Storage Workload Isolation via Tier Warming: How Models Can Help

JI XUE[1], FENG YAN[1], ALMA RISKA[2], EVGENIA SMIRNI[1]

[1] Computer Science Department
College of William and Mary

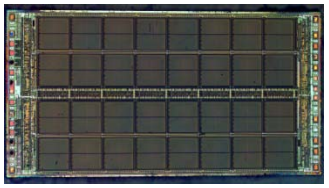
[2] EMC Corporation



EMC²
where information lives

Tiered Storage

- **Flat form**
  - Problem
    - Performance
    - Cost
- **Multi-tiered Storage System**
  - DRAM
    - $1/GB, <10 us, hundreds of GBs
  - SSD
    - $0.45/GB, 0.1 ms, tens of TBs
  - HDD
    - $0.05-0.10 /GB, 12 ms, multi PBs

DRAM 

SSD 

HDD 

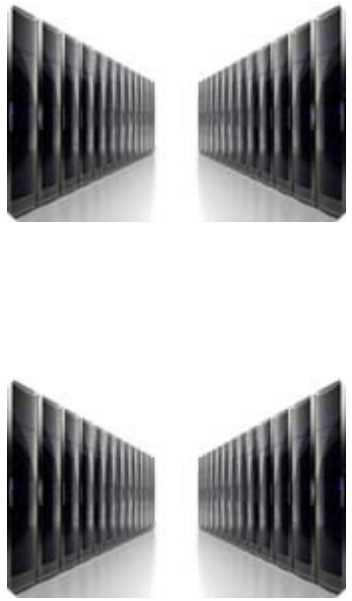
Traffic In Storage Systems

User Traffic
e.g. user upload, access …

System Traffic
e.g. replication, de-duplication…



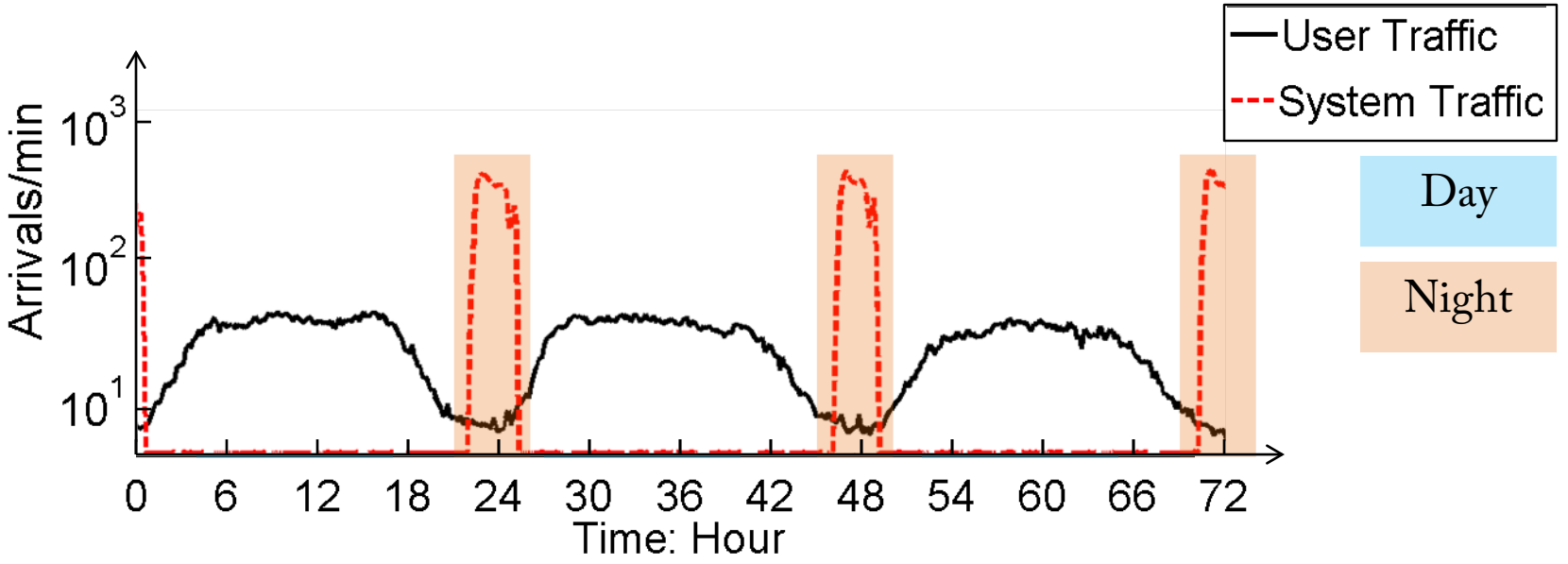




Traffic In Storage Systems

- **User Performance *V.S.* System Operation**
  - Overall Intensity
  - Working Set
    - **Right Place, Right Time**





Problem Definition

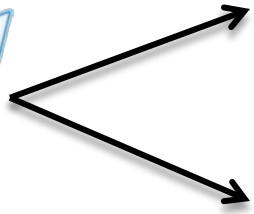
User Traffic
e.g. user upload, access …

V.S.

System Traffic
e.g. replication, de-duplication…

How to Schedule Properly?

