Information Models: Creating and Preserving Value in Volatile Resources

Chaojie Zhang, Varun Gupta, Andrew A. Chien

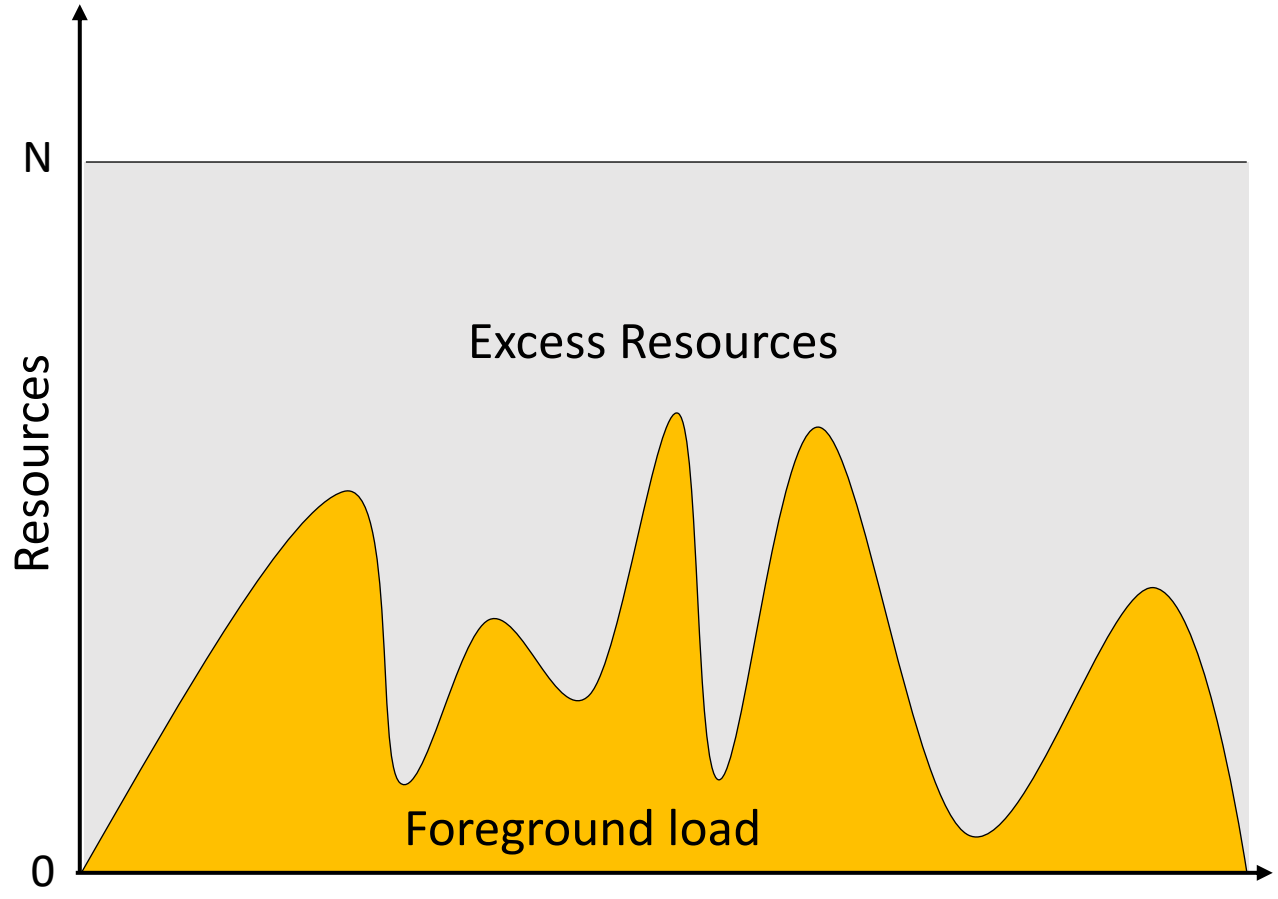
University of Chicago

June 25, 2019

ROSS Workshop

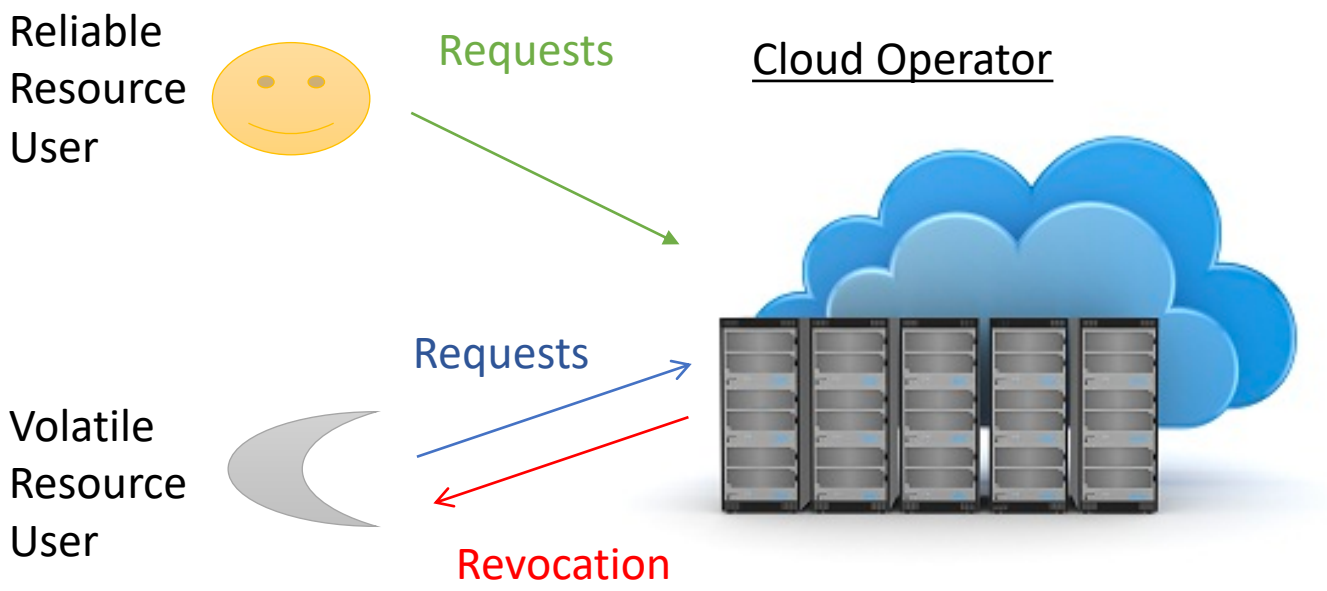
Excess Resources in the Cloud

- IaaS demand expanding
- Demand fluctuates
- Capacities must meet peak demand
- → excess resources
- Excess offered as volatile resources



