Application-Aware Latency Monitoring for Cloud Tenants via CloudWatch+

Liu Dapeng, Dan Pei, Youjian Zhao

Tsinghua University

+

Web

Latency matters for web!

+500ms lead to a revenue decrease of 1.2%

[Eric Schurman, Bing]

+100 to 400ms reduced #searches/user by 0.2% to 0.6%

[Jake Brutlag, Google]

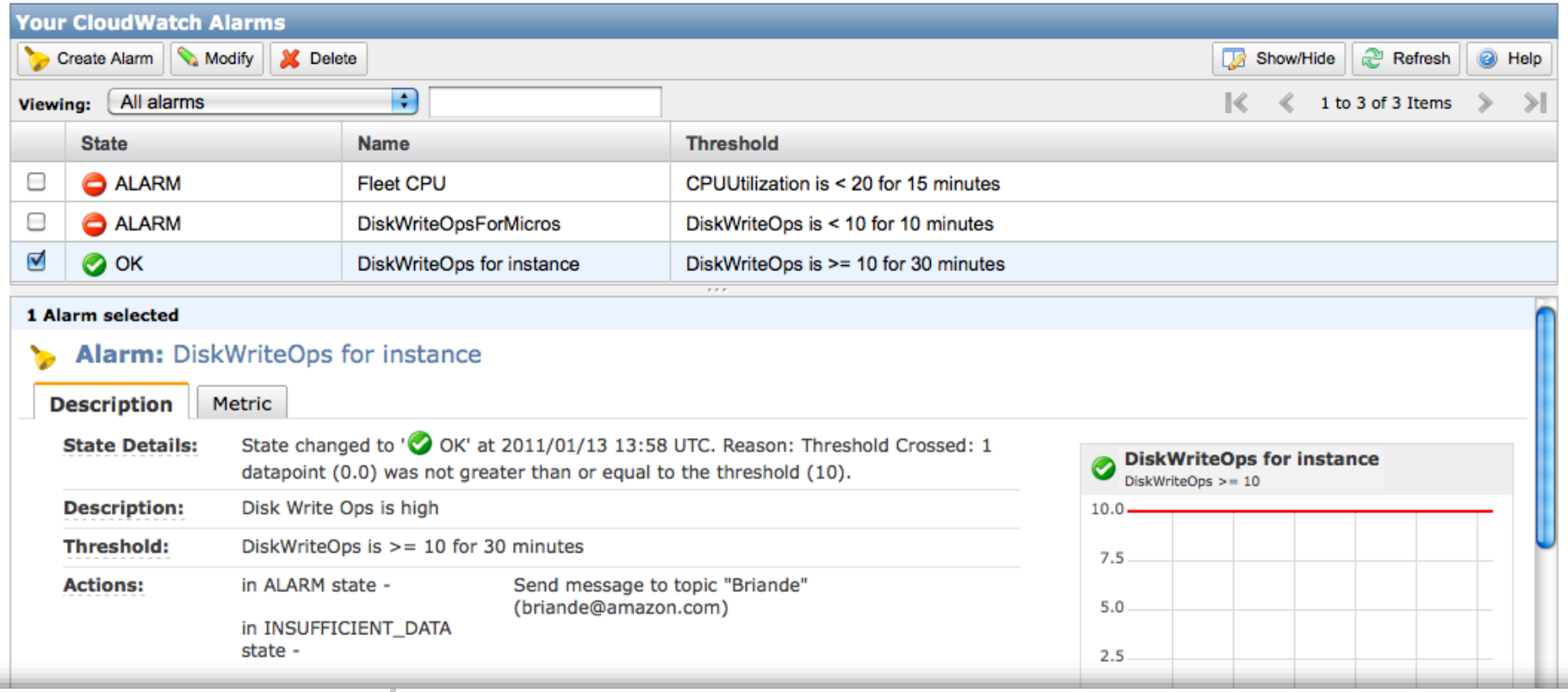
+100ms in latency lead to a 1% drop in sales

[Greg Linden, Amazon]

+1000ms reduced page view by 11%

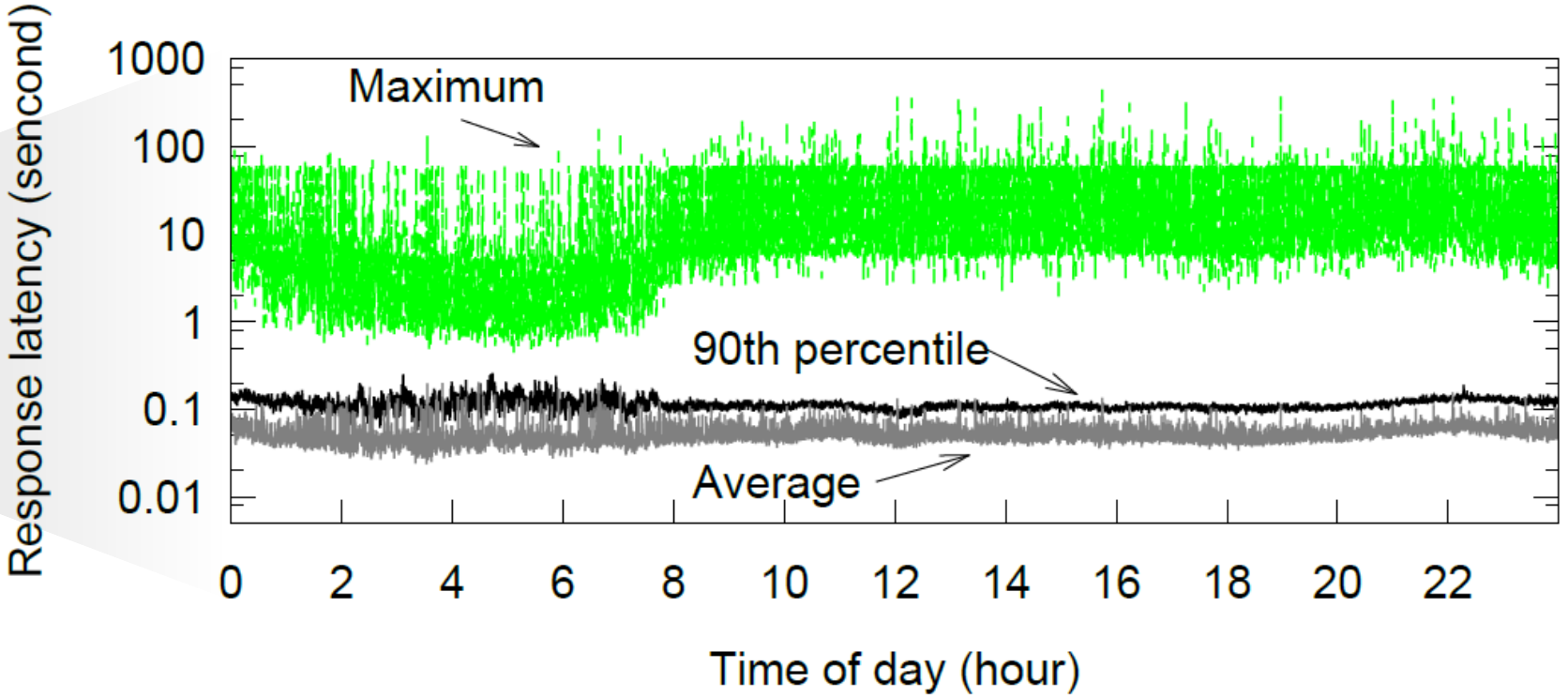
[Simic Bojan, Aberdeen]