Goal



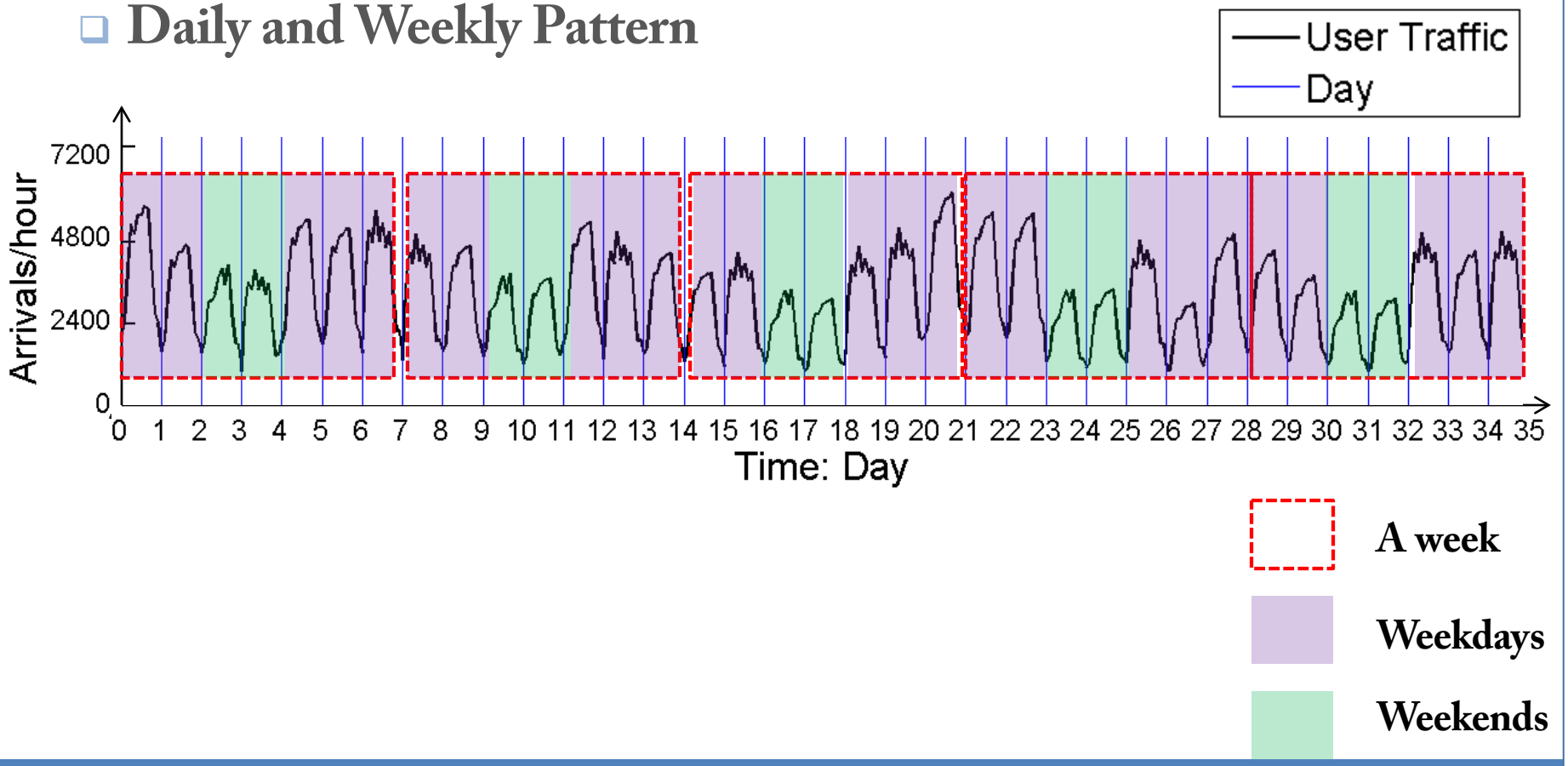
User Traffic Performance

System Traffic Performance

Motivation

Trace Overview

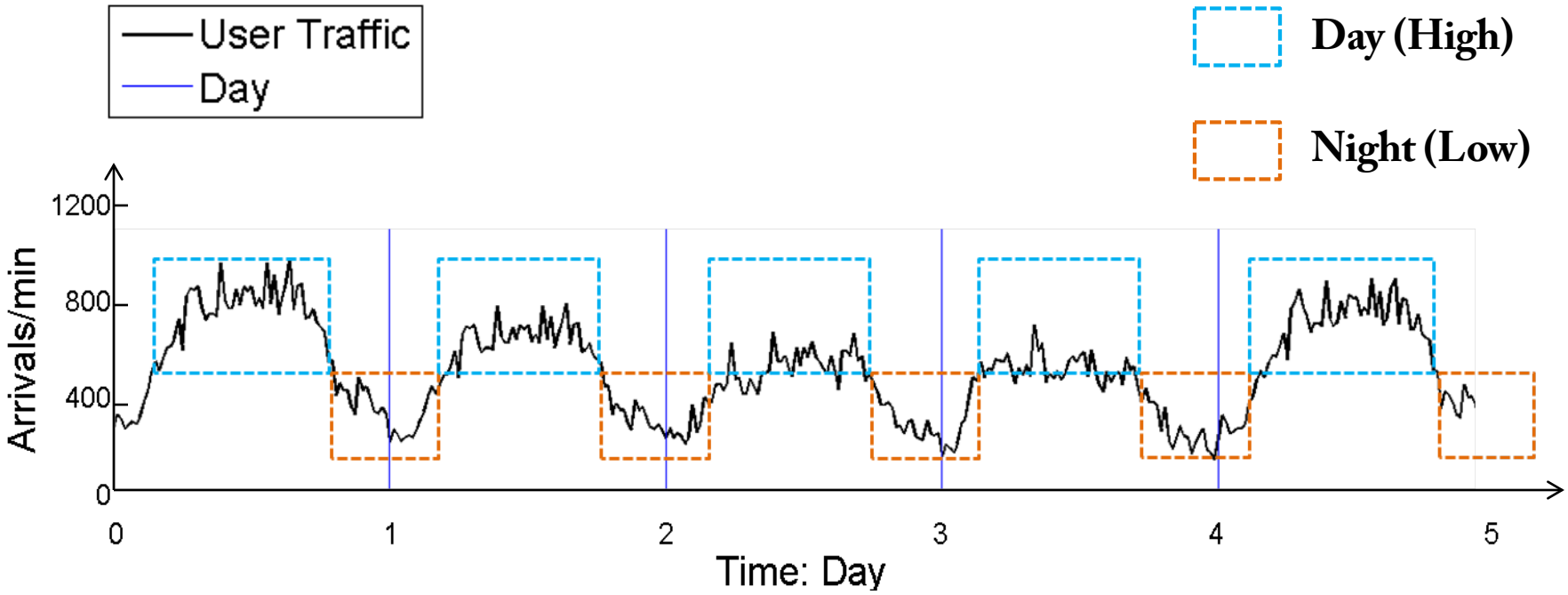
Daily and Weekly Pattern



Motivation

Trace Overview

Day-and-Night Pattern



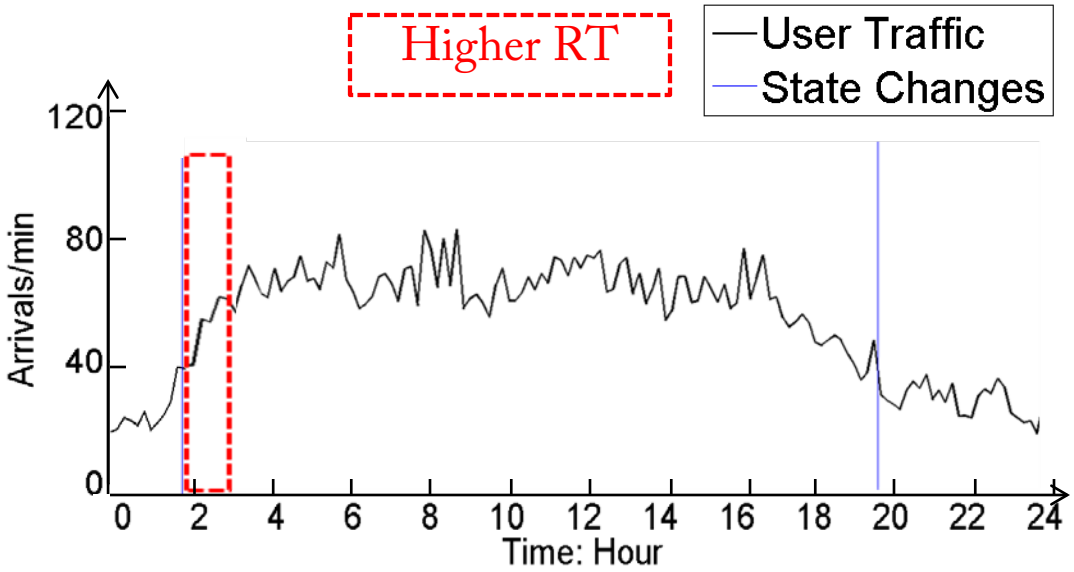
Motivation

- **Opportunity**
  - Predict *when* drastic changes happen
  - *Proactively* prepare the system for heavy user workload
    - E.g., warm up the fast tier with user data

Motivation

Three Scenarios

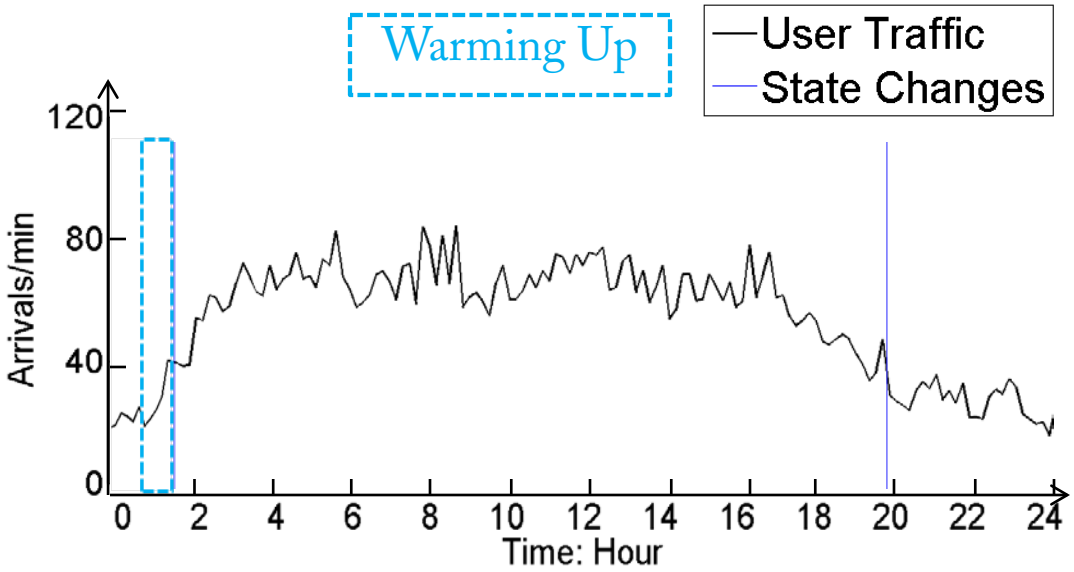
- User-Only
  - No system work
- **Reactive**
  - ***Online*** **detection**
  - **Brings user data when accessed *after* arrival of user high intensity periods**
- **Proactive w/ Future Knowledge**
  - ***A priori*** **knowledge of time when user high utilization starts**
  - **Warms up *before* user high intensity periods**



Motivation

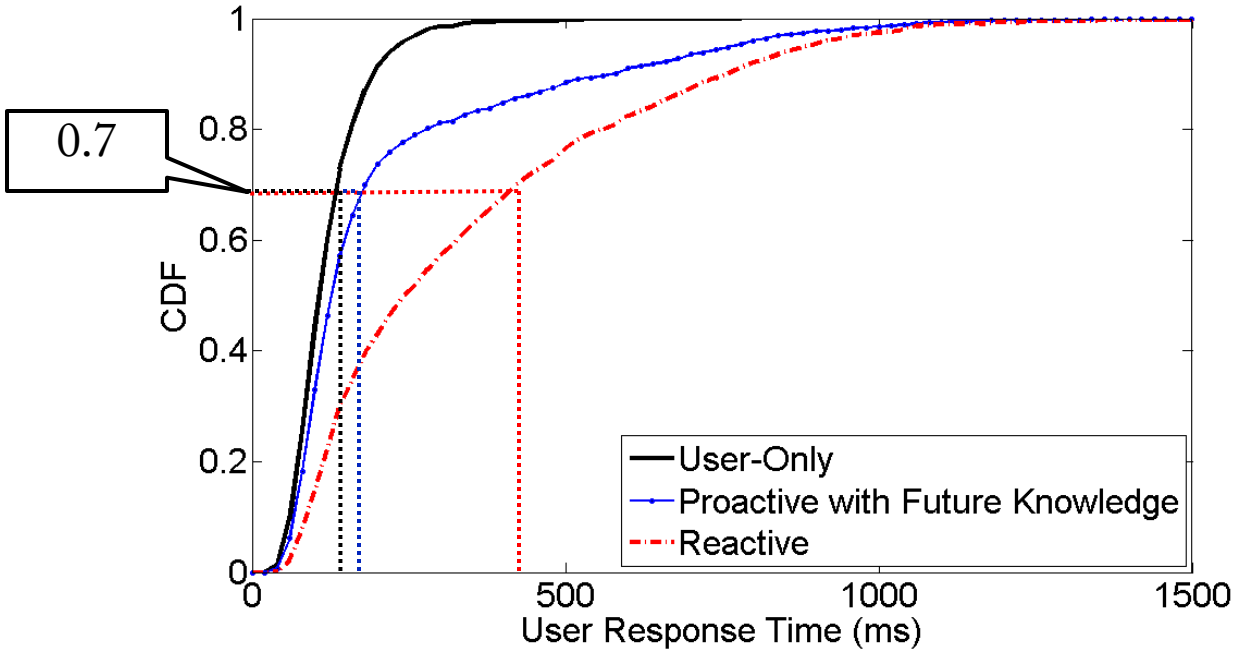
Three Scenarios

- User-Only
  - No system work
- Reactive
  - *Online* detection
  - Brings user data when accessed *after* arrival of user high intensity periods
- Proactive w/ Future Knowledge
  - *A priori* knowledge of time when user high utilization starts
  - Warms up *before* user high intensity periods



Motivation

- **Comparison**
  - CDF of user RT
    - User Only: base line
    - Reactive *versus* Proactive w/ Future Knowledge

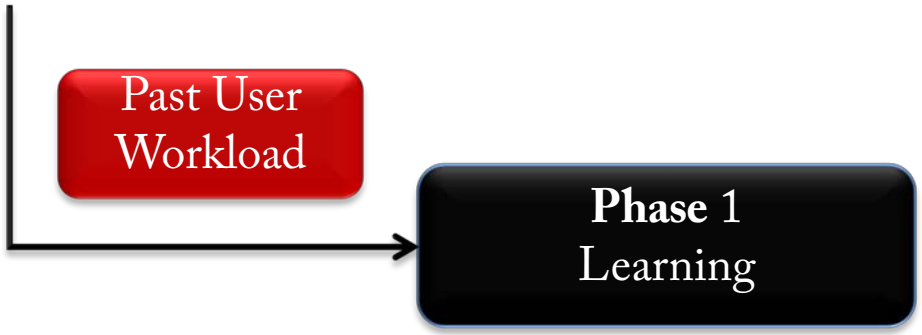


Challenge

Can our prediction be as (almost as) good as the a priori future knowledge?