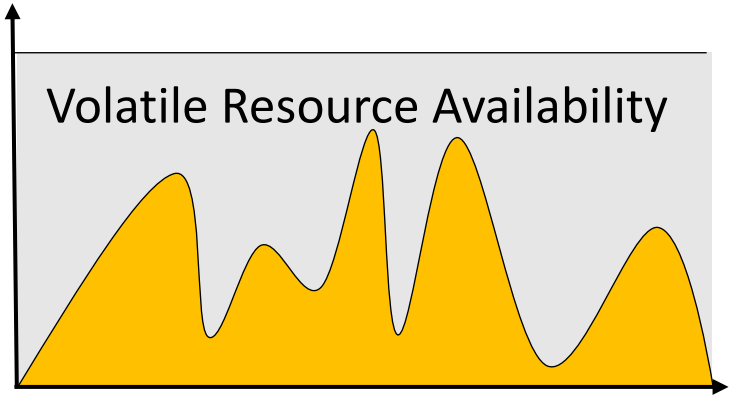
What are Volatile Resources?

- Unreliable, can be unilaterally revoked
- Examples
  - Google Preemptible VMs
  - AWS Spot Instances
- Consequences
  - Wasted work
  - Delayed critical path



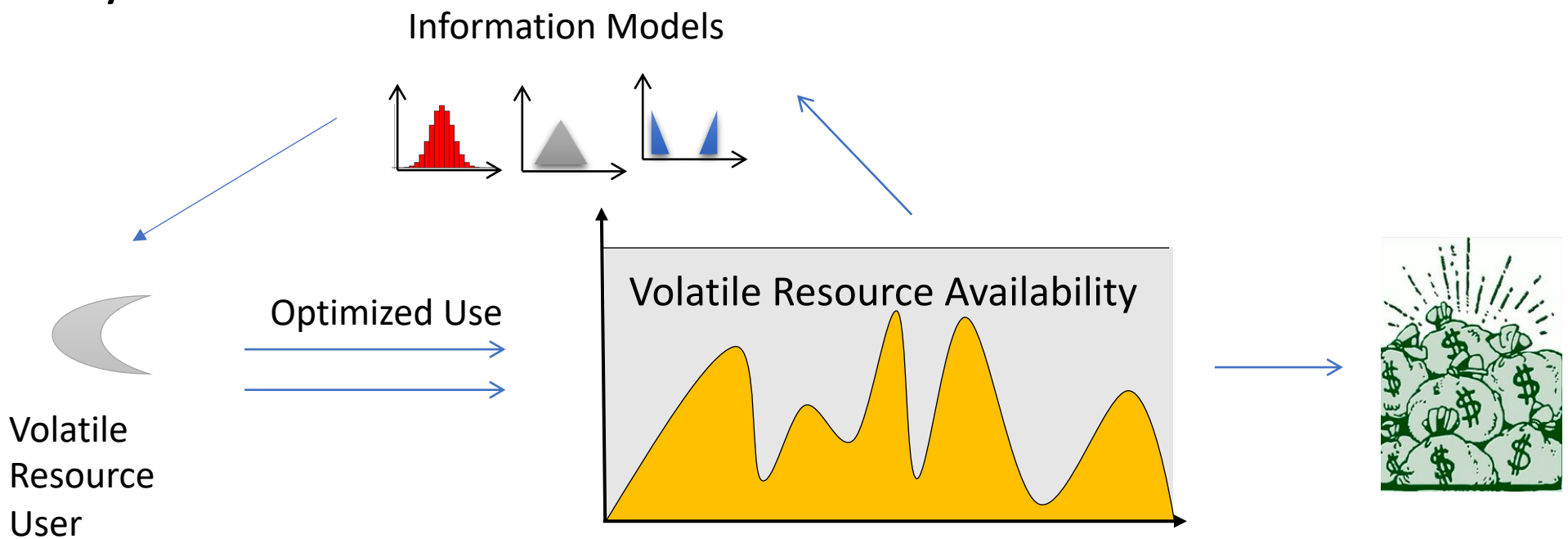
Arming Users with Information

Information Models (summary)



Maximizing Value of Volatile Resources

- What information model do users need to maximize their value of volatile resources
- Assume if user value maximized → cloud providers can sell for more money

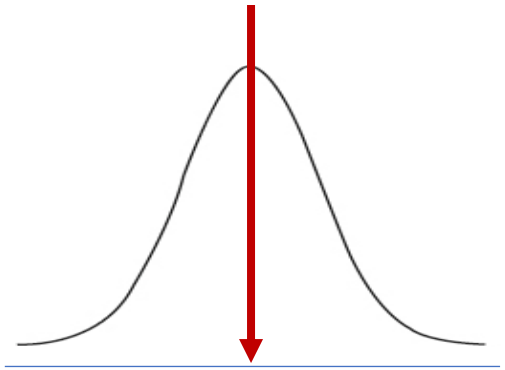


Main Contributions

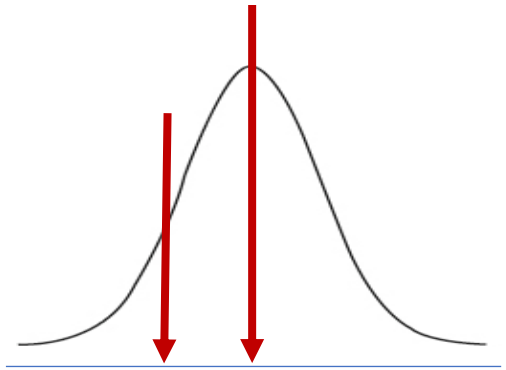
- Show a specific information model that dramatically increases users' ability to achieve value (small)
- Cloud providers can provide information models without compromising internal resource management flexibility
- Results are robust over 608 AWS Spot Instance pools
  - 4 regions, millions of CPUs

