

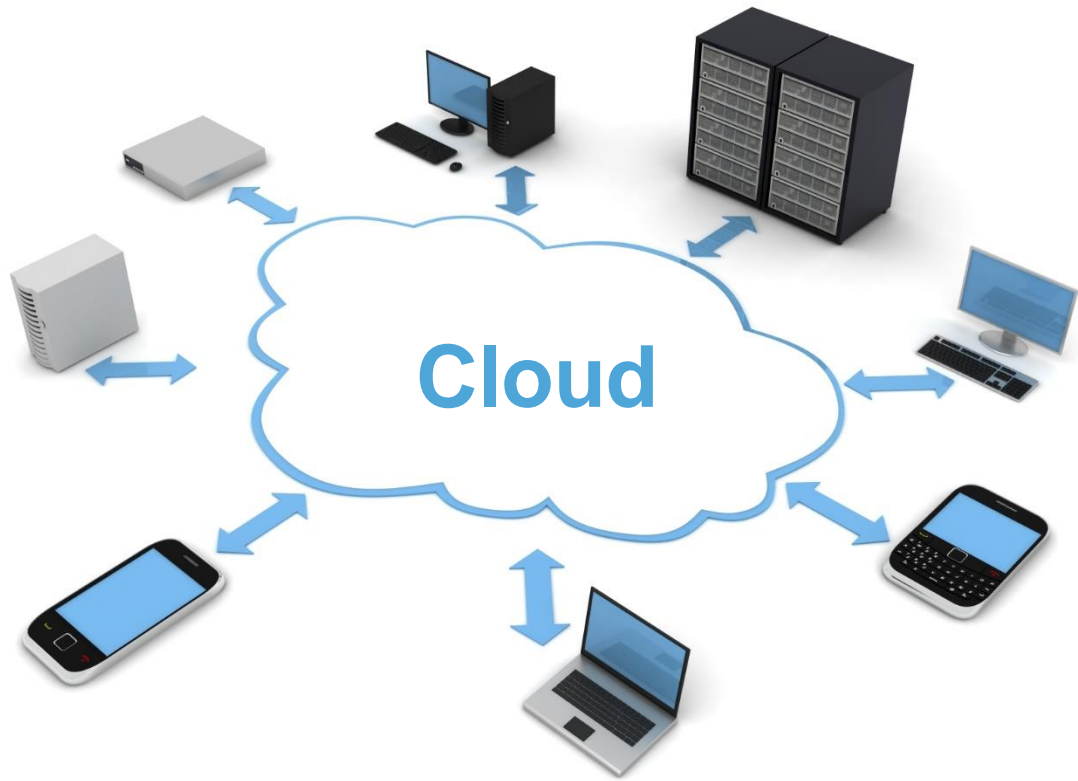
Application-Aware Latency Monitoring for Cloud Tenants via CloudWatch+



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

The screenshot shows the 'Your CloudWatch Alarms' interface. At the top, there are buttons for 'Create Alarm', 'Modify', and 'Delete'. Below that, a 'Viewing:' dropdown is set to 'All alarms'. A table lists three alarms:

	State	Name	Threshold
<input type="checkbox"/>	ALARM	Fleet CPU	CPUUtilization is < 20 for 15 minutes
<input type="checkbox"/>	ALARM	DiskWriteOpsForMicros	DiskWriteOps is < 10 for 10 minutes
<input checked="" type="checkbox"/>	OK	DiskWriteOps for instance	DiskWriteOps is >= 10 for 30 minutes

Below the table, the '1 Alarm selected' section shows details for the 'Alarm: DiskWriteOps for instance'.

Description | **Metric**

State Details: State changed to 'OK' at 2011/01/13 13:58 UTC. Reason: Threshold Crossed: 1 datapoint (0.0) was not greater than or equal to the threshold (10).

Description: Disk Write Ops is high

Threshold: DiskWriteOps is >= 10 for 30 minutes

Actions: in ALARM state - Send message to topic "Briande" (briande@amazon.com)
in INSUFFICIENT_DATA state -