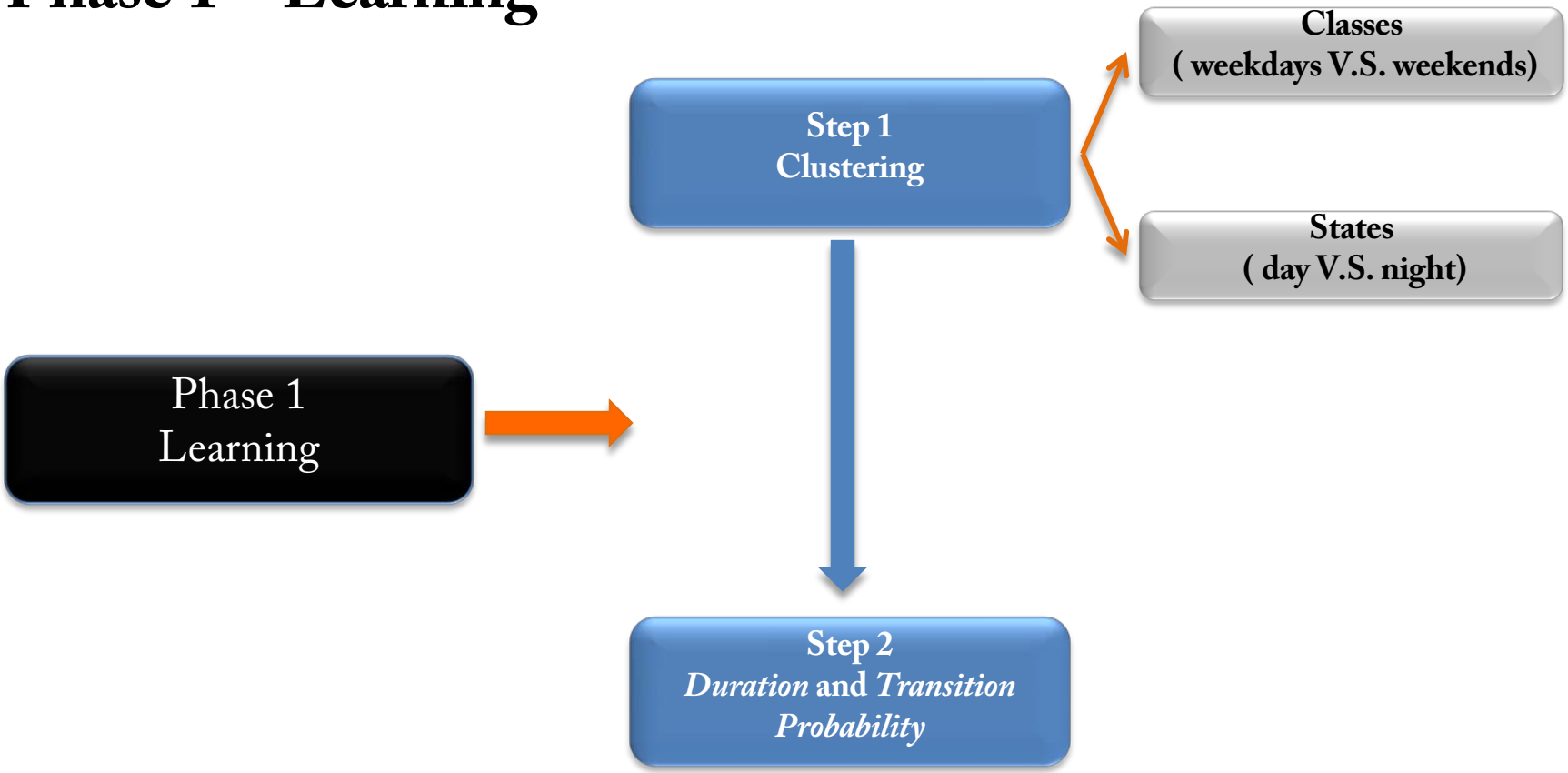
Algorithm: Prediction

- **Prediction-based Scheduling Policy**
  - **Phase 1**



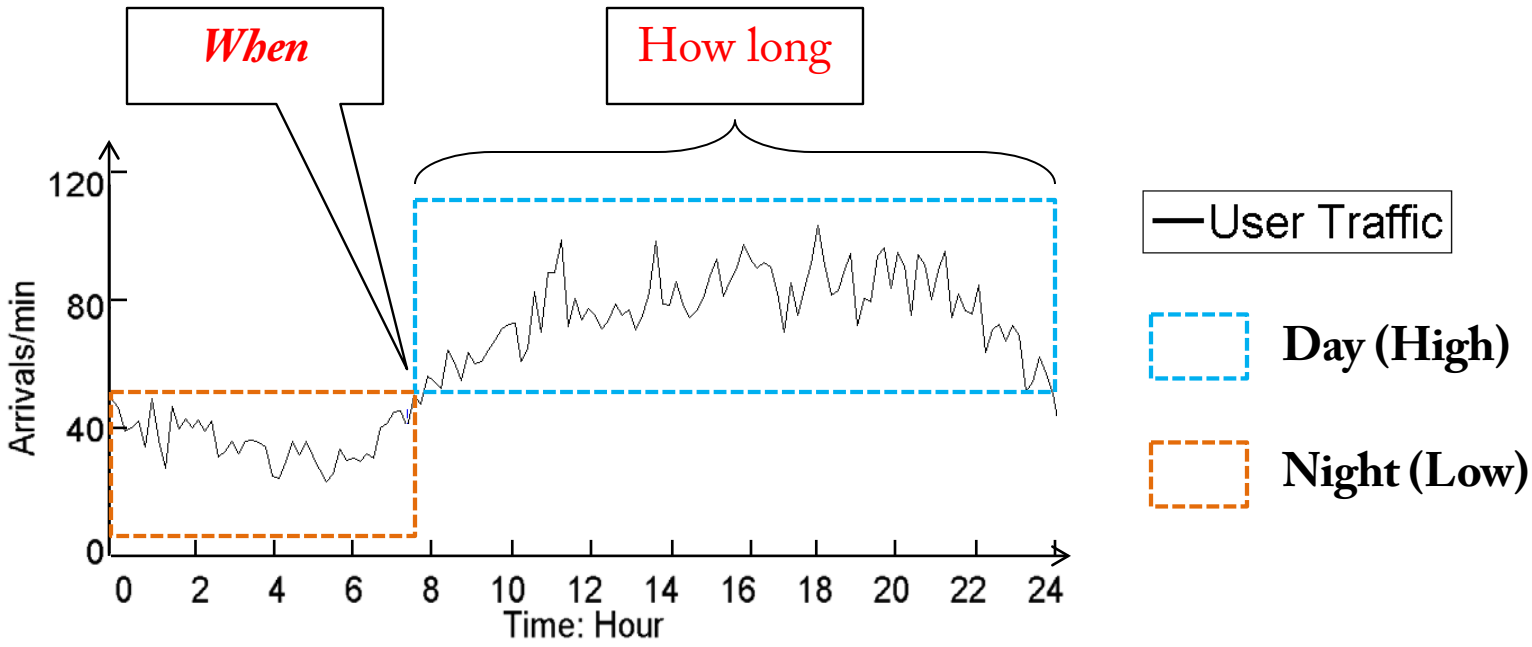
Algorithm: Prediction

- **Phase 1 – Learning**



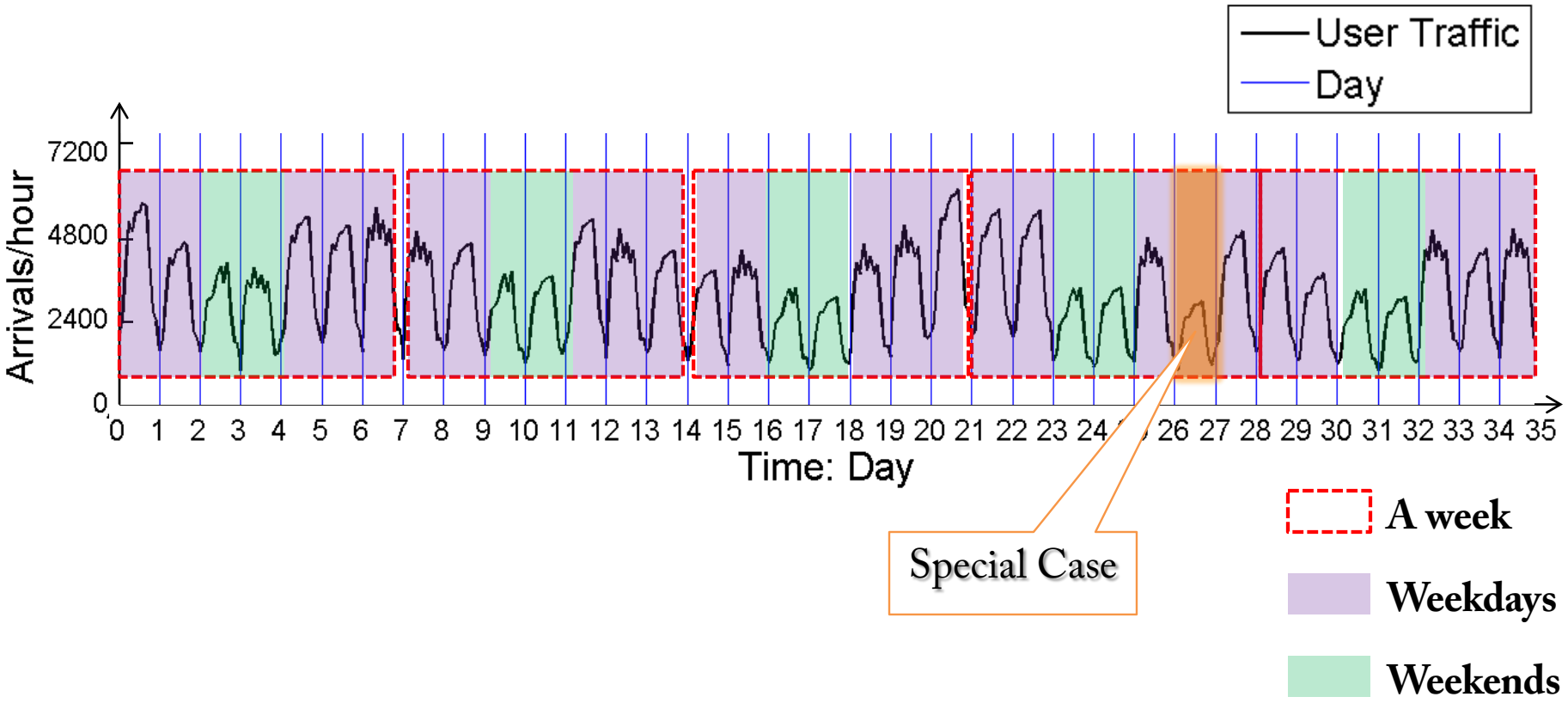
Traffic Intensity Prediction Model

- **Challenges**
  - Arrival Intensity Changes Overtime



Traffic Intensity Prediction Model

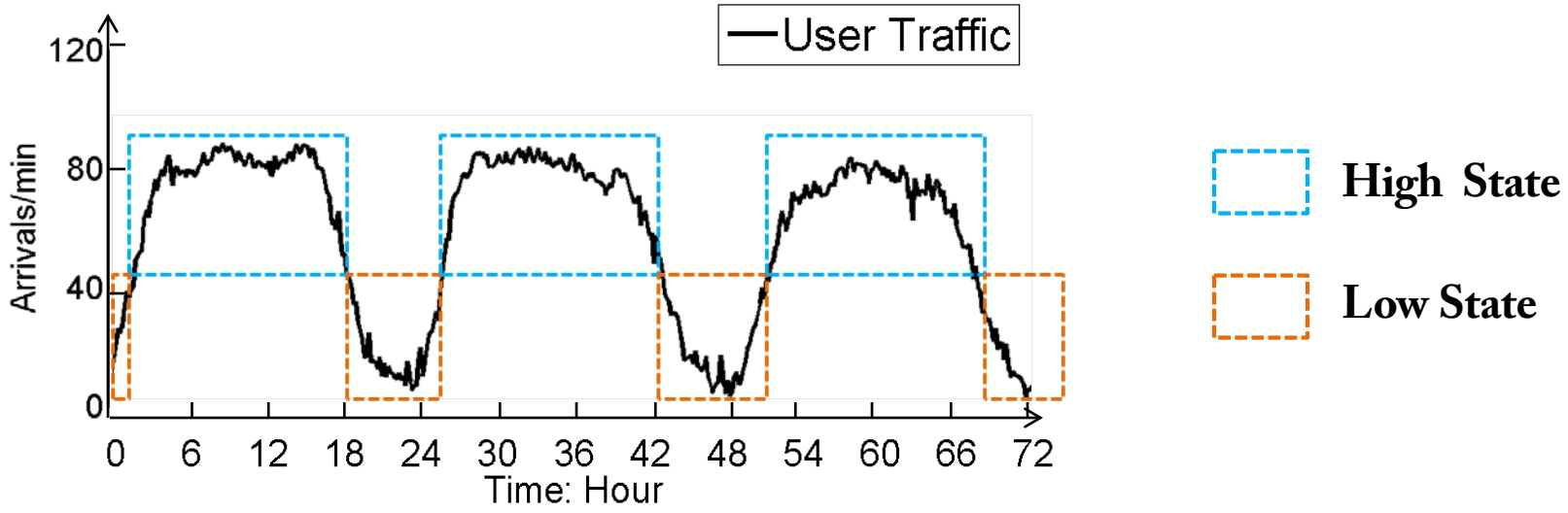
- **Challenges**
  - **Weekly Pattern with Exception**