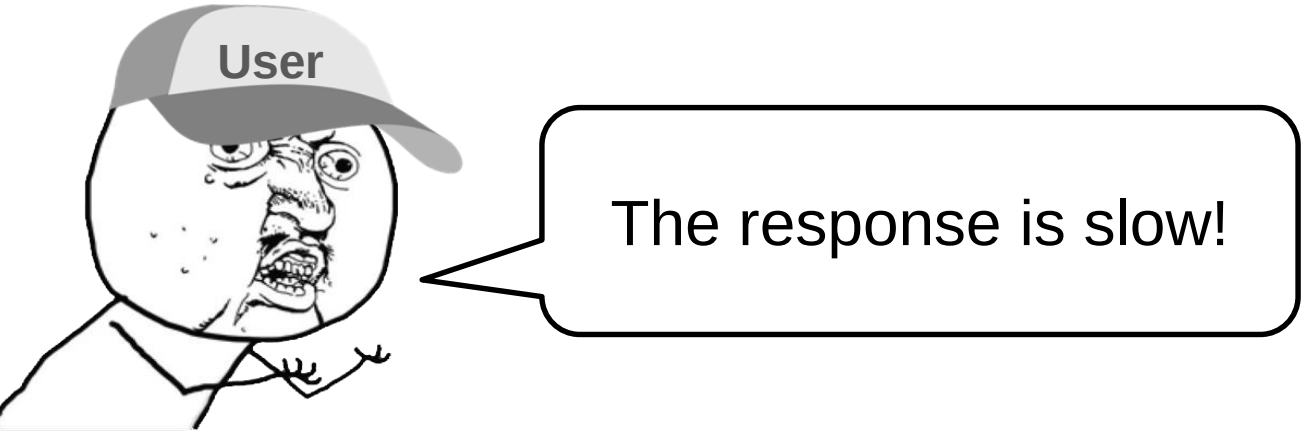
But … overall latency monitoring for tenants is insufficient
e.g., Amazon CloudWatch

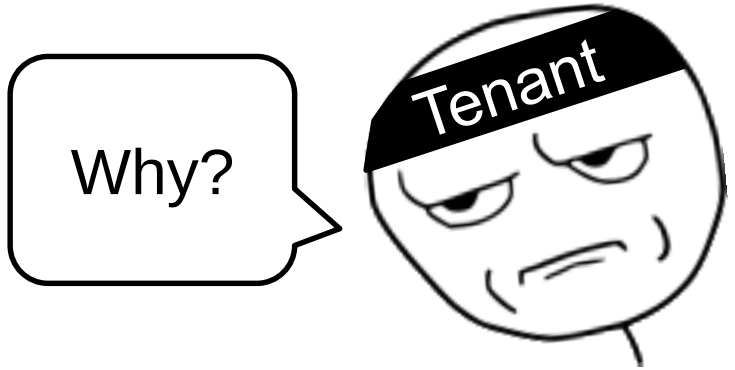


Latency	Measures the time elapsed in seconds after the request leaves the load balancer until the response is received. **Preferred statistic:** average

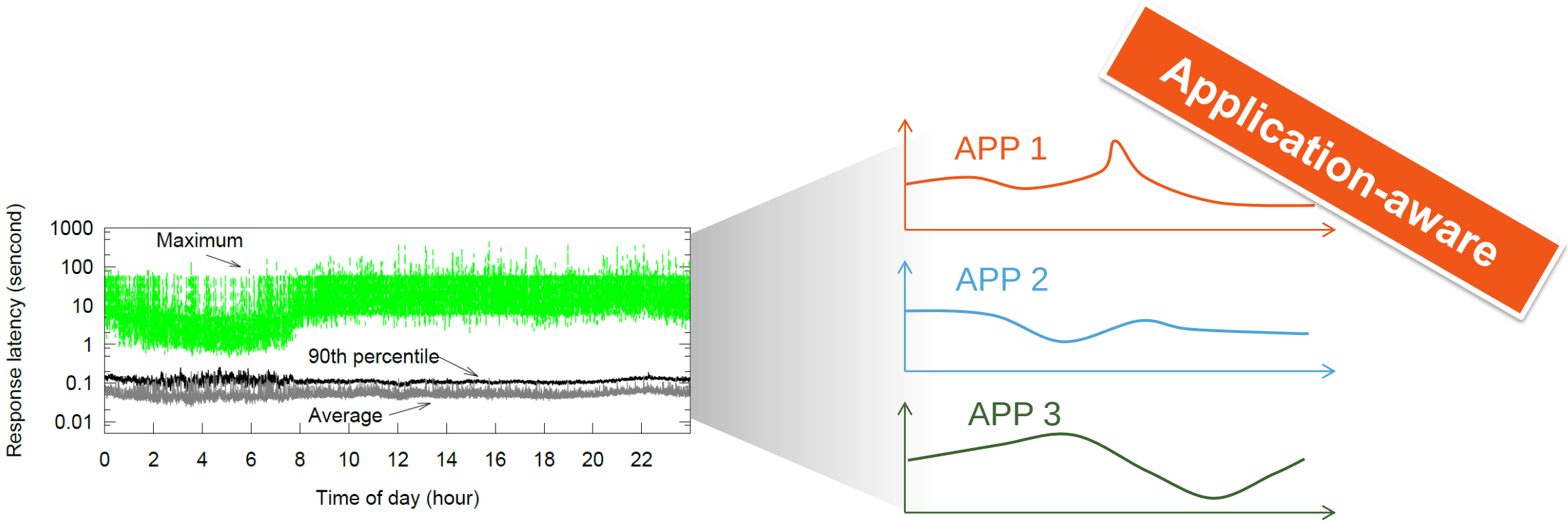
But … overall latency monitoring for tenants is insufficient
e.g., Amazon CloudWatch







Solution: CloudWatch+

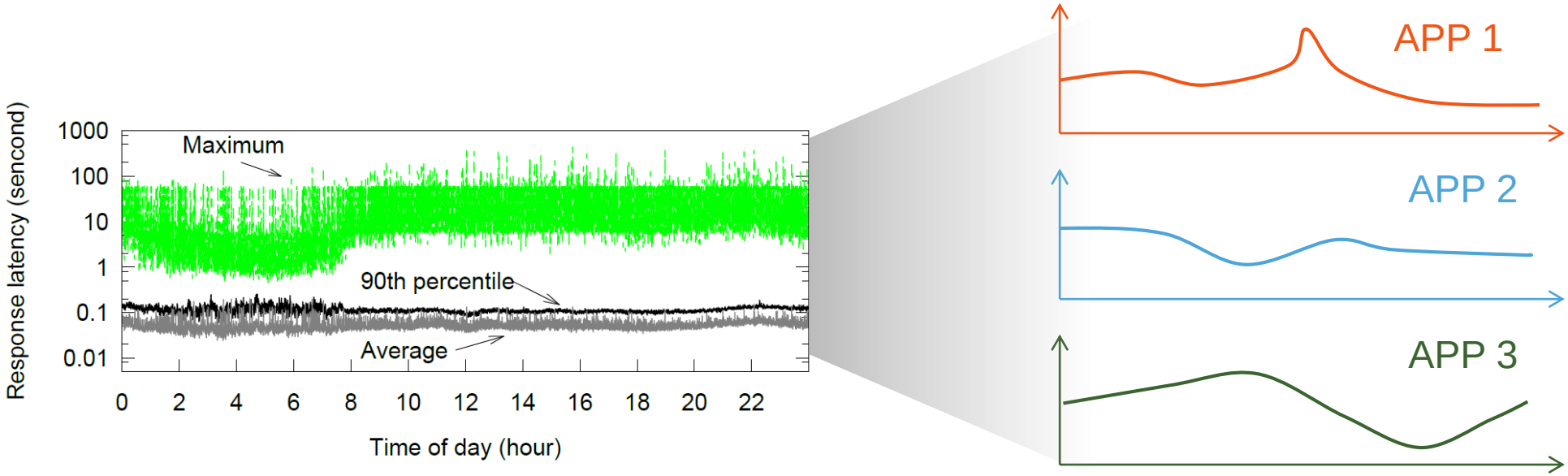


Outline

- Motivation
- Goals and Challenges
- Design
- Evaluation
- Conclusion

Goals

- Identifying web applications via a general indicator (e.g., URL)
- Online and realtime