DiskWriteOps for instance
DiskWriteOps >= 10

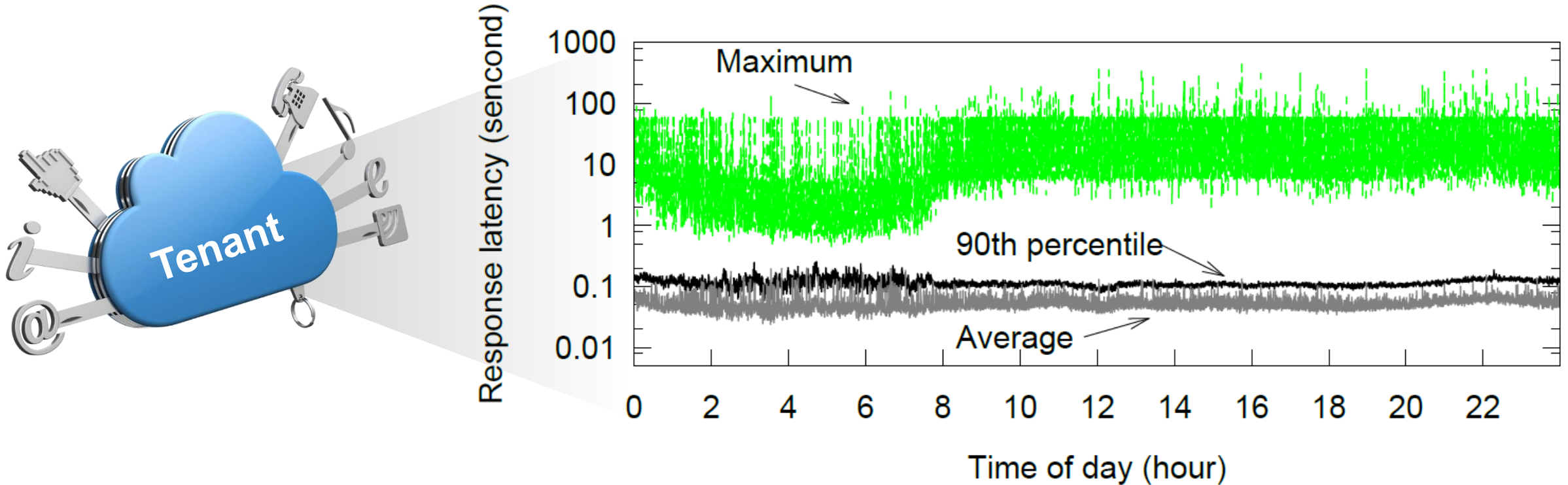
The chart shows a horizontal red line at 10.0 on a scale from 2.5 to 10.0, indicating the threshold.

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

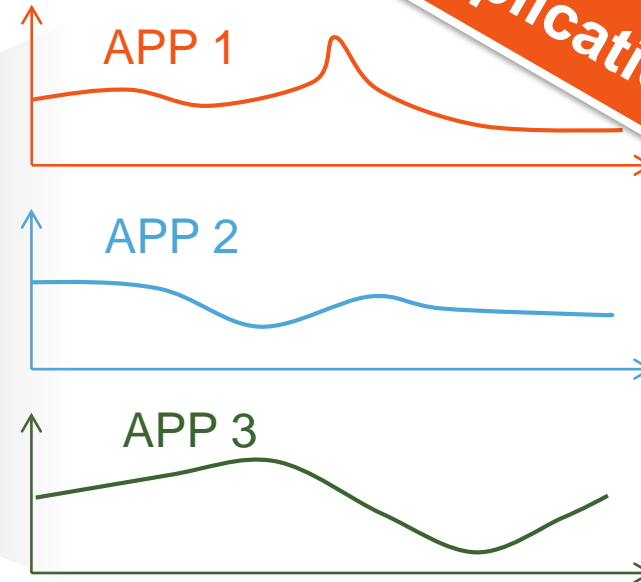
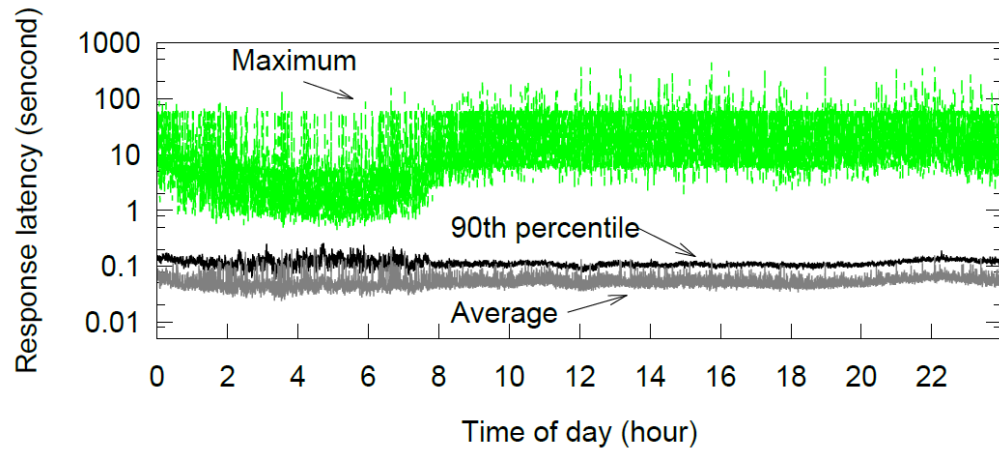


The response is slow!

Why?