Traffic Intensity Prediction Model

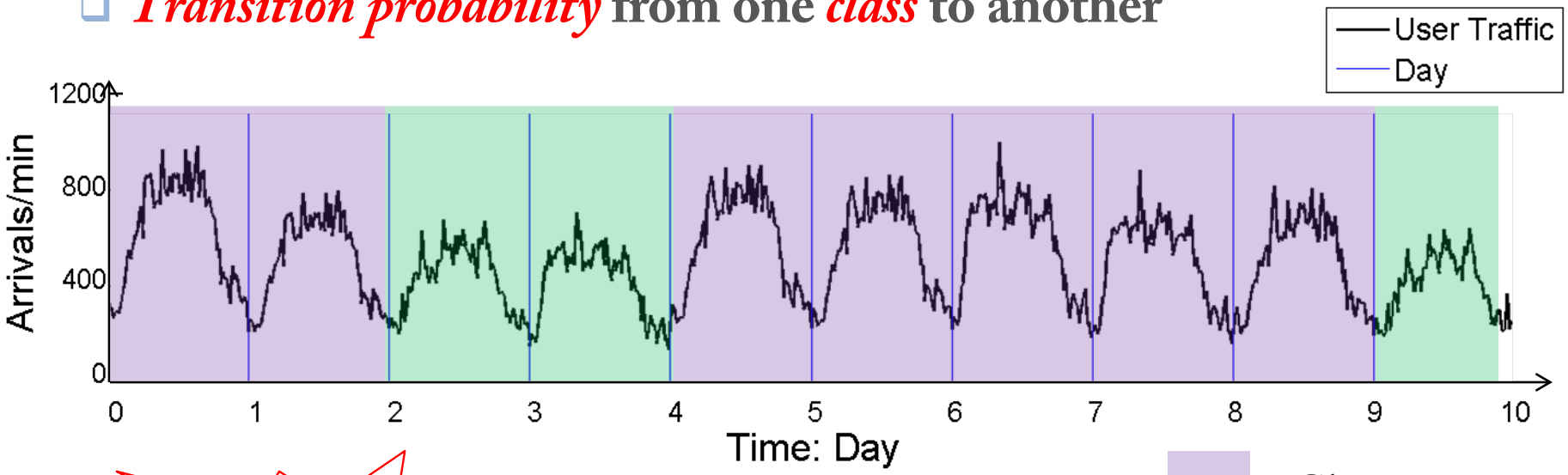
State Classification

- *High/Low state* of arrival intensities
  - Statistical analysis or user defined or other
- *Transition probability* from one *state* to the next
- *Duration* of each *state*



Traffic Intensity Prediction Model

- **Class Classification**
  - Different *types/classes* of overall arrival intensities
  - *Transition probability* from one *class* to another



