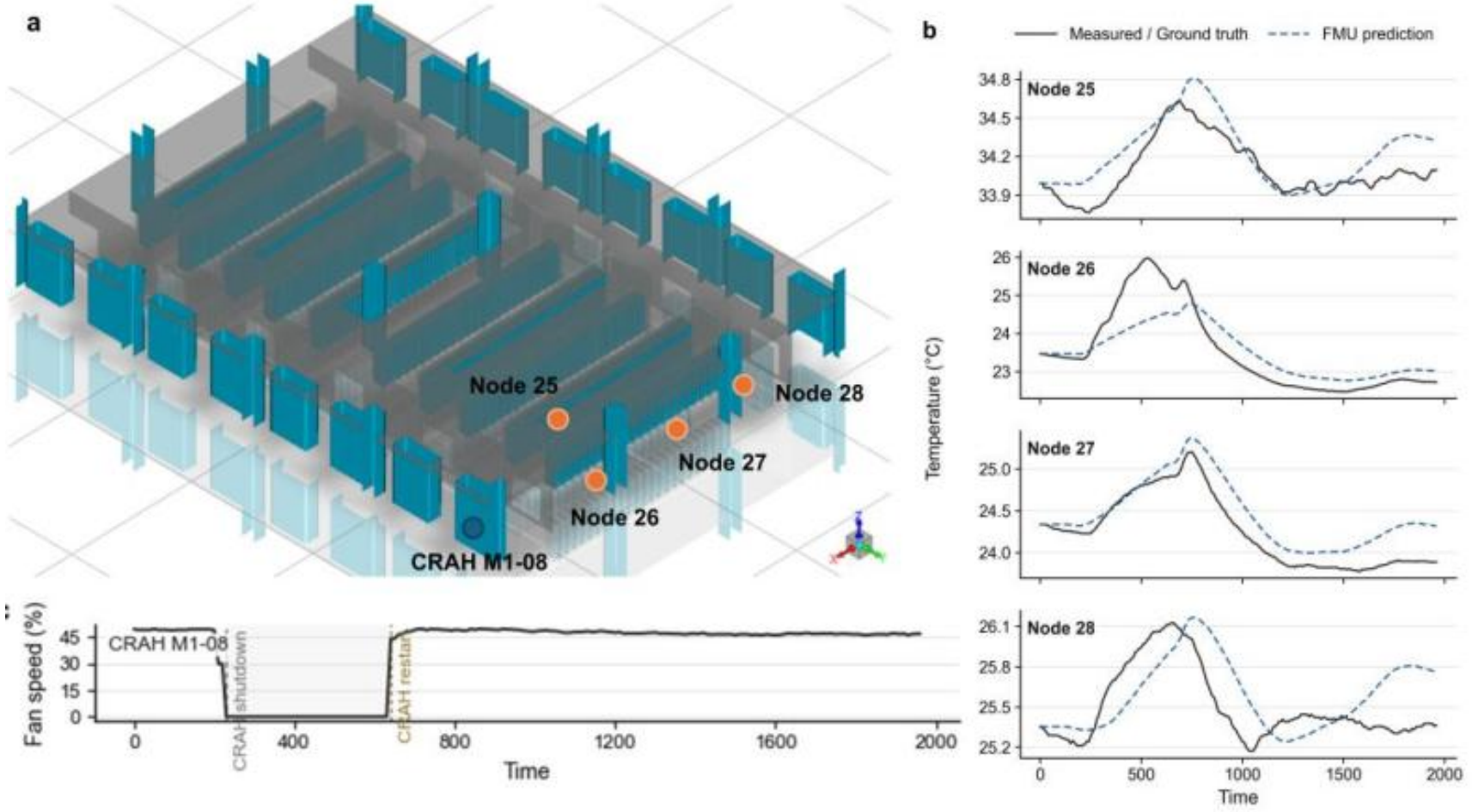
A CRAH shutdown creates a realistic cooling disturbance. It changes local cooling capacity, airflow distribution and return-air behavior.

Shutdown
230 s

Restart
640 s

197 replay steps | 10 s sampling | 43 evaluated nodes

The FMU tracks local temperature rise and recovery



What the curves show

- The prediction follows the dominant disturbance and recovery trend
- Different nodes respond with different amplitudes and timing
- Local peaks and recovery details still show deviations

Overall trend captured
Local transient details remain important

Accuracy and runtime support fast step-wise simulation

Accuracy

Overall RMSE

0.284 °C

Hotspot RMSE

0.192 °C

Overall MAE

0.197 °C

Return-air RMSE

0.373 °C

Runtime

Average step time

0.0656 s

Sampling interval

10 s

Real-time headroom

≈152×

10 s / 0.0656 s

Deployment capability and future integration

The current FMU already has the interface and runtime structure needed for engineering deployment.



Deployment and model-update path

Historical platform data
→ retraining and calibration
→ updated room-specific FMU resources

Operating changes to cover

Equipment aging
Seasonal drift
New rack installation or layout change

Next coupling direction

MPC or model-based RL
→ coordinated cooling actions
→ higher energy-use efficiency

Thank you

Questions and discussion welcome

An FMI-Based Fast Thermal Prediction FMU for Real-Time Digital Twins of Data Center Server Rooms

Haobo Guo · Xi’an Jiaotong University · AMFC 2026