Solution: CloudWatch+

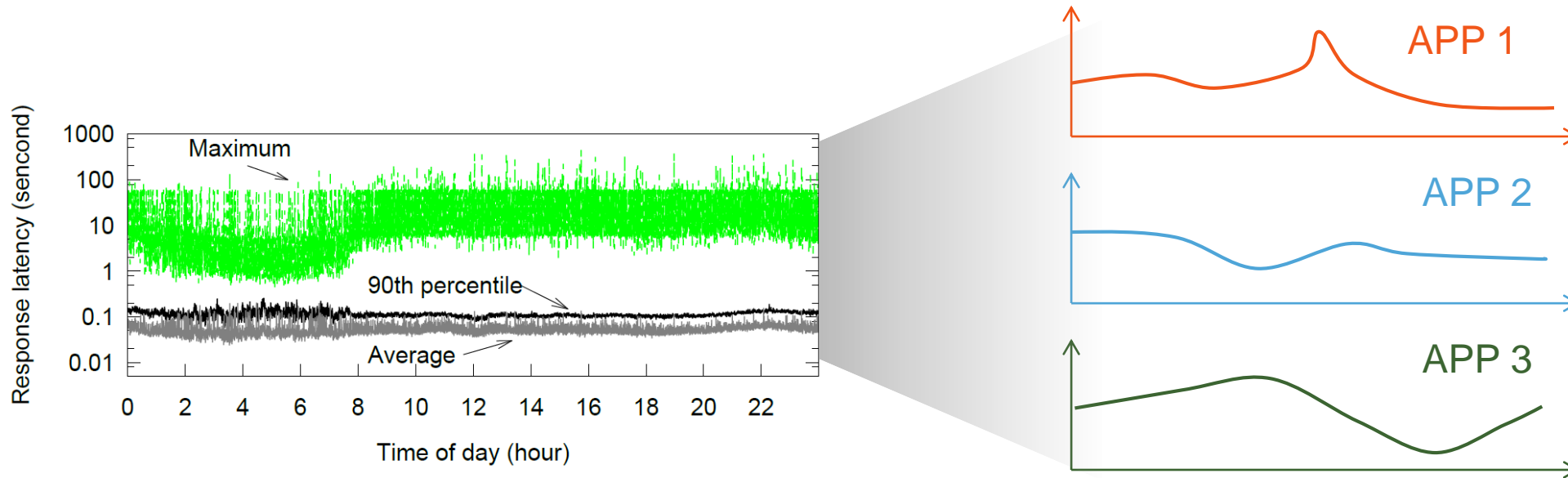


Application-aware

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

Goals

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



- 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>

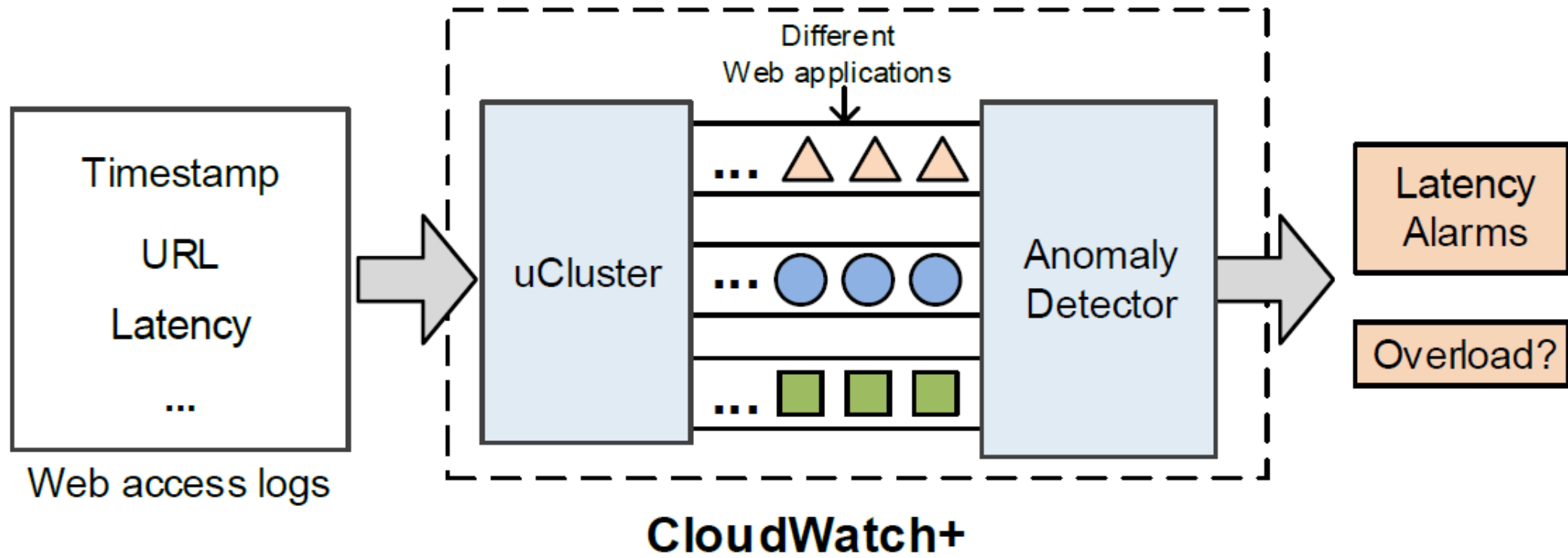
<https://www.xxx.com/news/Rio>

} Application: ??