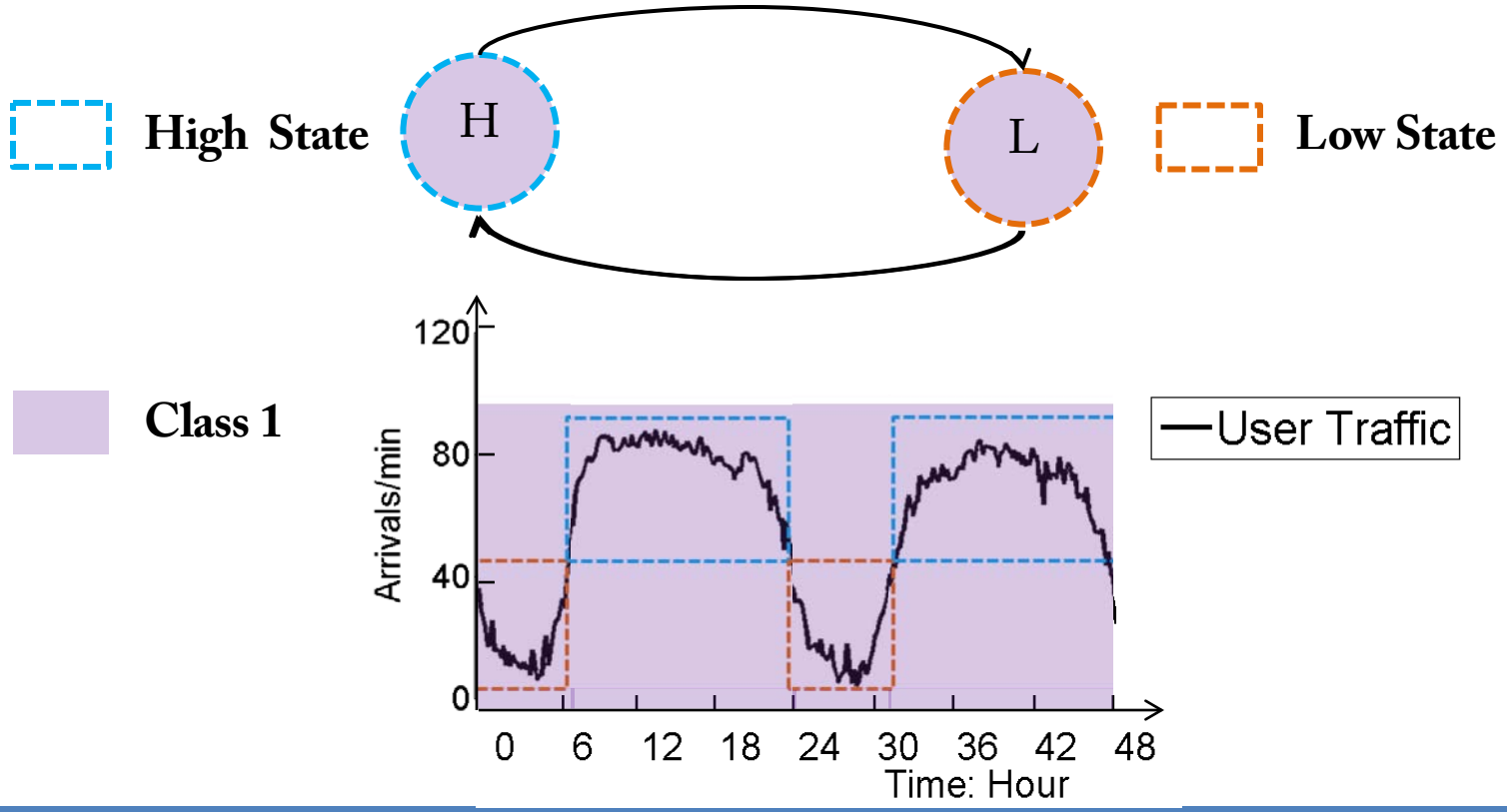
Any number of classes

Class 1

Class 2

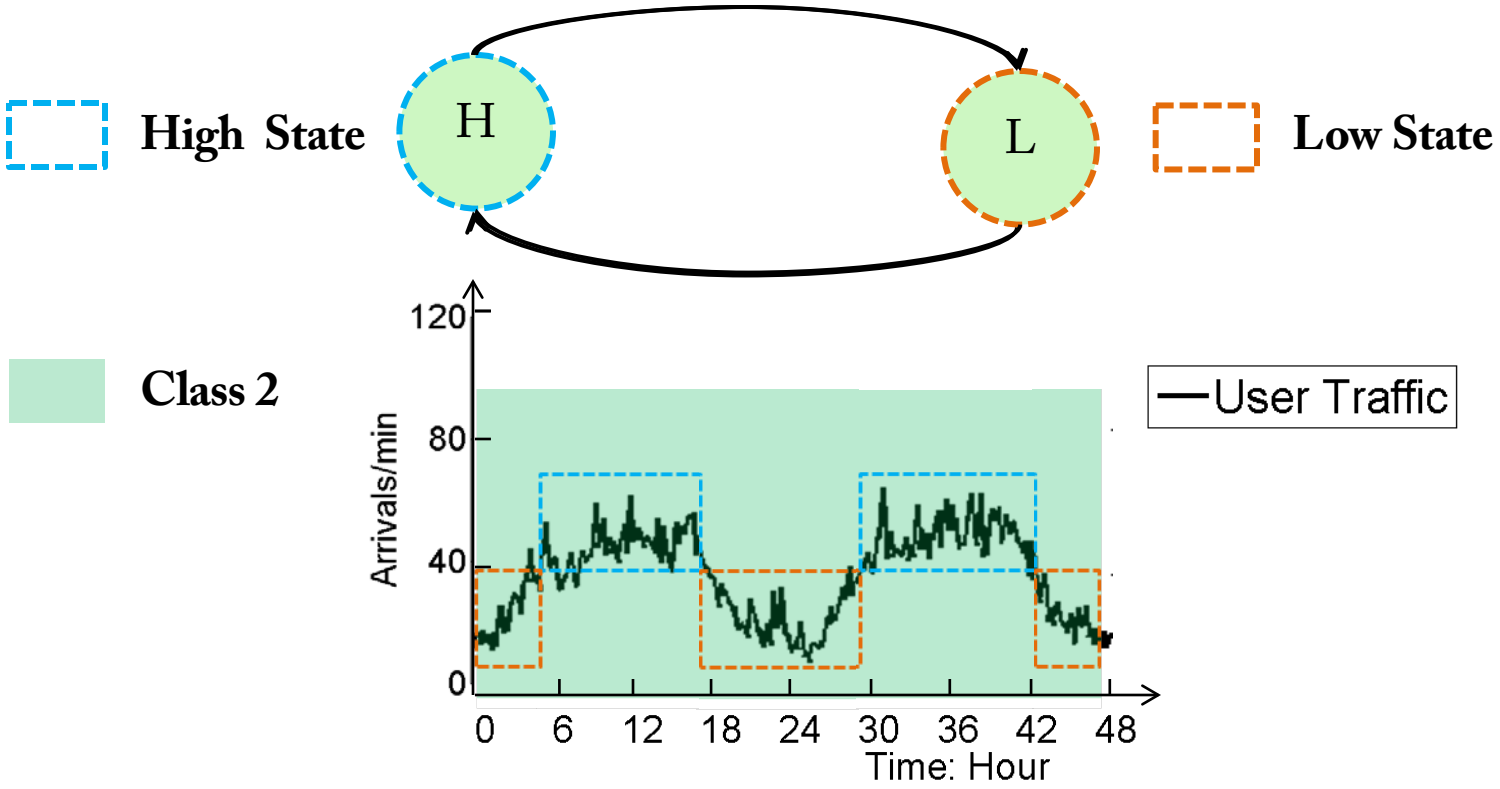
Traffic Intensity Prediction Model

- **Markovian Model**



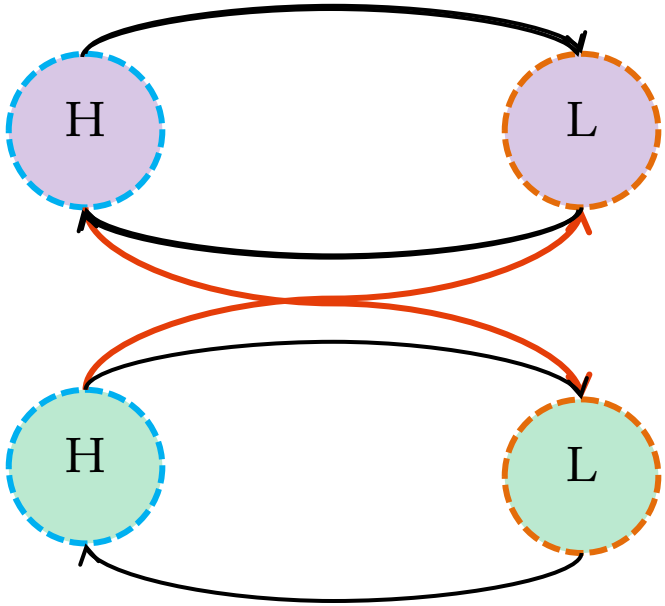
Traffic Intensity Prediction Model

- **Markovian Model**

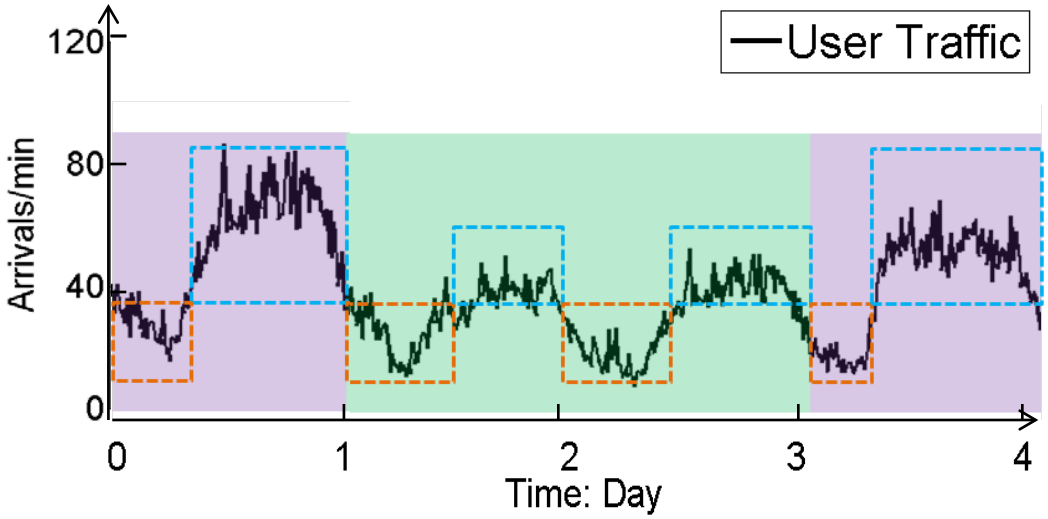


Traffic Intensity Prediction Model

- **Markovian Model**