Challenges

- **Application and parameter fields cannot be distinguished easily**

RFC 3986

https://www.xxx.com/news?title=CNSM
https://www.xxx.com/news?title=Rio

} Application: news

URL Rewrite

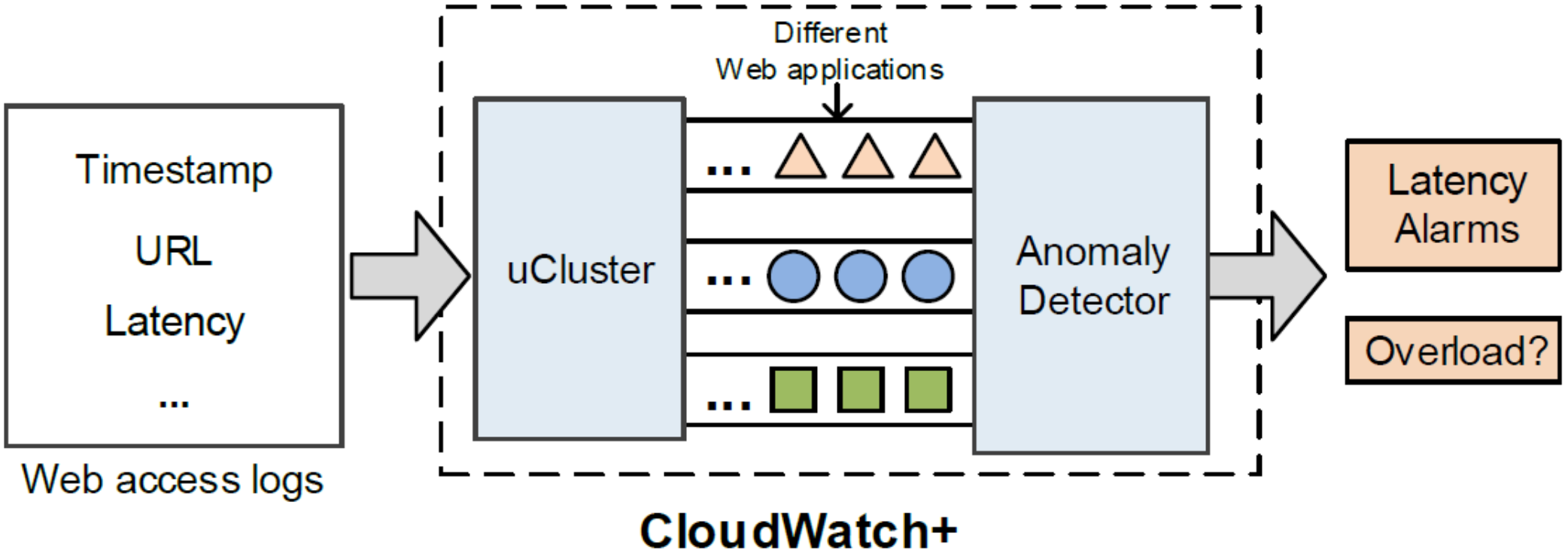
https://www.xxx.com/news/CNSM
https://www.xxx.com/news/Rio

} Application: ??

Outline

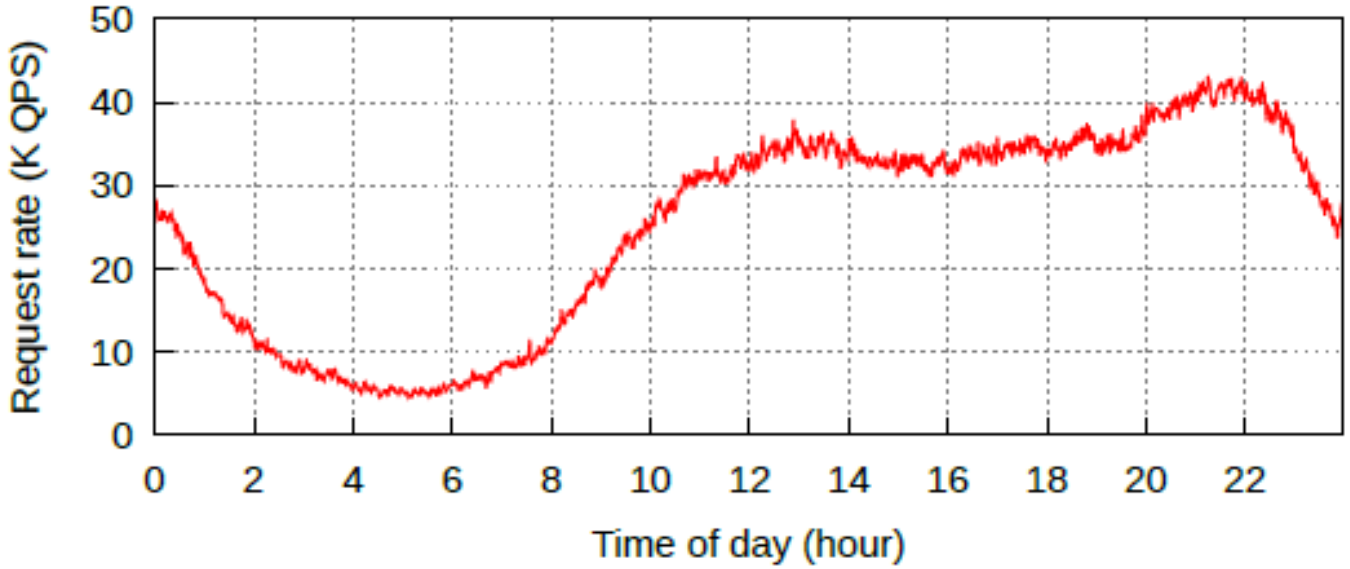
- Motivation
- Goals and Challenges
- Design
- Evaluation
- Conclusion