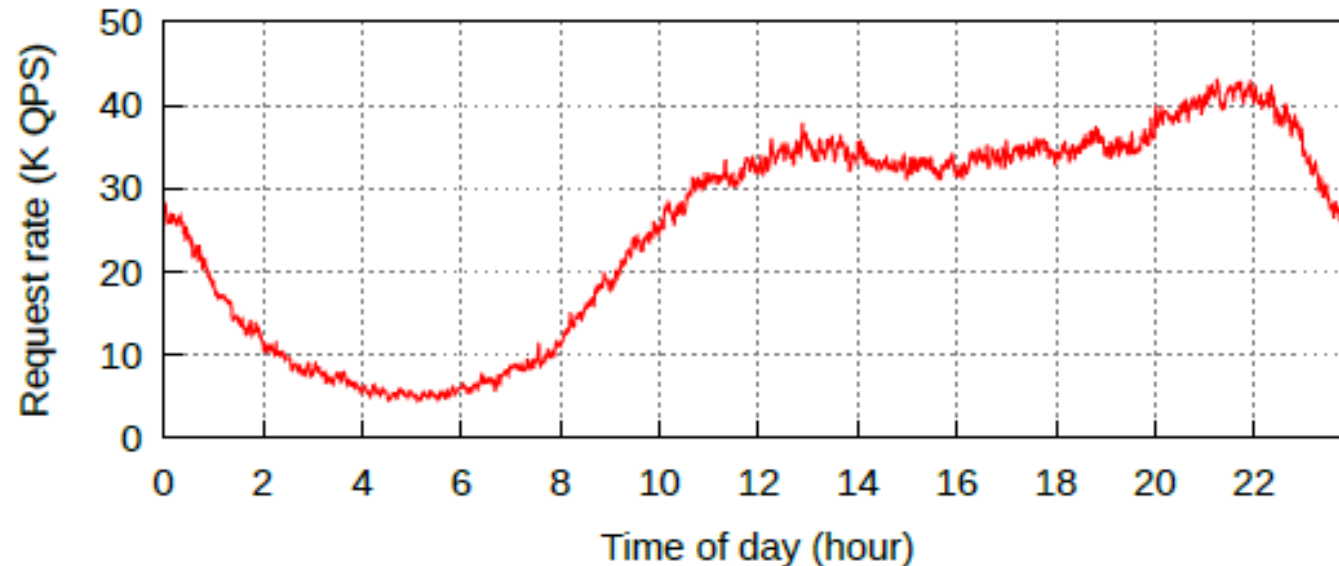
Outline

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

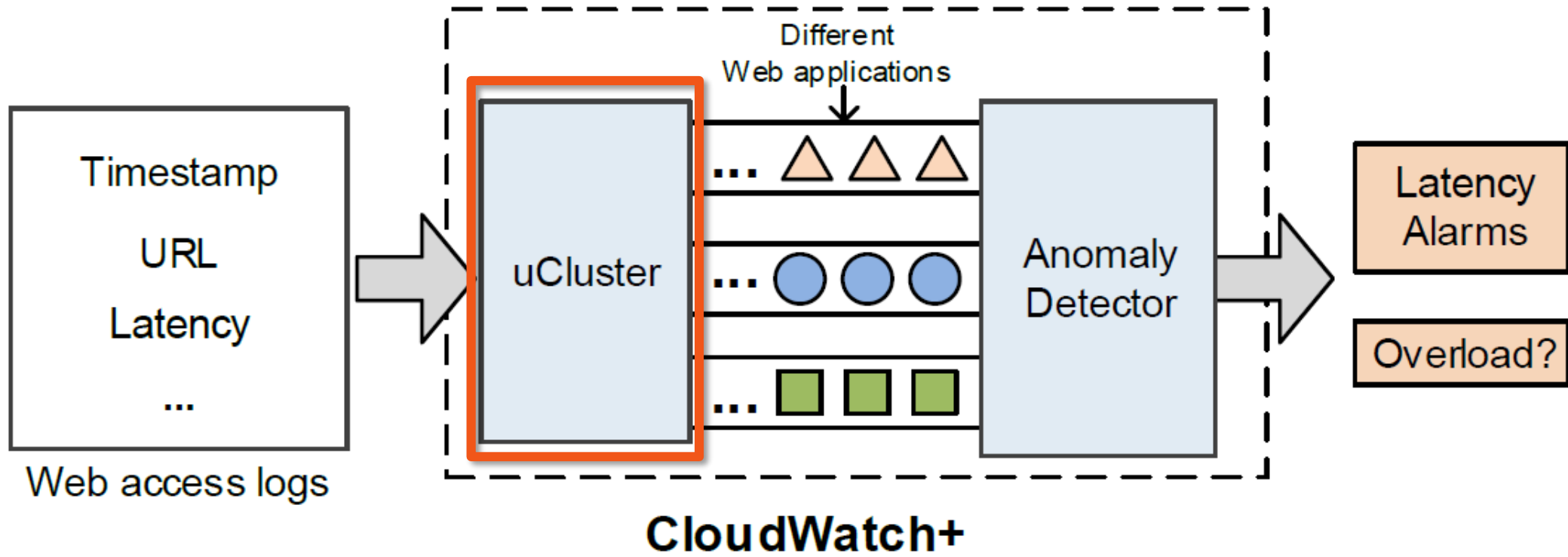
Architecture



- 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