Information Models

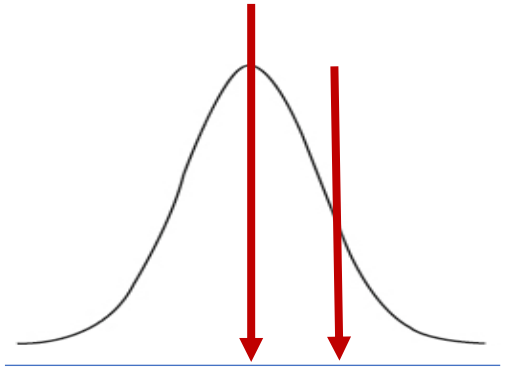
- *What information enables users to target volatile resources to extract most value?*
  - *Interval duration PDF's*



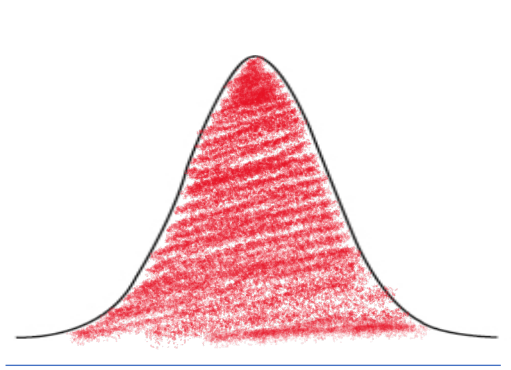
1. MTTR



2. 10pctile



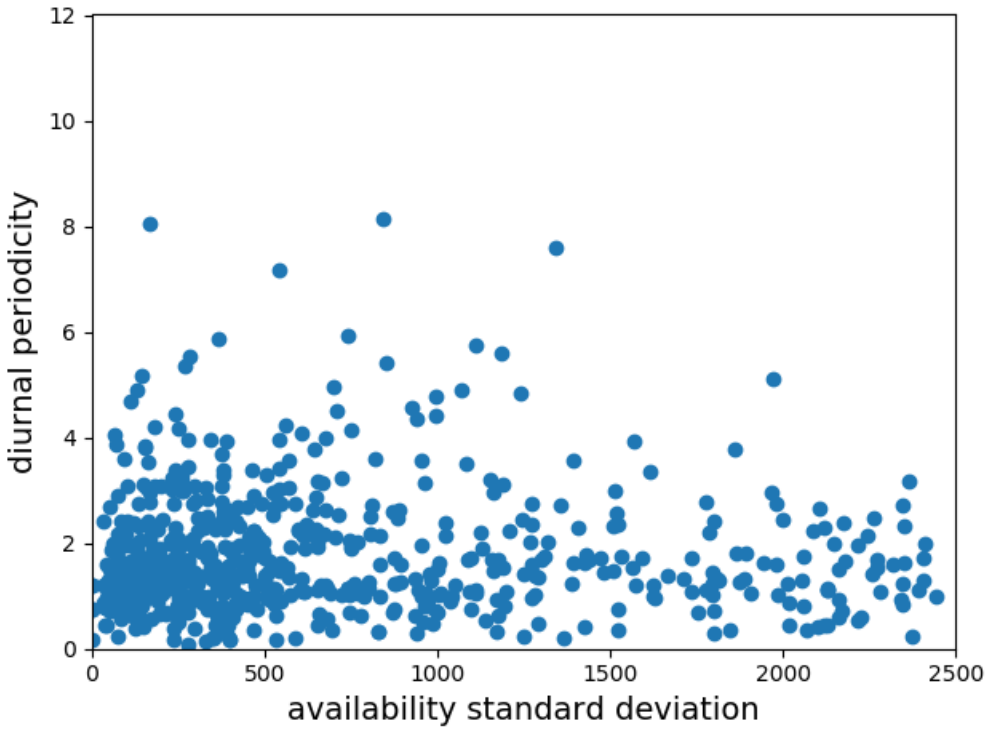
3. 90pctile



Full

Evaluation of Information Models

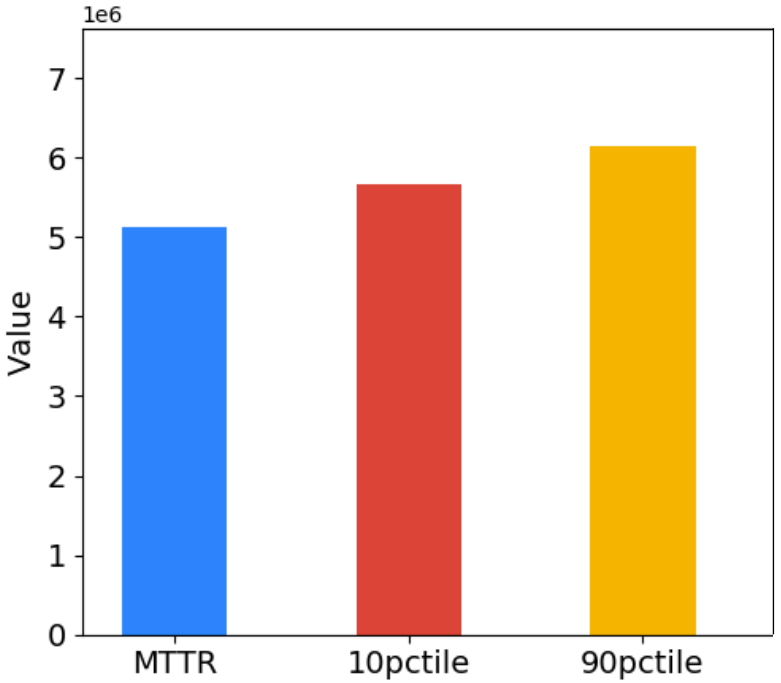
- Resource Dynamics: 3-month 608 AWS Spot Instance pools
  - 5 minute intervals, 15 million data points
- User behavi
  - Match con
  - Maximize
- Utility funct
  - Step funct
- Metrics:
  - Total User