High State Low State

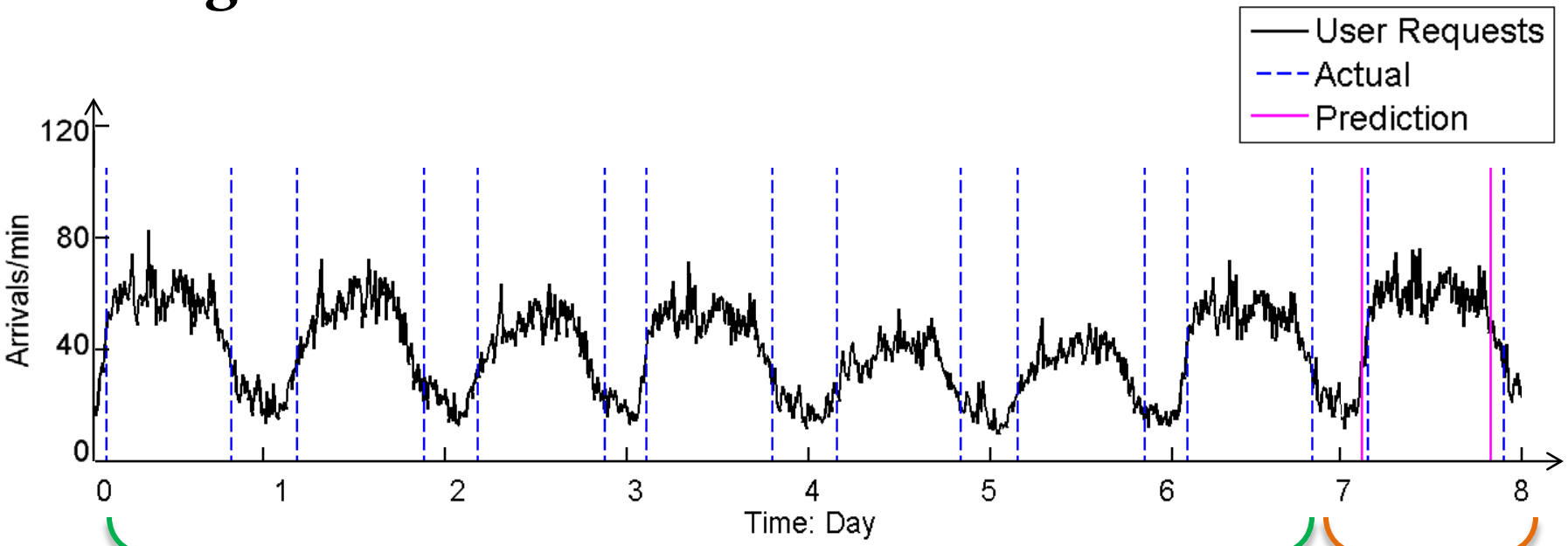


Class 1

Class 2

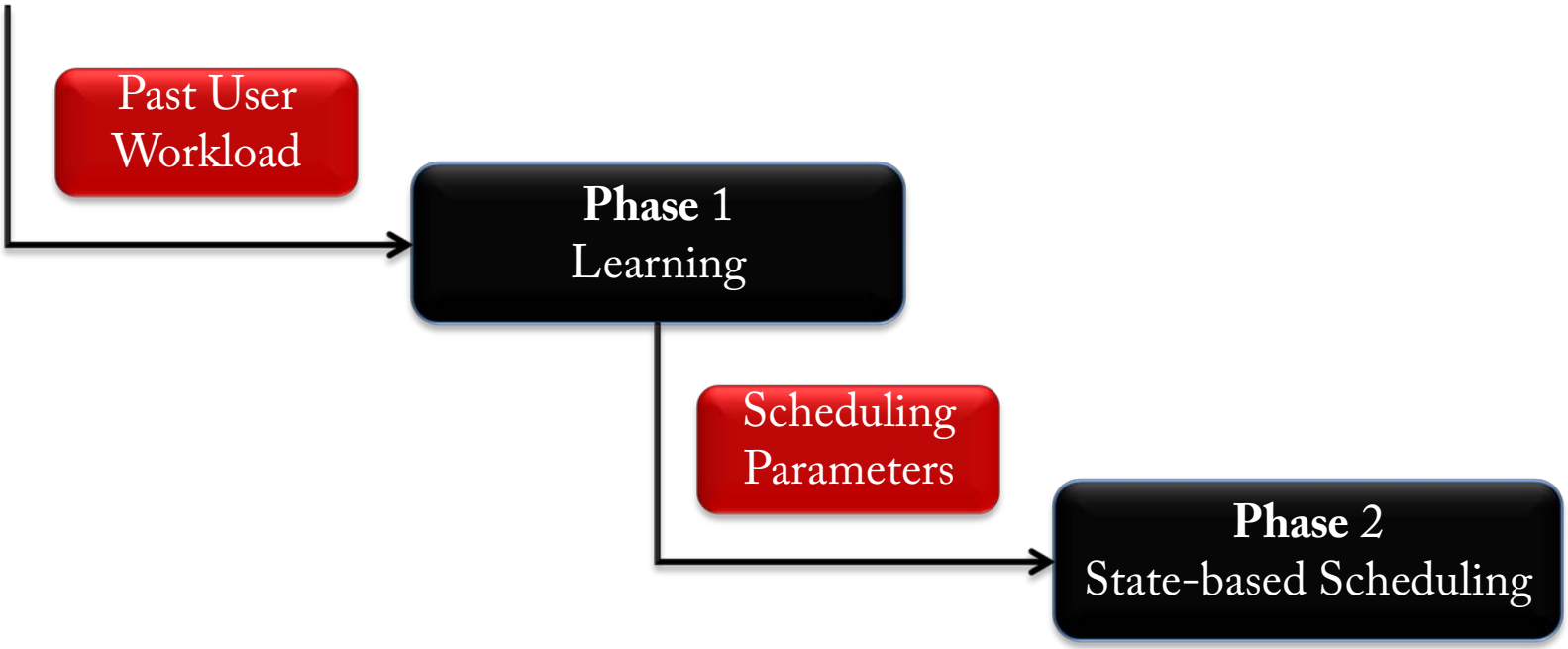
Traffic Intensity Prediction Model

- Comparison between actual and predicted state changes



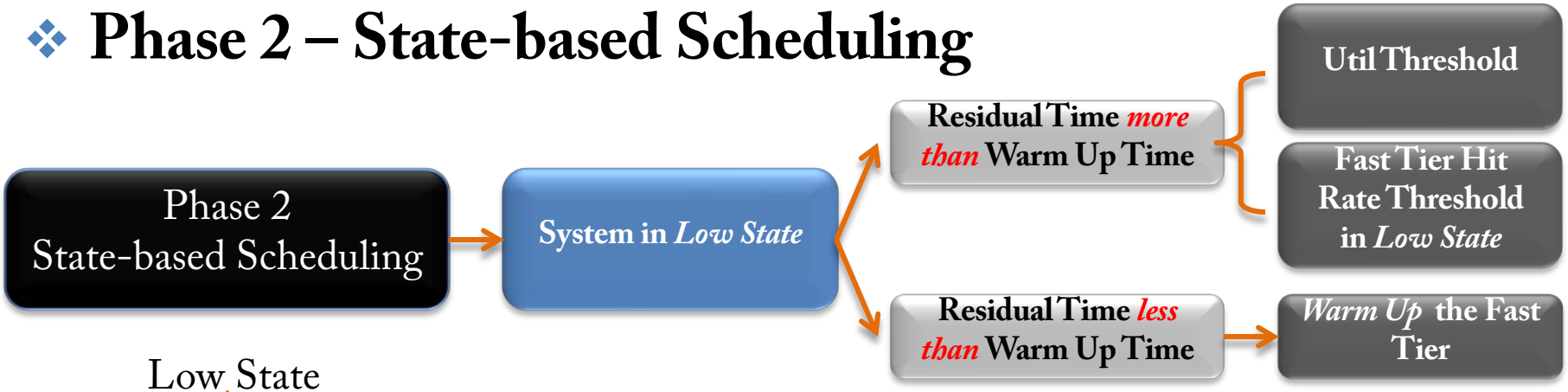
Algorithm: Scheduling

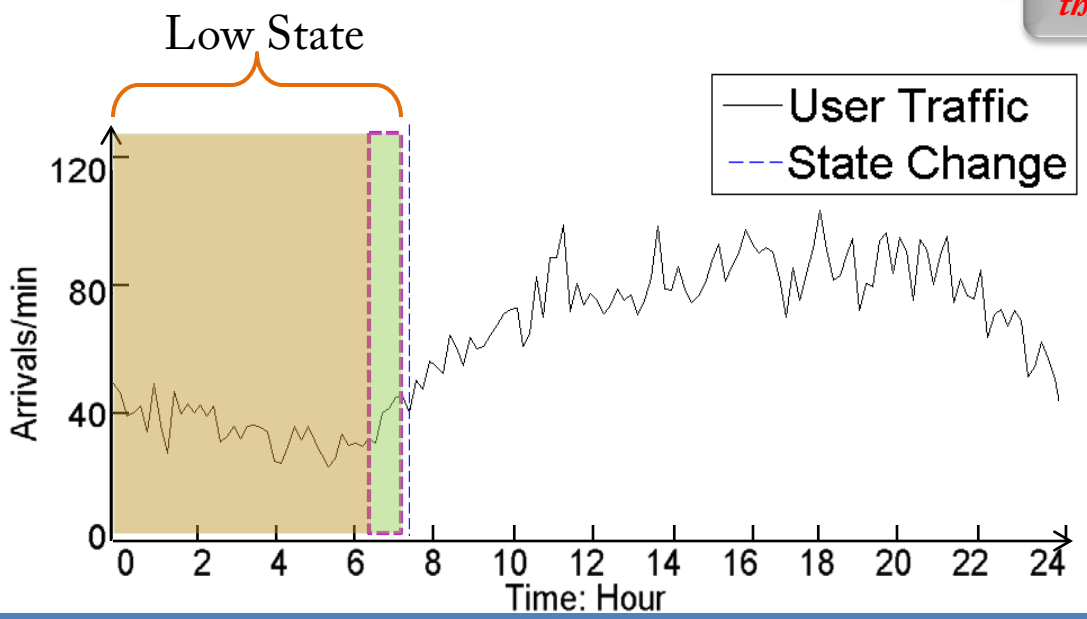
- **Prediction-based Scheduling Policy**
  - **Phase 2**