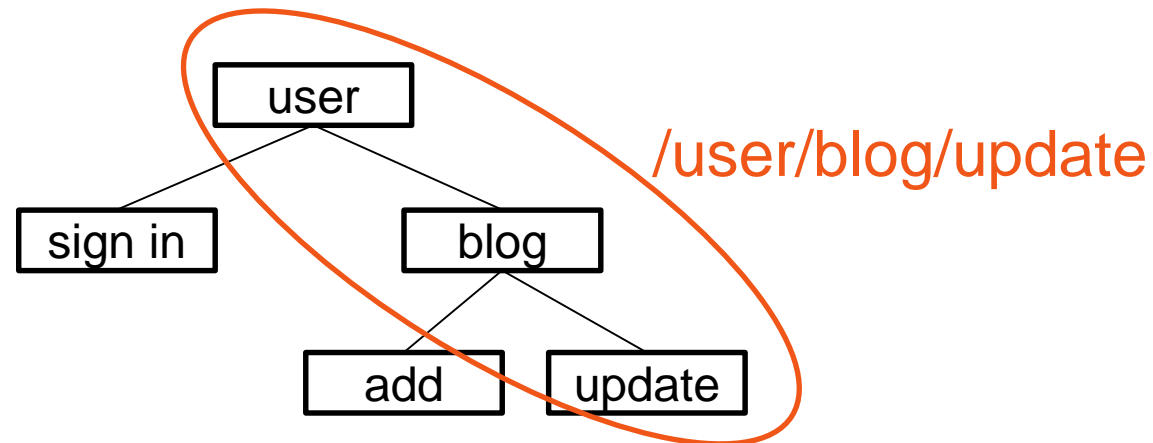


Intuitions

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

`/news/id` → `/news/000001`
`/news/000002`
.....
`/news/999999`