Architecture

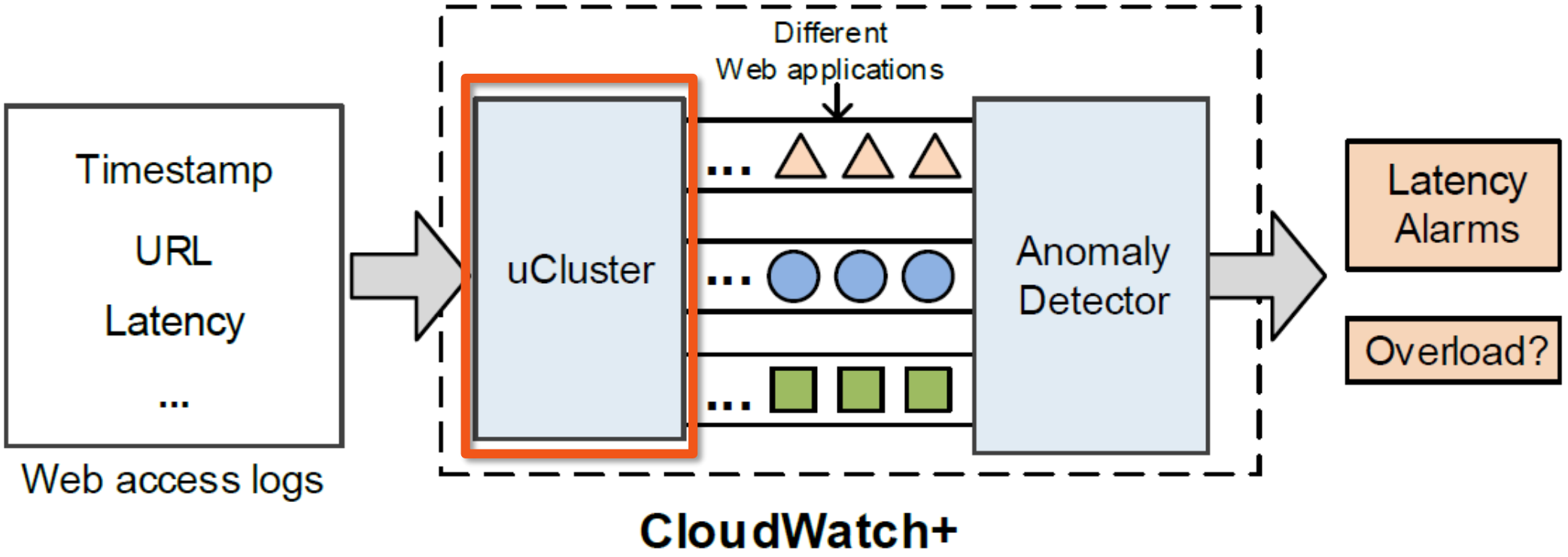


Data set

- One-day access log from a cloud data center
  - More than 200 tenants (we focus on the top 64)
  - 33 million records (after sampling by 2%)
  - 42,000 QPS at peak-hour



Architecture

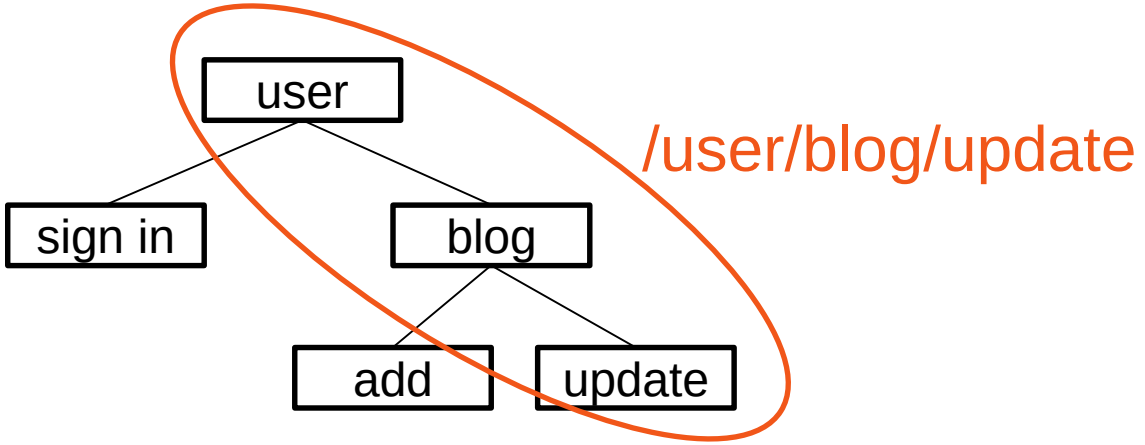


Intuitions

- Intuition 1: parameter fields can generate more different URLs.

/news/*id* → /news/000001, /news/000002, ..., /news/999999

- Intuition 2: URL is hierarchical in nature



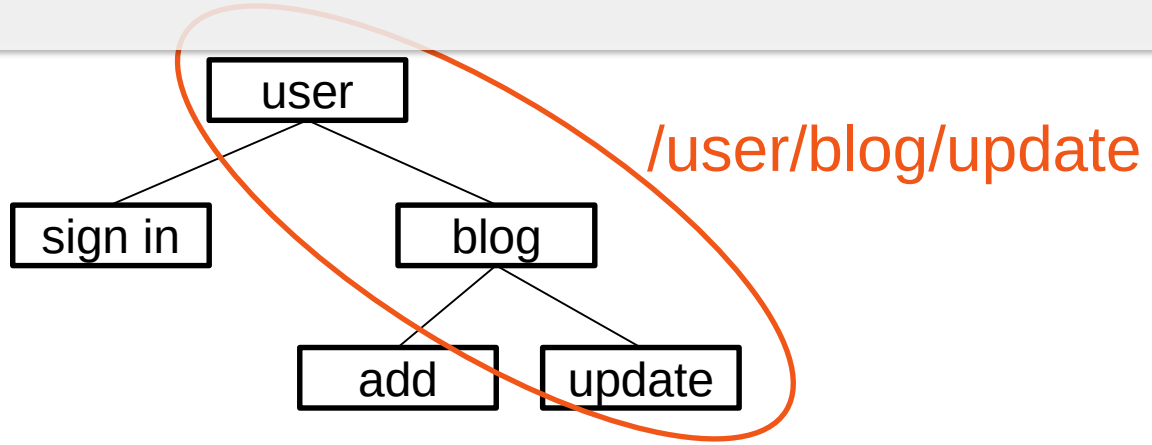
Intuitions

- Intuition 1: parameter fields can generate more different URLs.

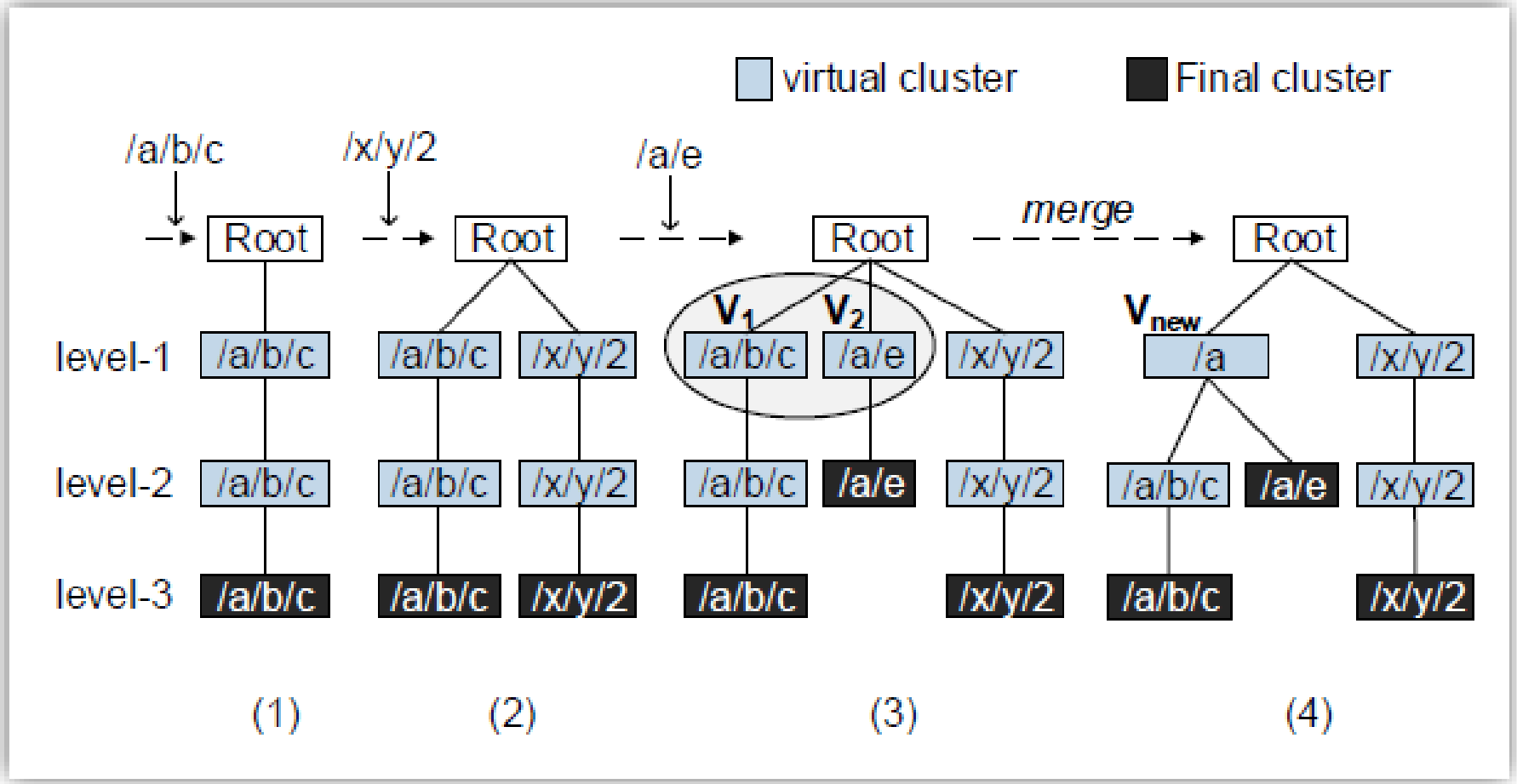
/news/*id* ⇒ /news/2013_05_13_000001
/news/2013_05_13_000002
.....
/news/2013_05_13_999999