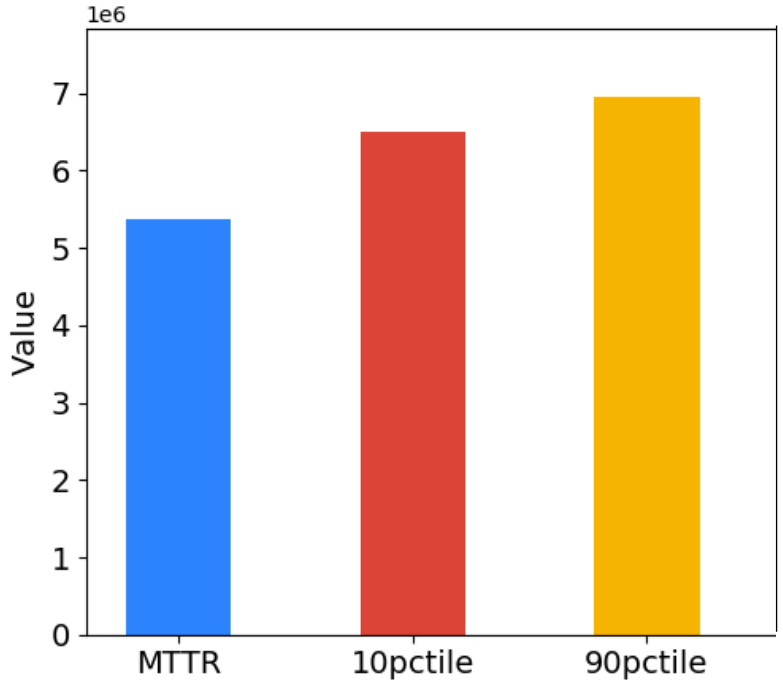


to revocation)

on the intervals

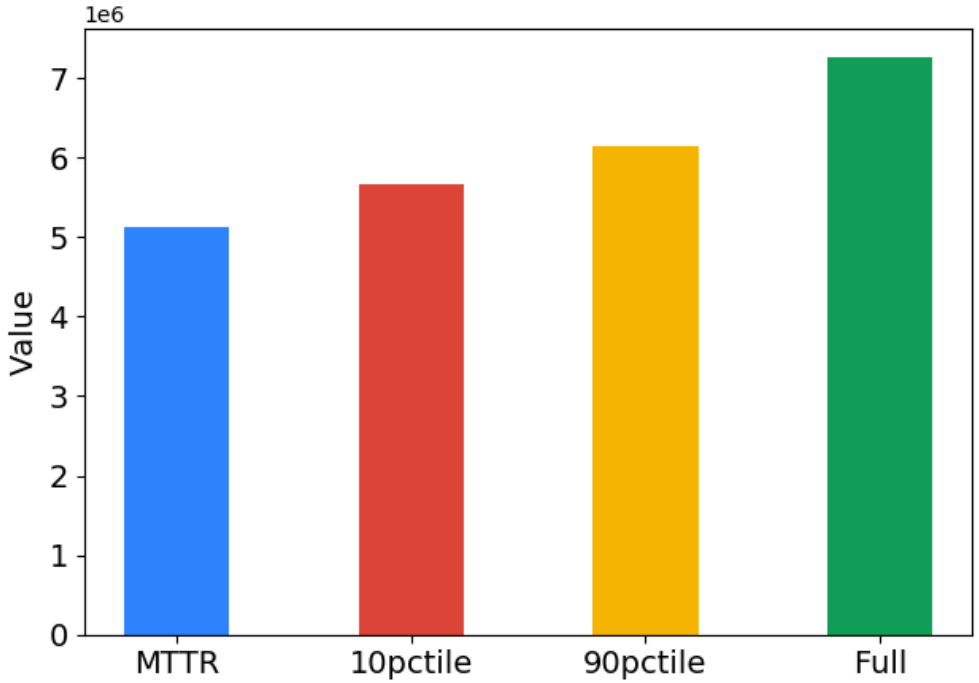
Evaluation: Total Value vs. Information Models

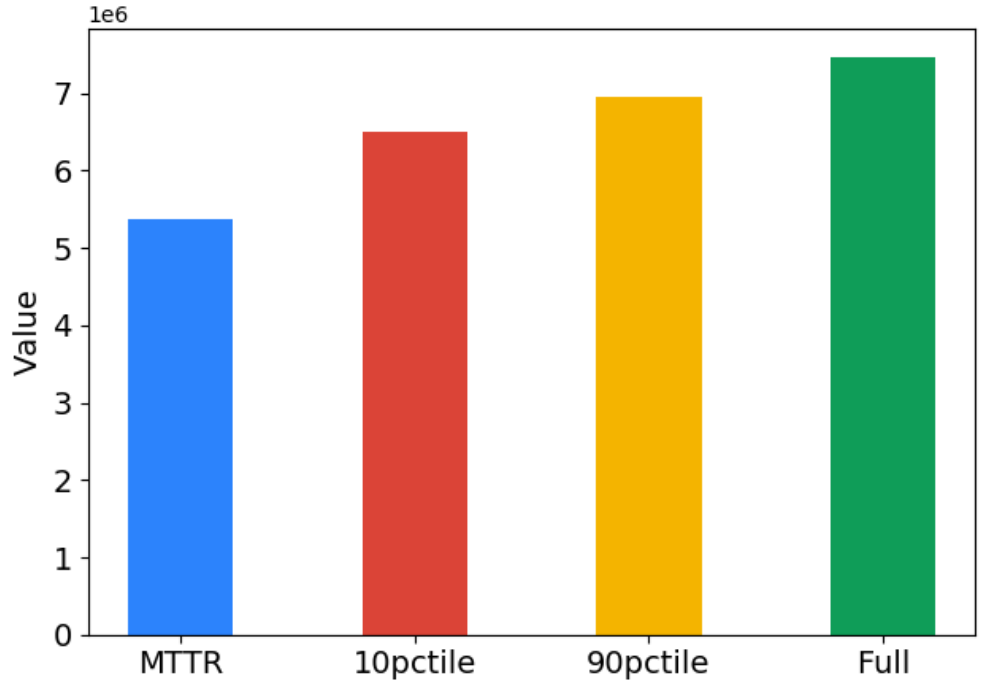




- Comparing three information models
- **90pctile** gives best results
  - 30% value increase