- **Intuition 2:** URL is hierarchical in nature

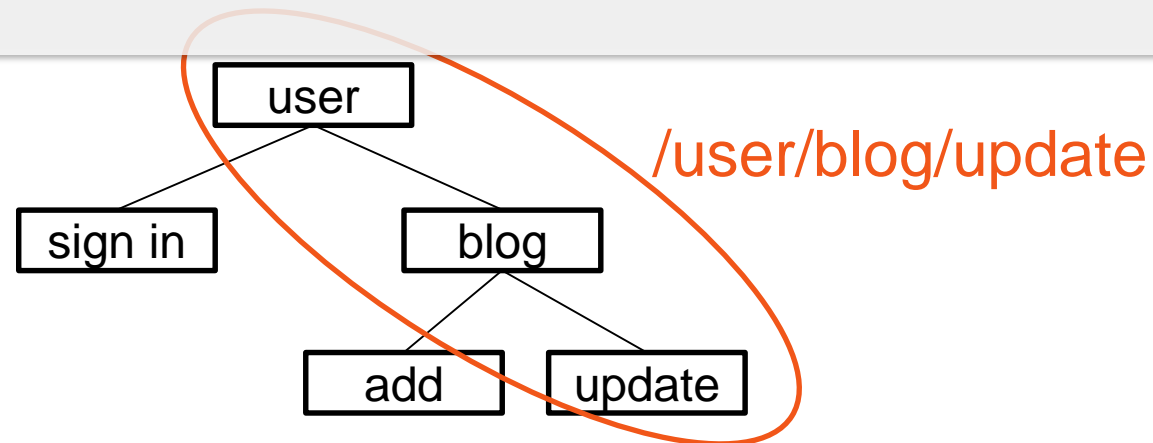


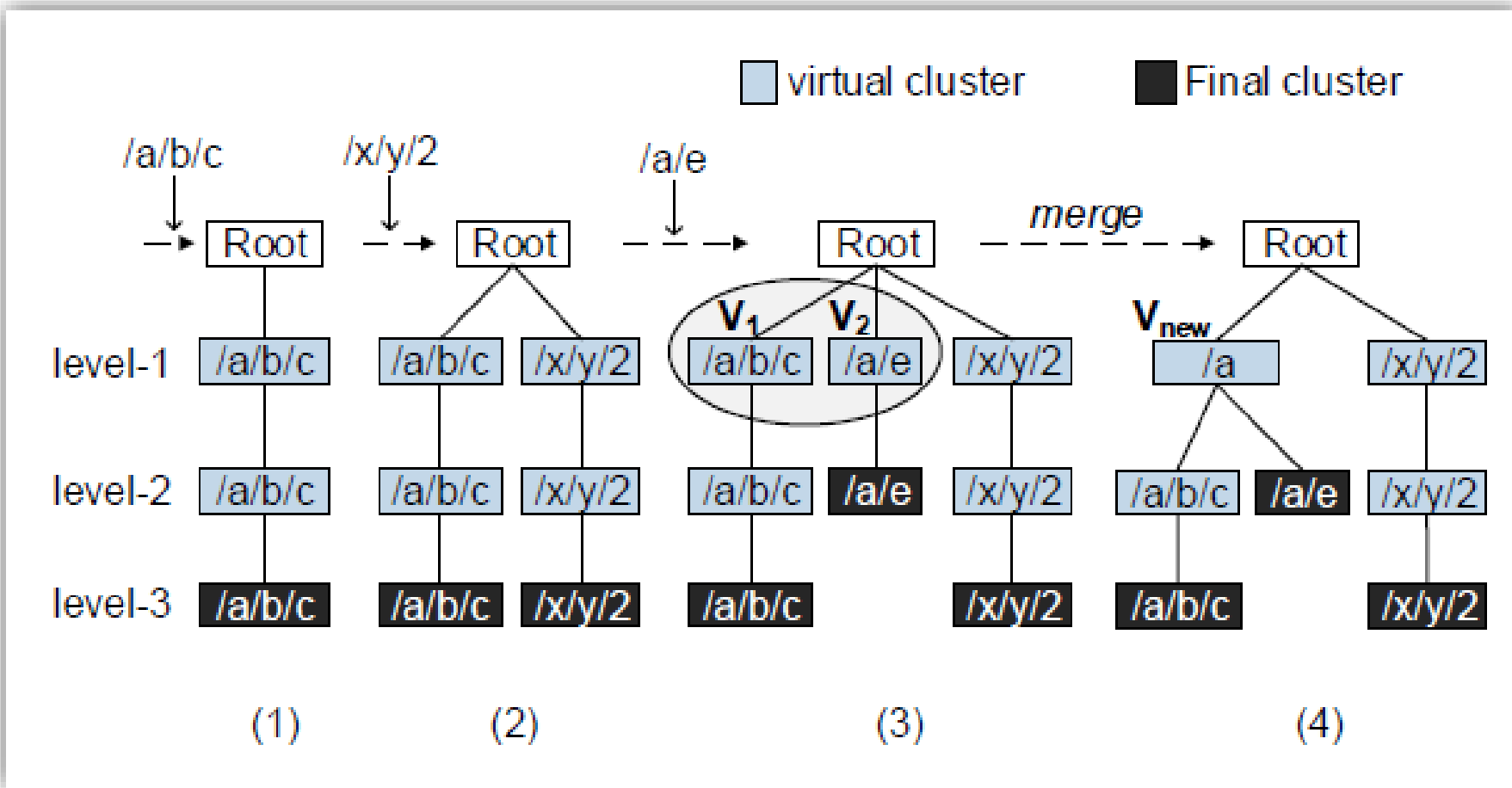
- **Intuition 1:** parameter fields can generate more different URLs.