Hierarchically Frequent Pattern Mining

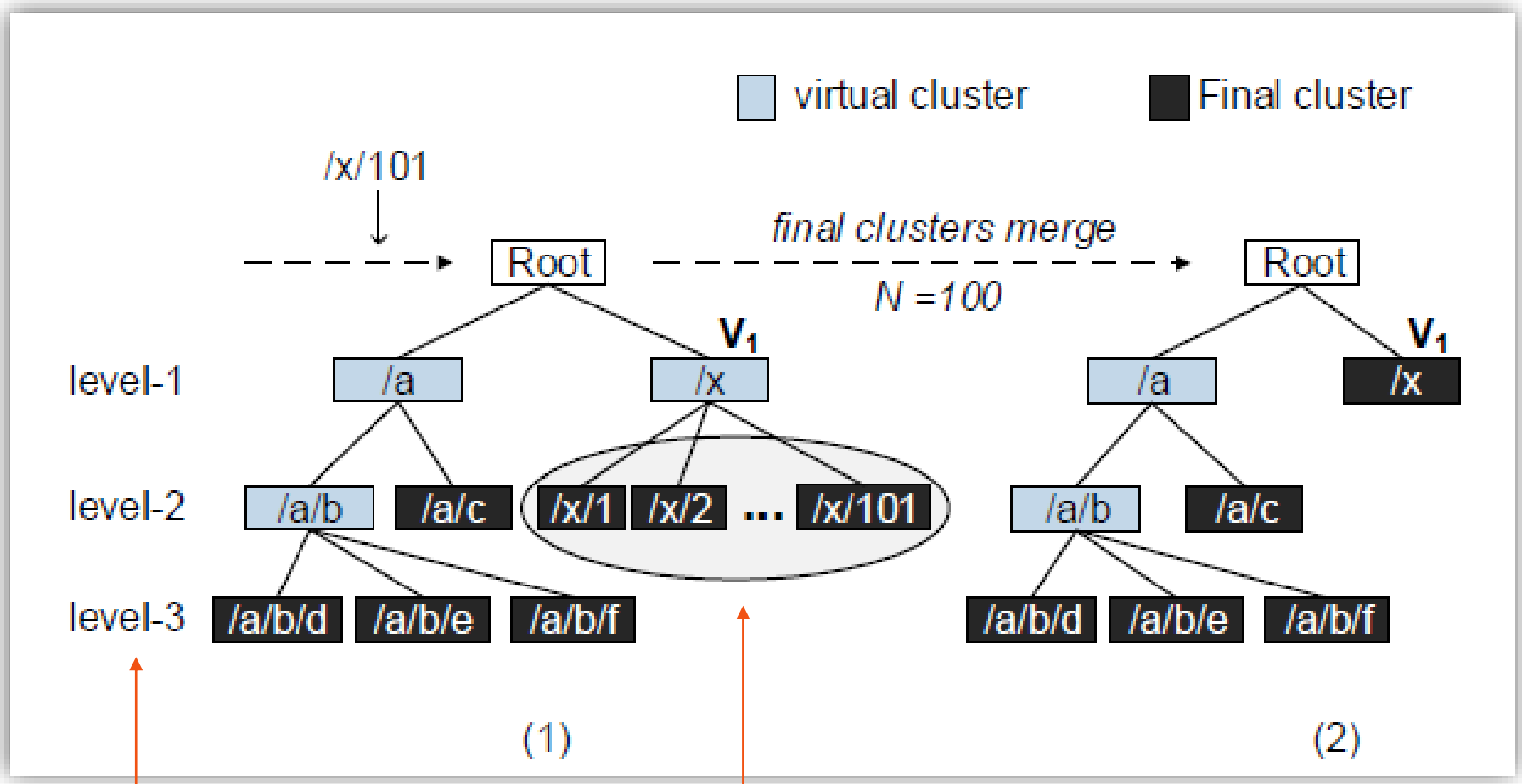
- Intuition 2: URL is hierarchical in nature



uCluster


virtual cluster
Final cluster
/a/b/c
/x/y/2
/a/e
merge
Root
Root
Root
Root
level-1
level-2
level-3
/a/b/c
/x/y/2
/a/e
/a
V1
V2
Vnew
(1)
(2)
(3)
(4)