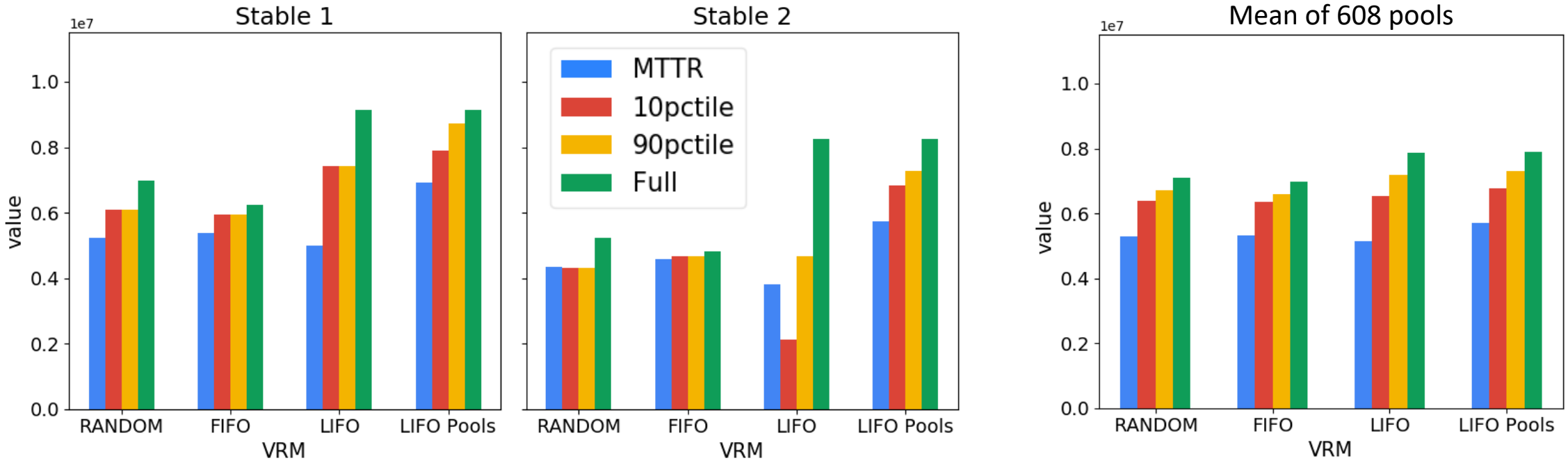
Evaluation: Total Value of Information Models





- Comparing three information models, and **Full** is a reference
- **90pctile** gives best results
  - 30% value increase
- Limited information models can achieve most of the benefit of **Full**, 90%
- Results are robust over vast majority of 608 instance pools

Evaluation: Robustness of Info Model Benefit



- But, cloud providers use a range of volatile resource management (VRM, revocation) policies?
- Information Model benefit and ordering is robust across
  - A range of VRMs
  - All 608 instance pools

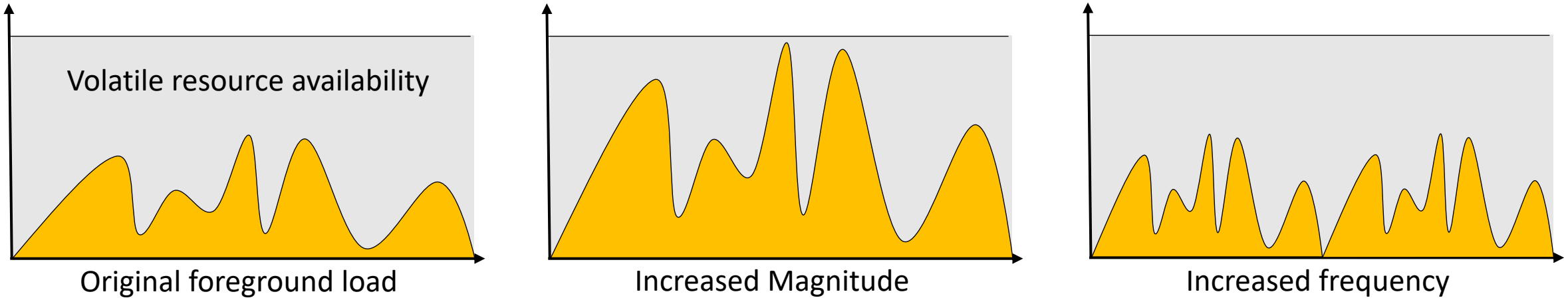
Information Models: Summary

- It's hard for users to maximize value with no information, and cloud providers afraid of sharing too much
- With just limited information (mean + 90th percentile) dramatically increase user value