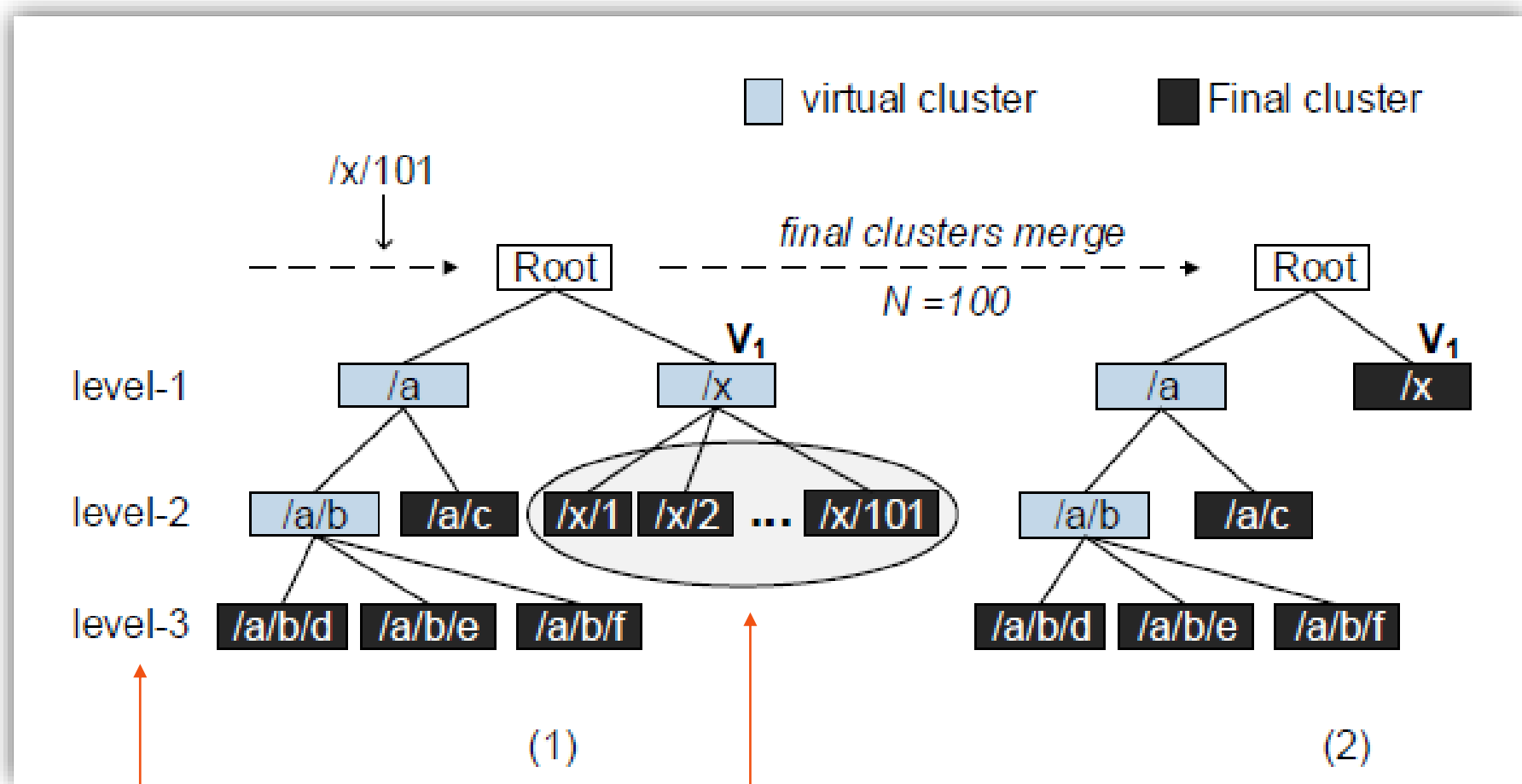
`/news/id` → `/news/2013_05_13_000001`
`/news/2013_05_13_000002`
.....
`/news/2013_05_13_999999`