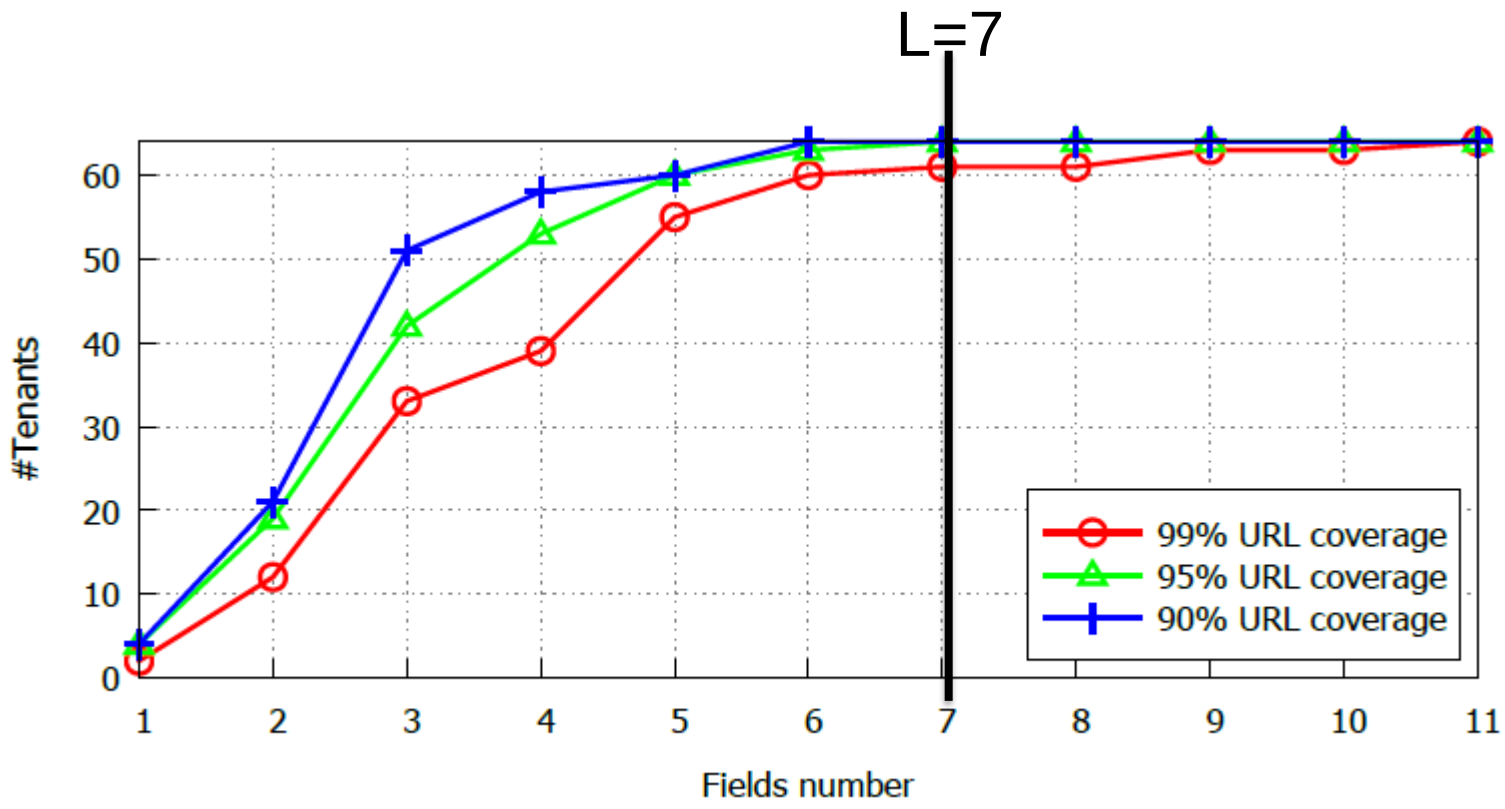
uCluster



Arguments: **L** **N**

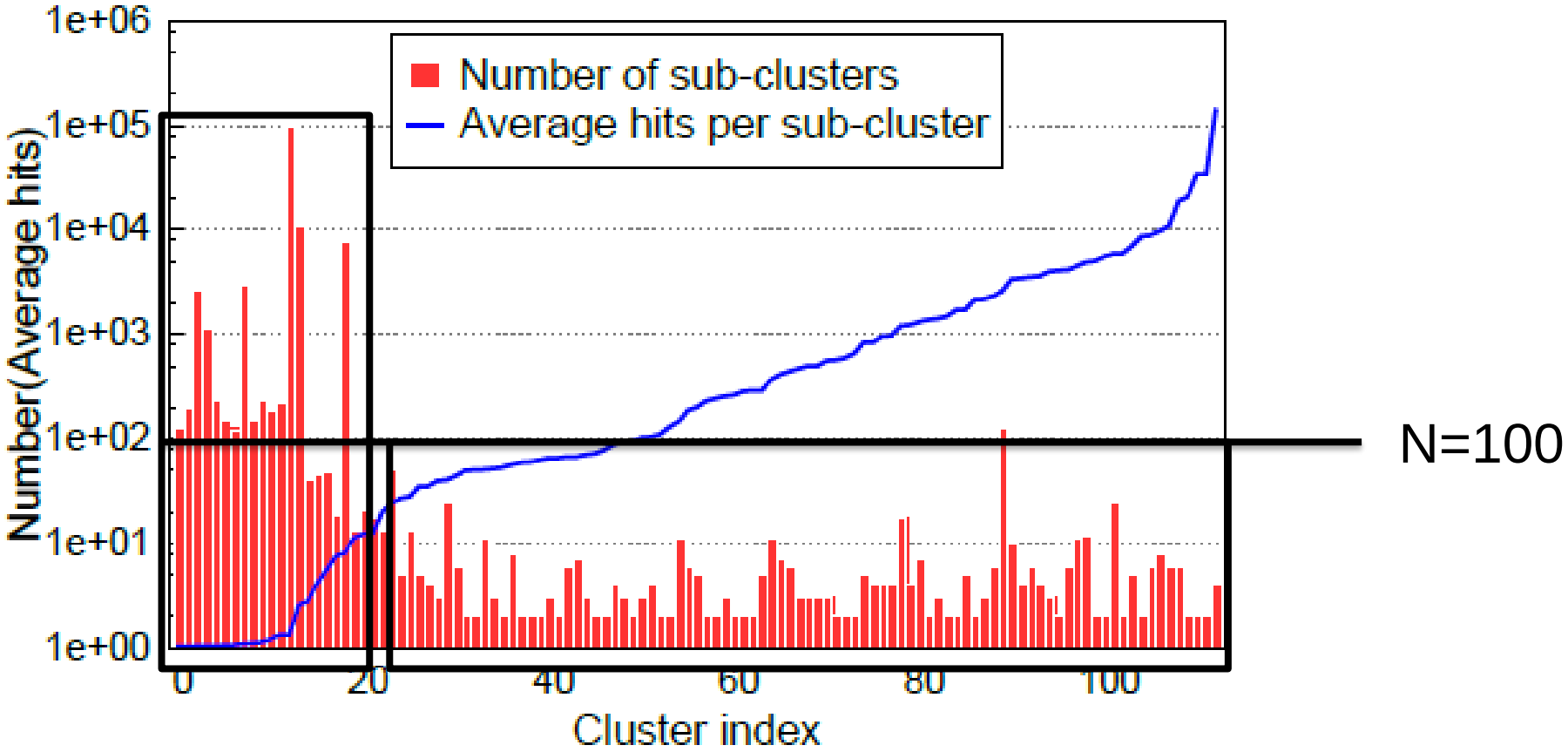
Arguments selection

- L: how many fields are there in URLs?