Algorithm: Scheduling

Phase 2 – State-based Scheduling





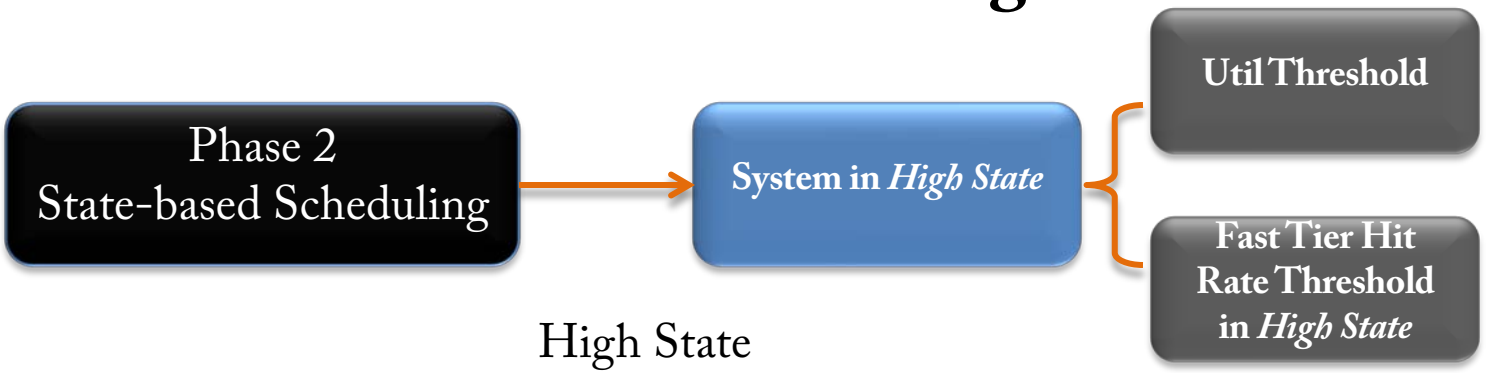
Warming Up

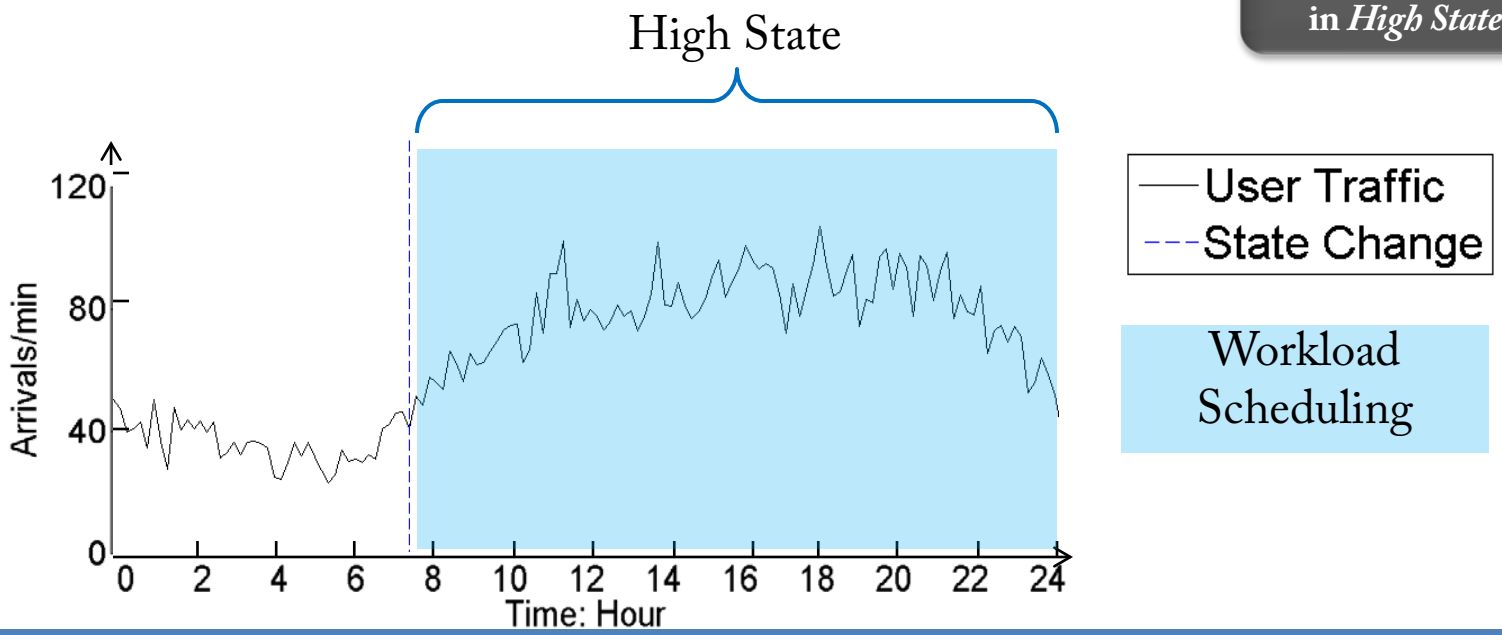
Residual Time

Schedule System & User Traffic

Algorithm: Scheduling

- **Phase 2 – State-based Scheduling**





Fast Tier Hit Rate Model

- **Challenge**
  - *When* to warm up
- **Solution**
  - System measurements
  - A *linear estimation model* for the fast tier hit rate
    - Serving system work *slows down* the service rate of user traffic
    - Stopping serving system work, user service rate *increases*

Experimental Evaluation

- **Experimental Environment**
  - 12 GB Memory
  - Disk enclosure containing 12 SATA 7200RPM HDDS of 3TB each
- **Workload Generation**
  - fio
  - 1 GB user working set
  - 24 GB system working set

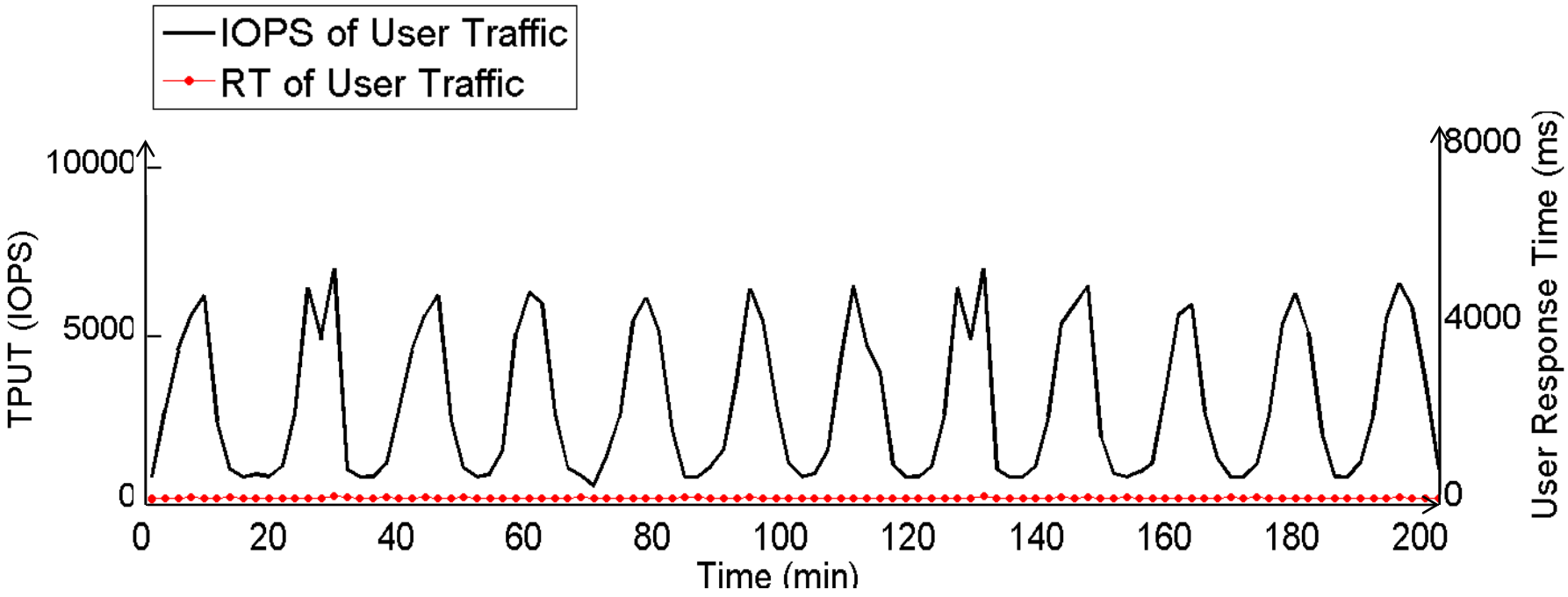
Experimental Evaluation

- **Comparison of Scheduling Policies**
  - User-only
    - No system work scheduled
    - A baseline
  - Feedback-based
    - A reactive policy
    - State-of-the-practice
  - Prediction-based: this work