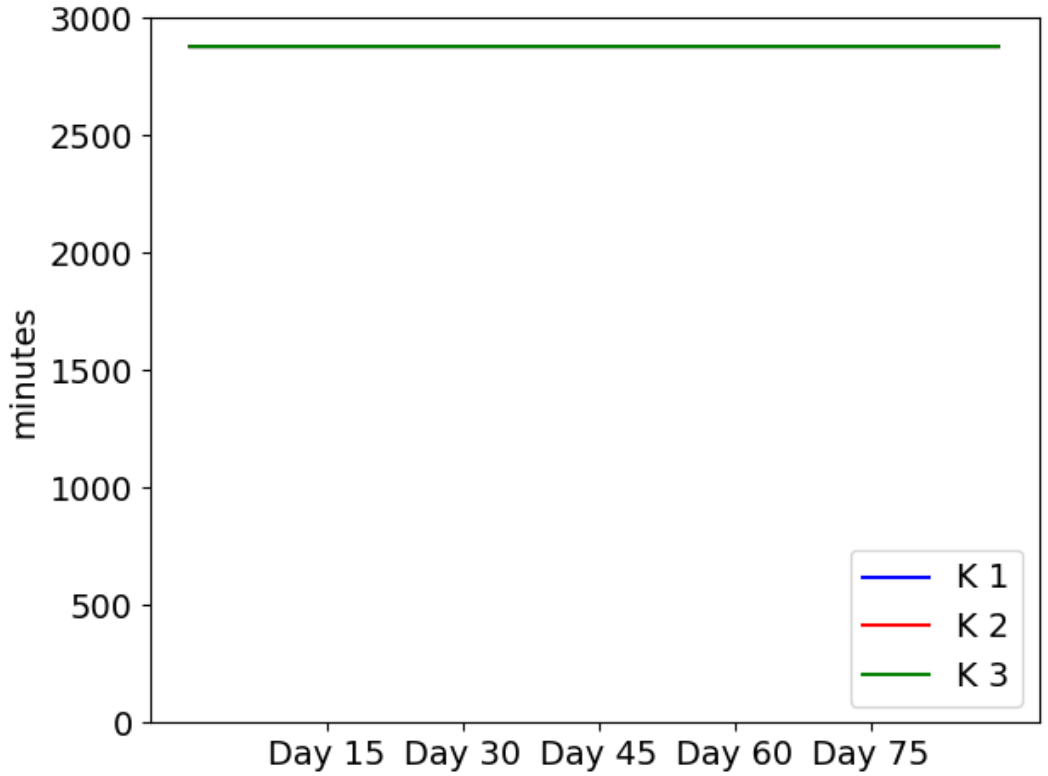
- However, cloud providers worry that information model will constrain resource management

Challenge: Statistical Guarantees and Resource Management "Freedom"

- So, if we gave out an information model (statistical guarantee) : *Does it constrain resource management?*
- Changed foreground load → Changed statistics

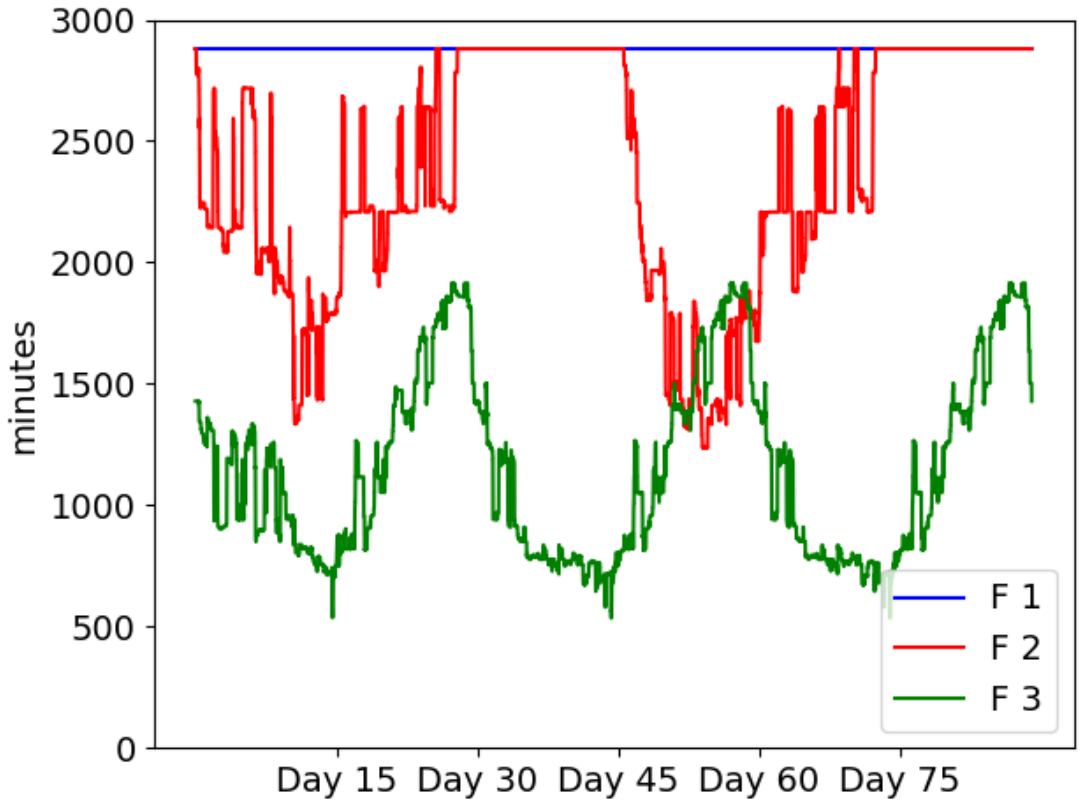


What about a Change in Magnitude?



- Consider drastic reduction in volatile resources (1->1/K)
- K = 1, 2, 3
- How does this affect 90pctile?
  - 2-week sliding window
- **Magnitude change has no impact on 90pctile statistical guarantees → No constraint!**

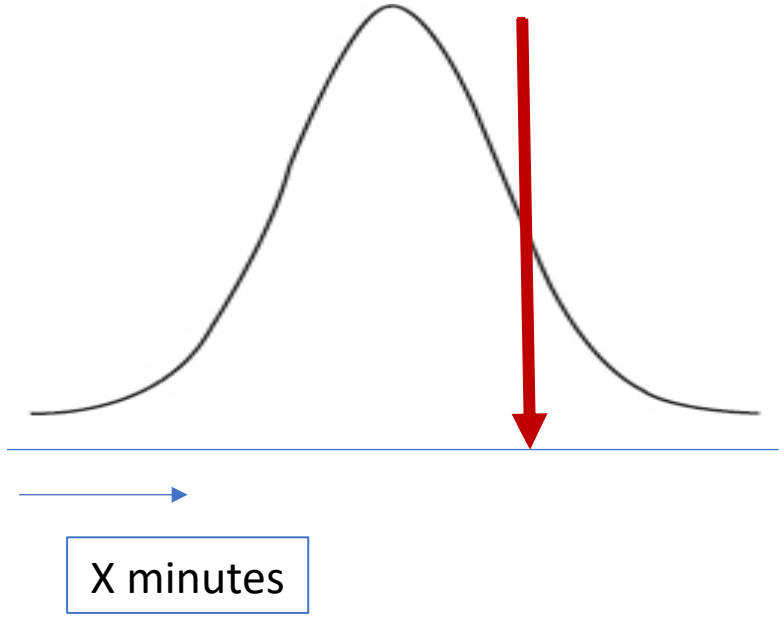
What about a Change in Frequency?