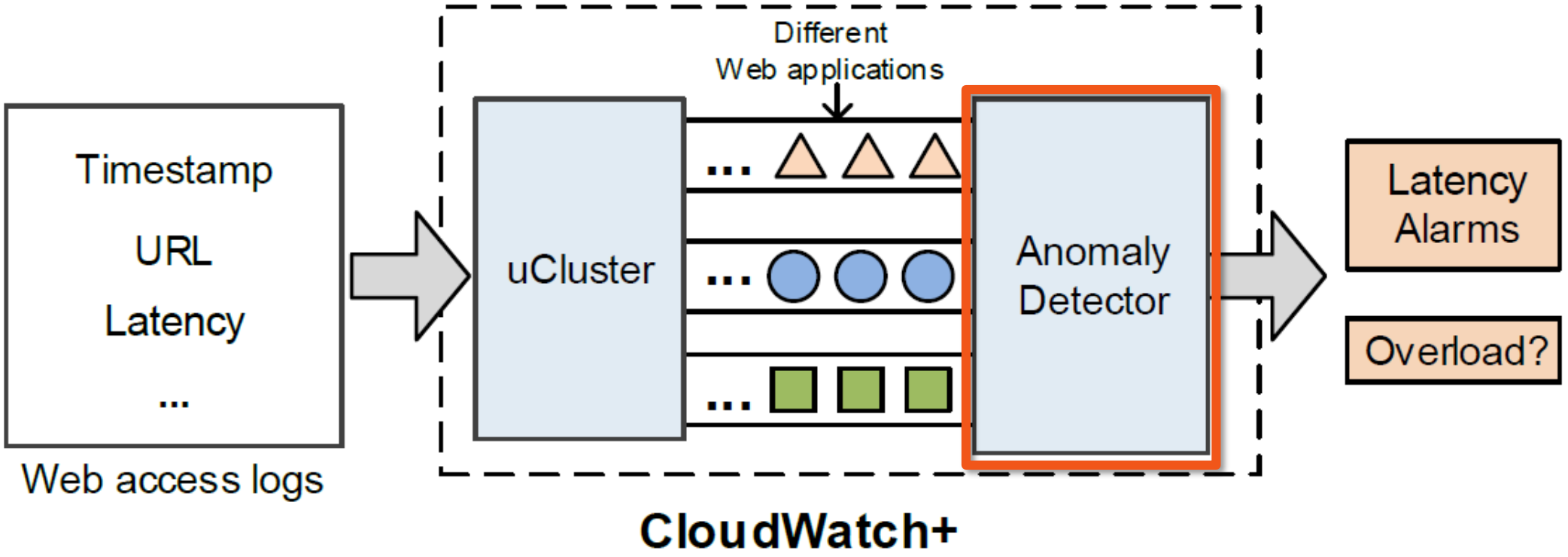
Arguments selection

- N: how many sub-clusters should be treated as a parameter field?

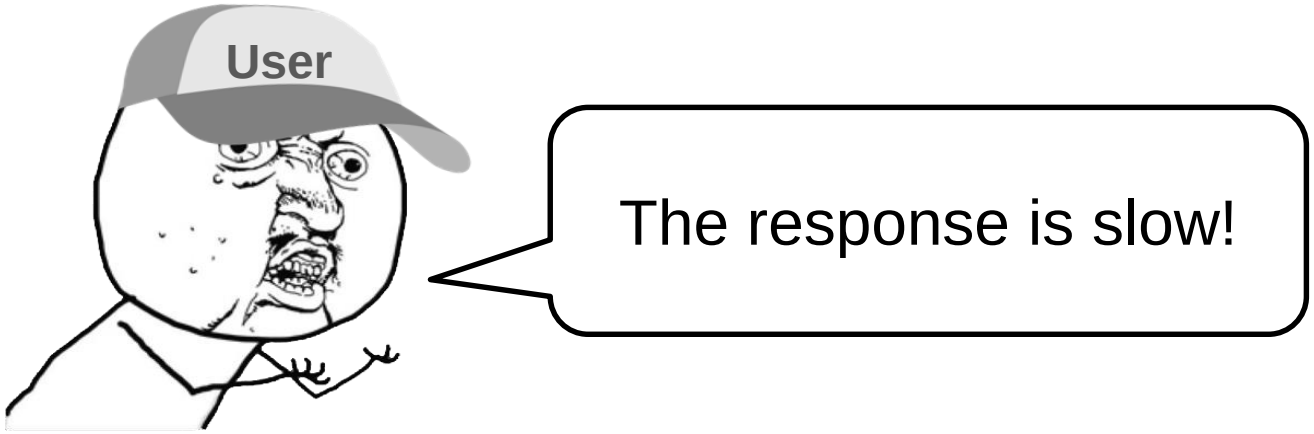


(a) Tenant 1.