Experimental Evaluation

- **Measurement Results**
  - User-only

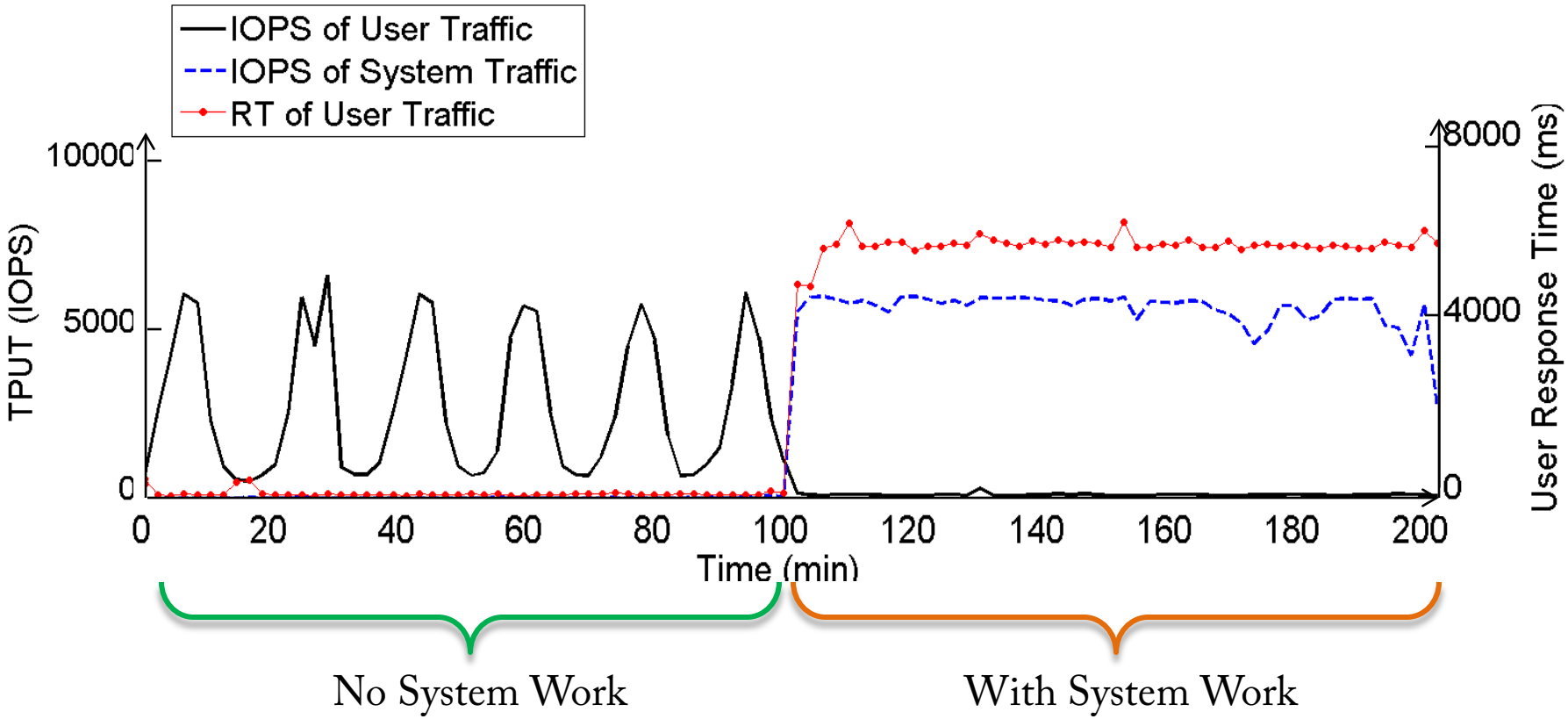
Access Speed Difference: 100x



Experimental Evaluation

- **Measurement Results**

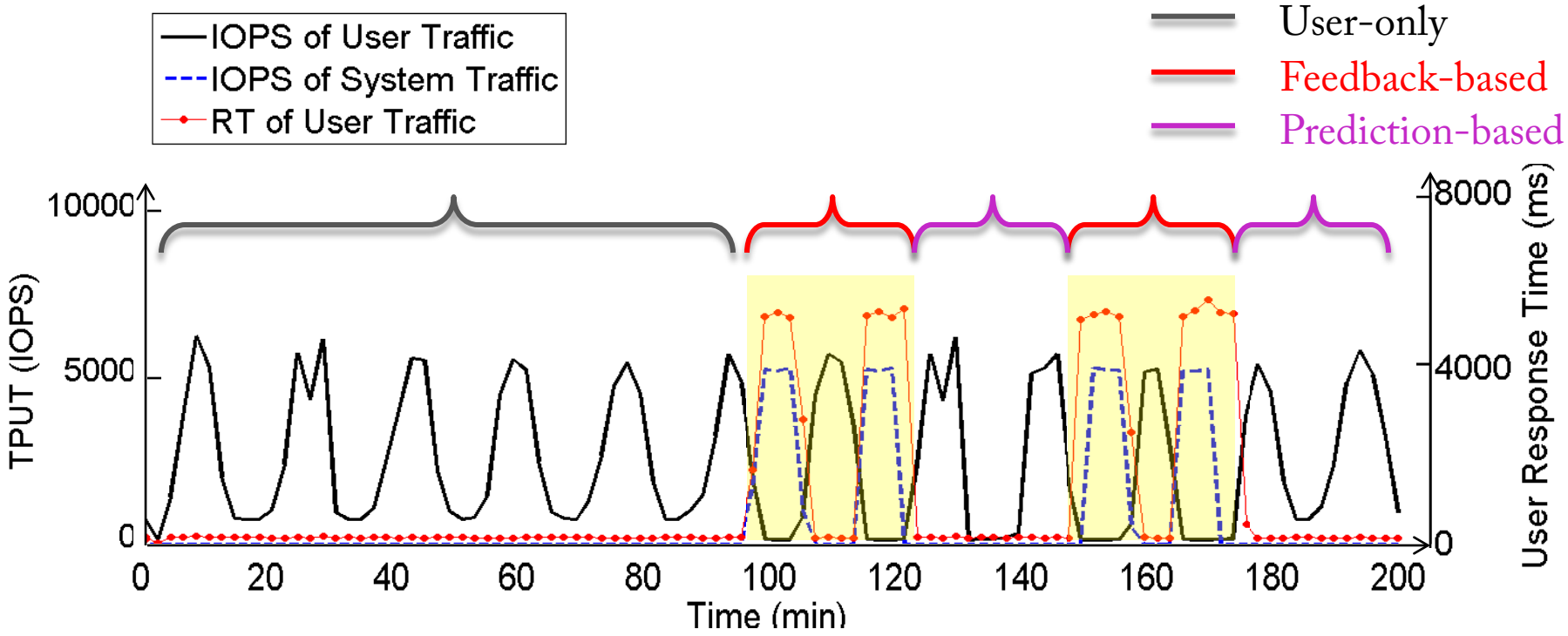
Access Speed Difference: 100x



Experimental Evaluation

- **Measurement Results**
  - User-only
  - Feedback-based *vs* Prediction-based

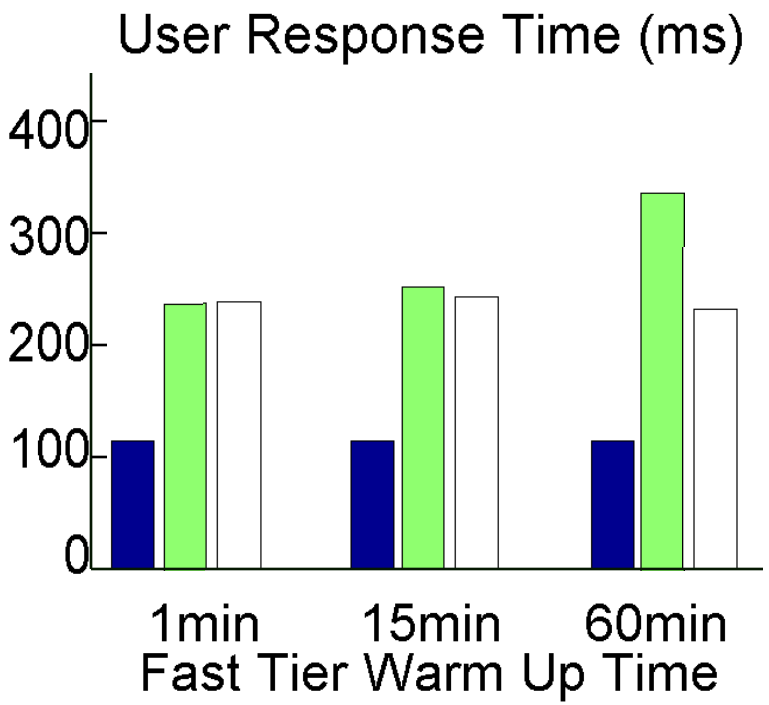
Access Speed Difference: 100x

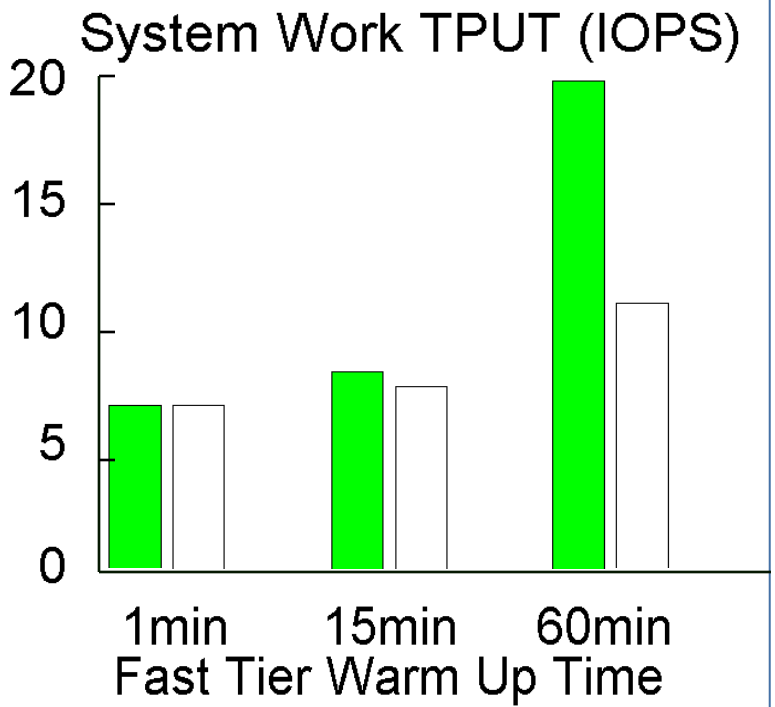


Experimental Evaluation

- **Simulation Results**

Access Speed Difference: 10x

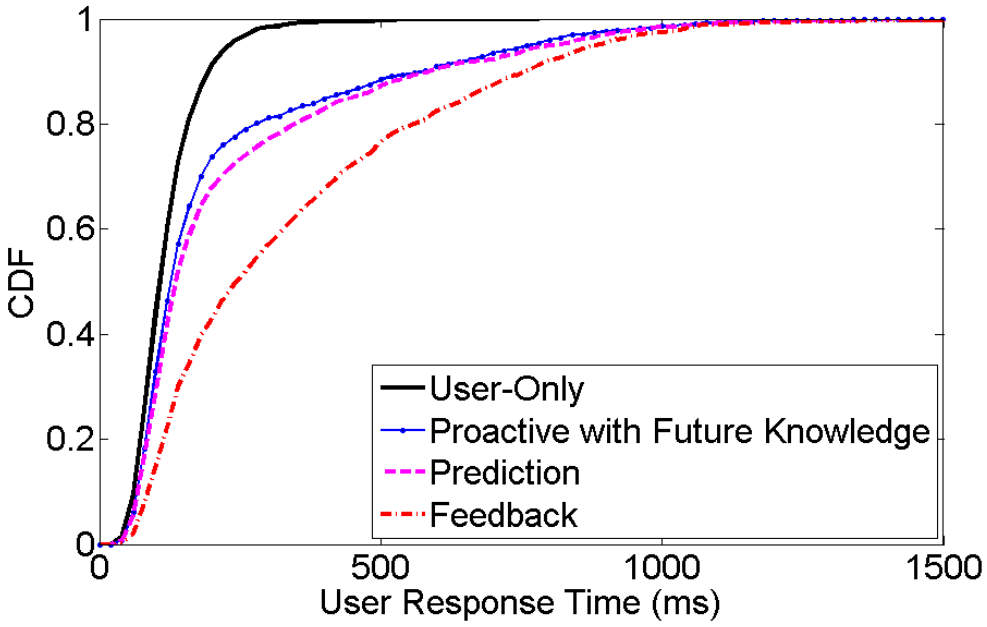




User-Only Feedback Prediction

Experimental Evaluation

- **Simulation Results**



Warm Up Time 60min

Summary

- A Markovian-based model.
- A prediction-based scheduling policy.
- Experiments show effectiveness and robustness.

Future Work

- More Tiers
- More Sensitivity Analysis
  - Different workloads
  - Different warm up times
  - Access speed differences between tiers
  - …

THANK YOU!

Questions?