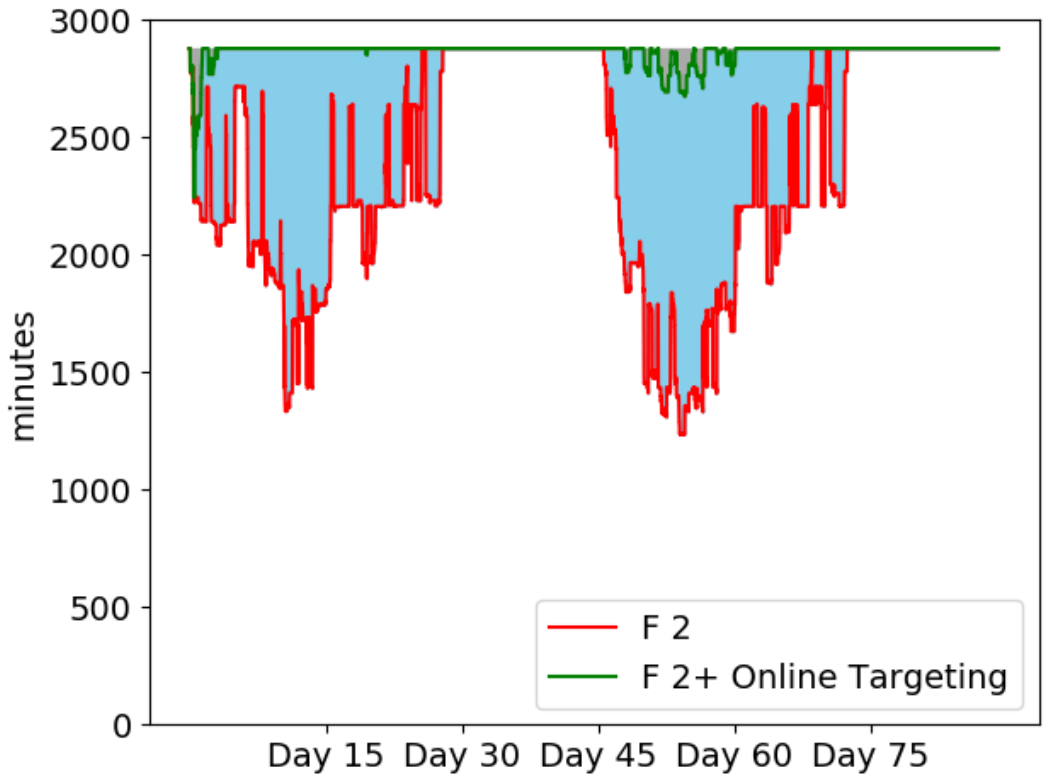
- Increase volatile resource variation frequency by contracting time base (1->1/F)
- F = 1, 2, 3
- How does this affect 90pctile?
  - 2-week sliding window
- **Frequency change reduces 90pctile dramatically**
- **Violates the guarantee!**

Can We Preserve the Guarantee?

- Idea: Guarantee-Preserving Resource Management
  - Maintain 90pctile guarantee under frequency change
- Offline Static Algorithm
  - Reshape the distribution by withholding each interval for X minutes
    - kills short intervals, shortens long intervals
  - What is the best X?
    - Find smallest X that preserves guarantee