Architecture



Anomaly Detection



Code bugs

User's problems

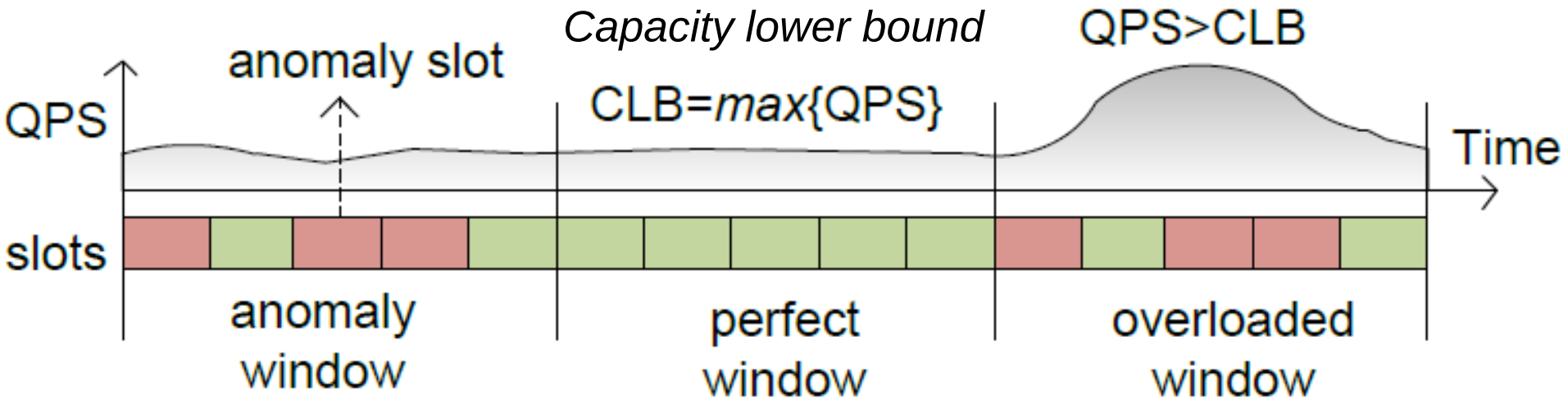
Cloud failures

Overload

Useful information for deciding elastic scaling

Anomaly Detection

- For any application



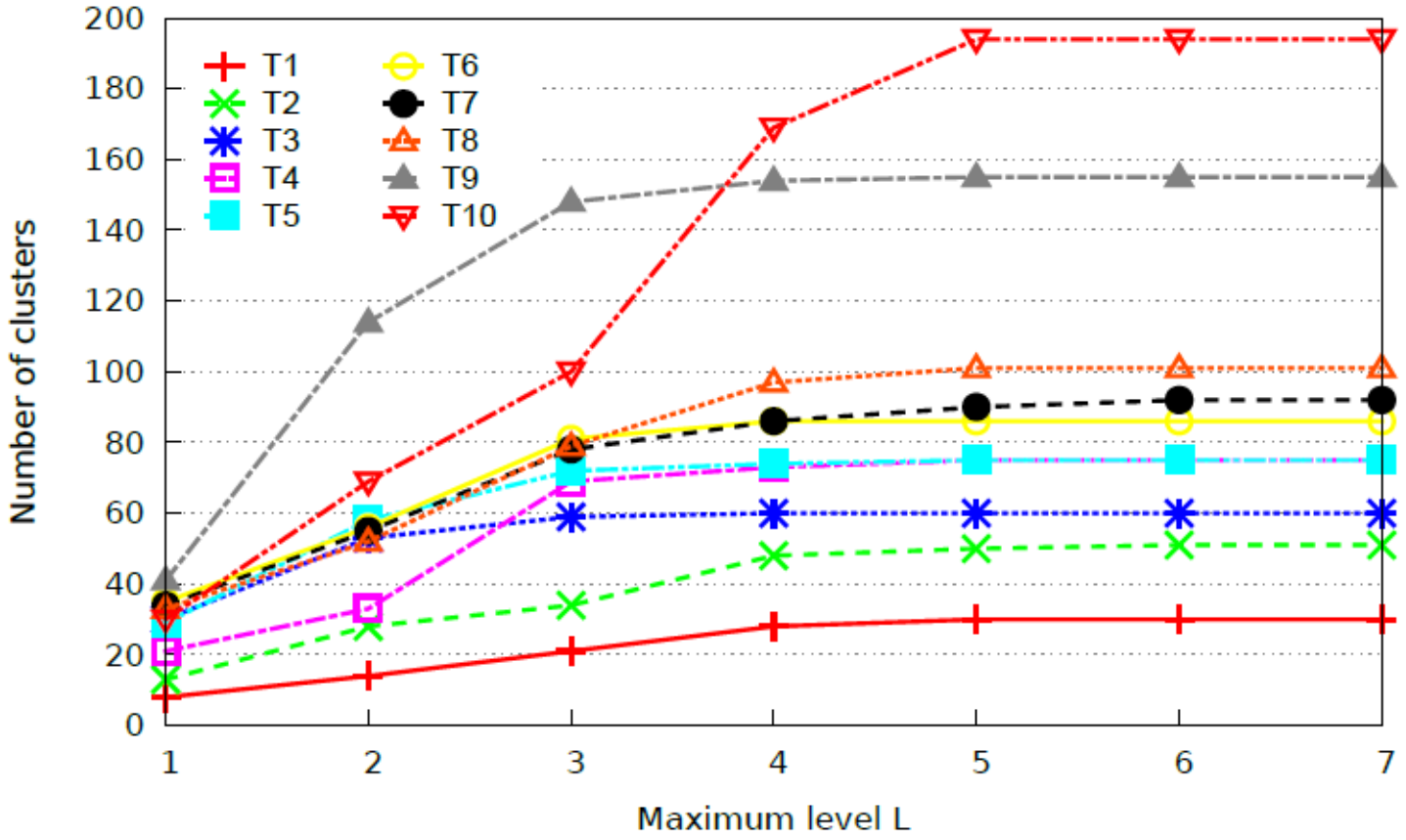
Detection window length W=5
Alarm threshold n=3

Outline

- Motivation
- Goals and Challenges
- Design
- Evaluation
- Conclusion

Clustering Results

- For the top 10 most visited tenants

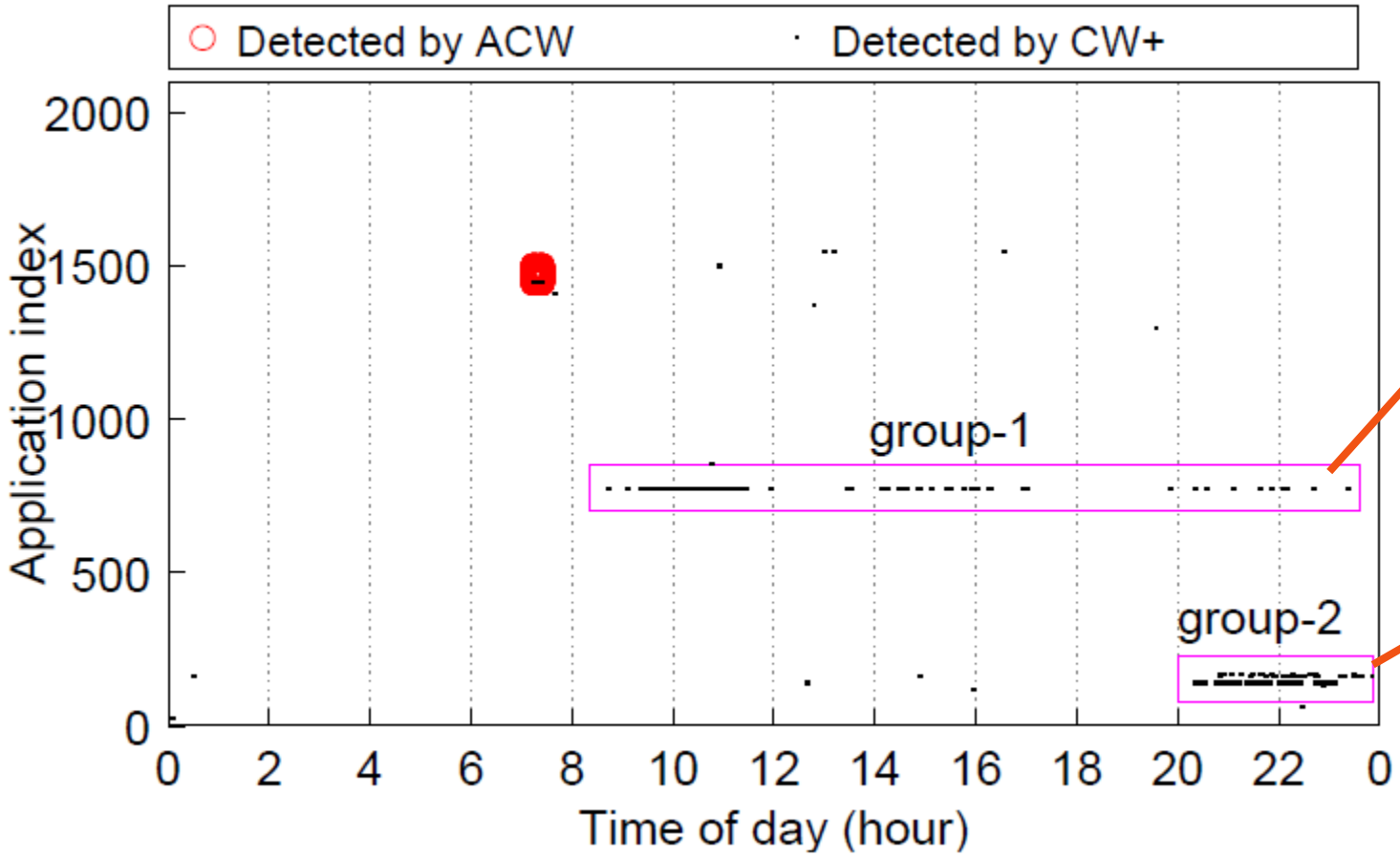


160,000- 6,000,000 Unique URLs

10-200 Clusters (applications)

Detection Results

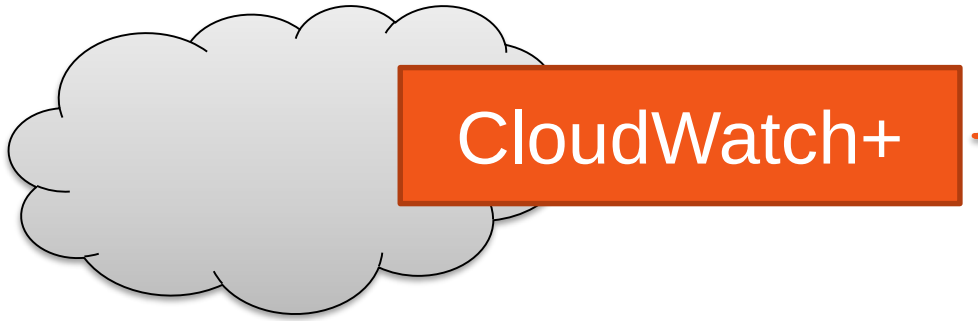
- For the top 64 most visited tenants



"upload" application
of a file sharing service

"sign in" application
of an online social network

Conclusion



- Automatically learn applications
- Monitor applications latency separately
- Suggest whether a latency anomaly is caused by overload

Thank you
Q&A

Backup

Performance and Overhead

- Runtime for clustering and detecting one day records (33 million)
  - 875 seconds
- The number of virtual and final clusters
  - 10,000