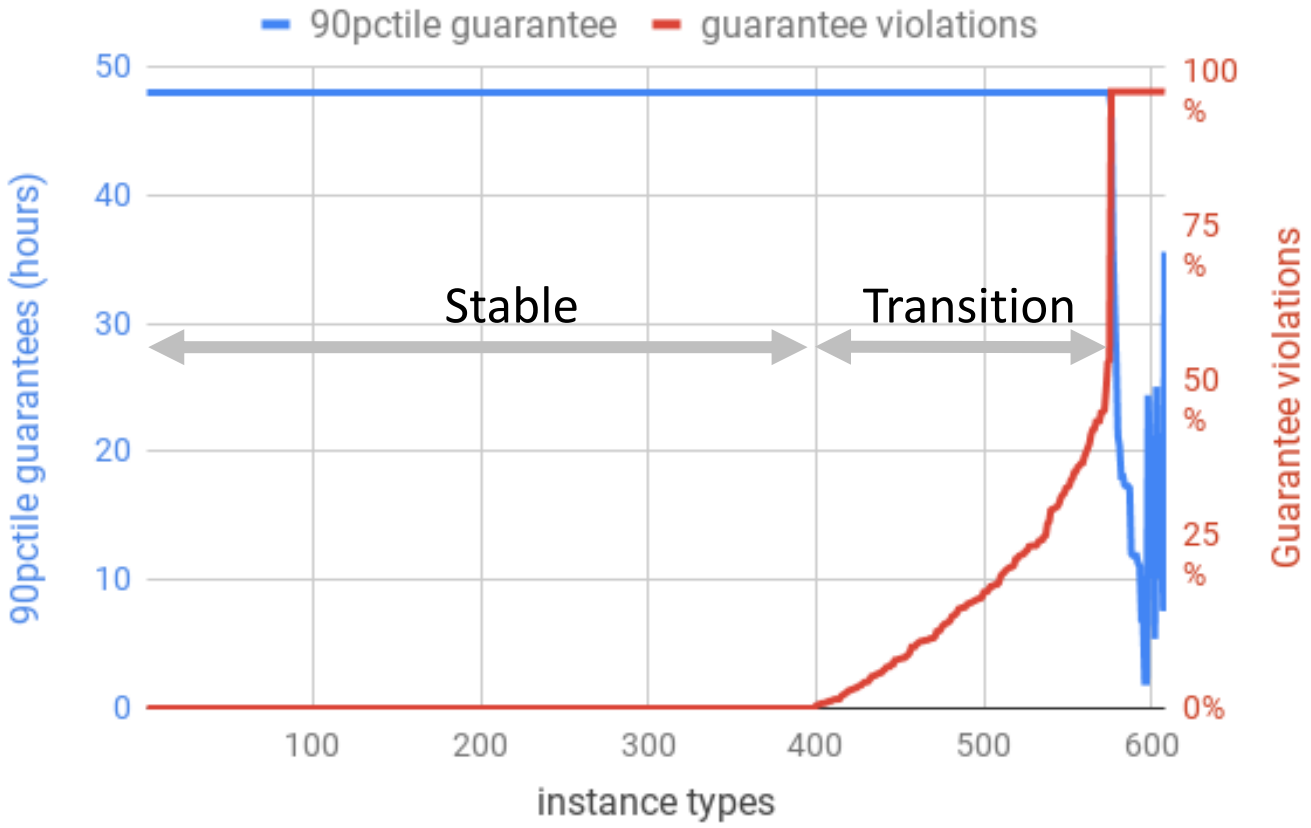


Online Dynamic Algorithms



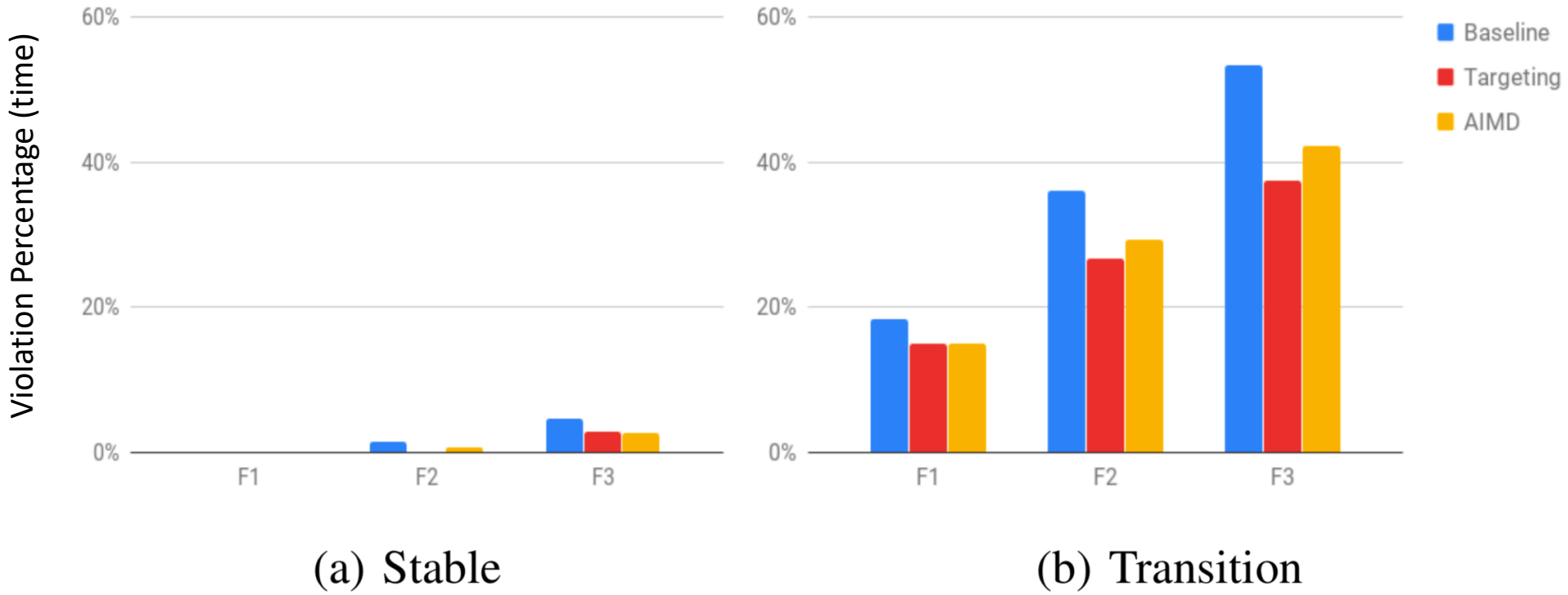
- Idea: AIMD, Online Targeting
- Doubles the 90pctile – preserves the guarantee and reduces job failures
- Info Model => Good user value
- Preserving RM => Providers' flexibilities

Classifying 608 Instance Types



- 3 Classes of Instance Types
  - Stable, Transition, Unstable
- 400 Stable
  - The 90pctile is consistent
- 177 Transition
  - 90pctile guarantee is matched most of the time
- 31 Unstable
  - 90pctile unstable, low, unusable

Evaluation: Preserving 90pctile Guarantees



(a) Stable (b) Transition

- Guarantee Preserving Algorithms
  - Effective for Stable pools
  - Helpful for Transition pools

Related Work

- Volatile Resource Characterization
  - Characterization of price [Javadi 2011, Tang 2012, Wolski 2017], revocation behavior [Chohan 2010]
- Engineering Reliable Resources
  - Checkpointing [Khatua 2013], replication [Voorsluys 2012, Xu 2016], migration [Yi 2013, Jung 2013]
  - Construct an “economy class” of nearly reliable resources [Carvalho 2014]
- Value of Information
  - Transient guarantee [Shastri 2016]
- Guarantee Preserving Algorithms
  - None

Summary & Future Work

- Small information model → large increase in user value
  - 90pctile info model: two numbers
  - 30% average increase, up to 2X
  - 90% of the benefit of full disclosure
- Guarantee preserving algorithms can preserve guarantees and maintain cloud provider's flexibility
- Results robust over 608 AWS Spot Instance pools

- For more information: http://zccloud.cs.uchicago.edu/ and
- *Chaojie Zhang, Varun Gupta, and Andrew A. Chien, Information Models: Creating and Preserving Value in Volatile Cloud Resources , in the IEEE International Conference on Cloud Engineering (IC2E), June 2019, Prague, Czechoslovakia*.