Hierarchically Frequent Pattern Mining

- **Intuition 2:** URL is hierarchical in nature





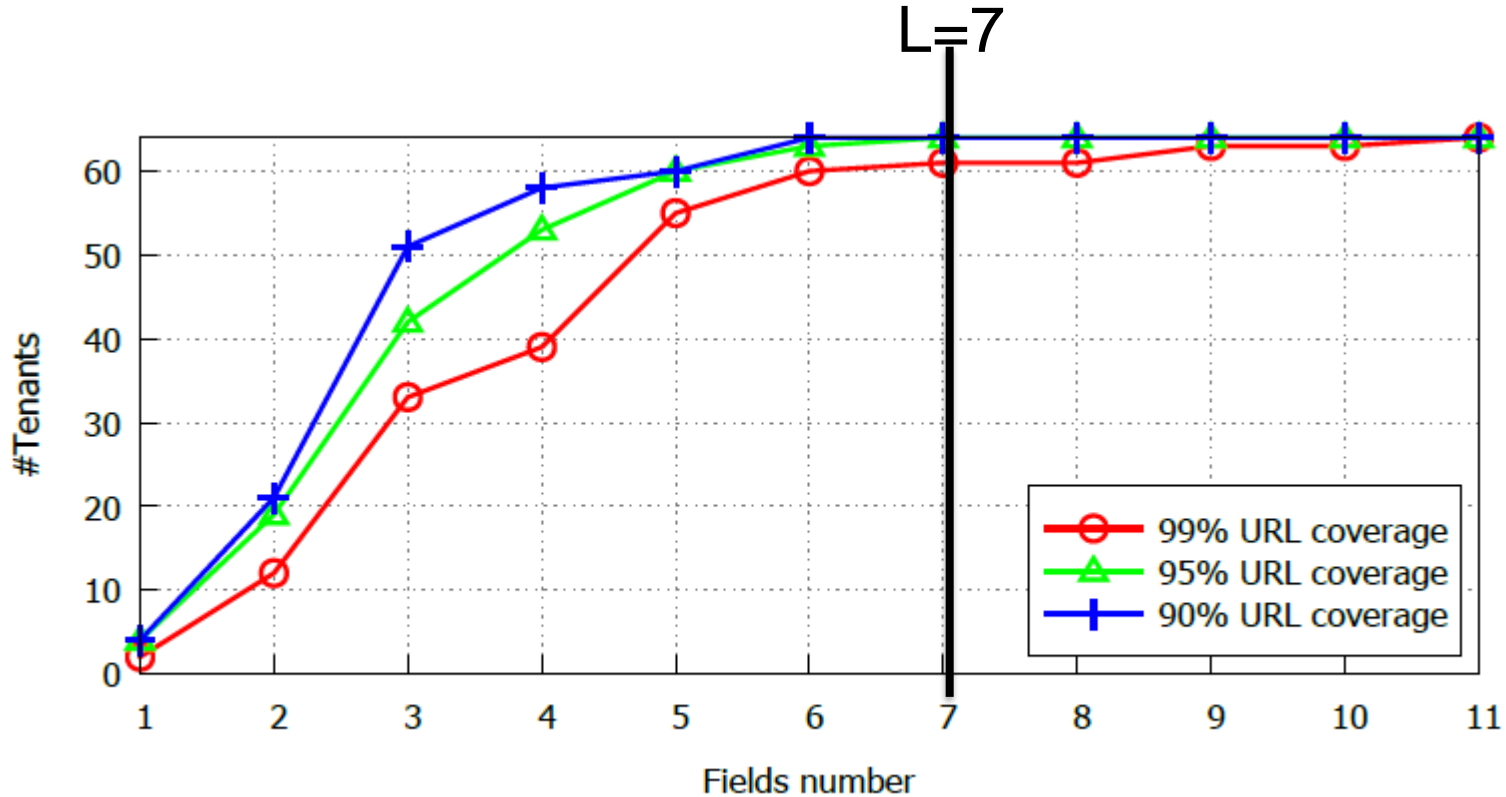


Arguments: L

N

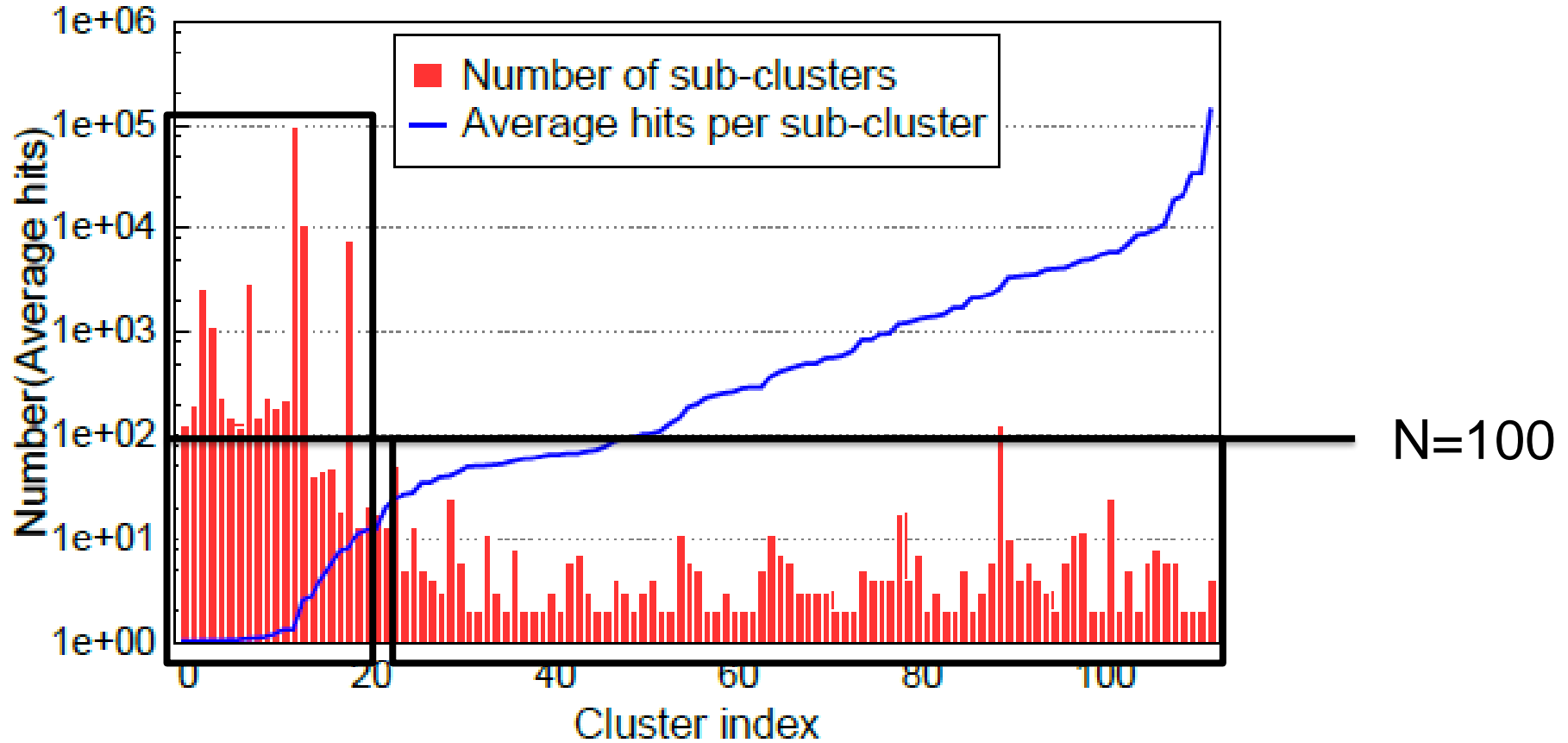
Arguments selection

- L: how many fields are there in URLs?



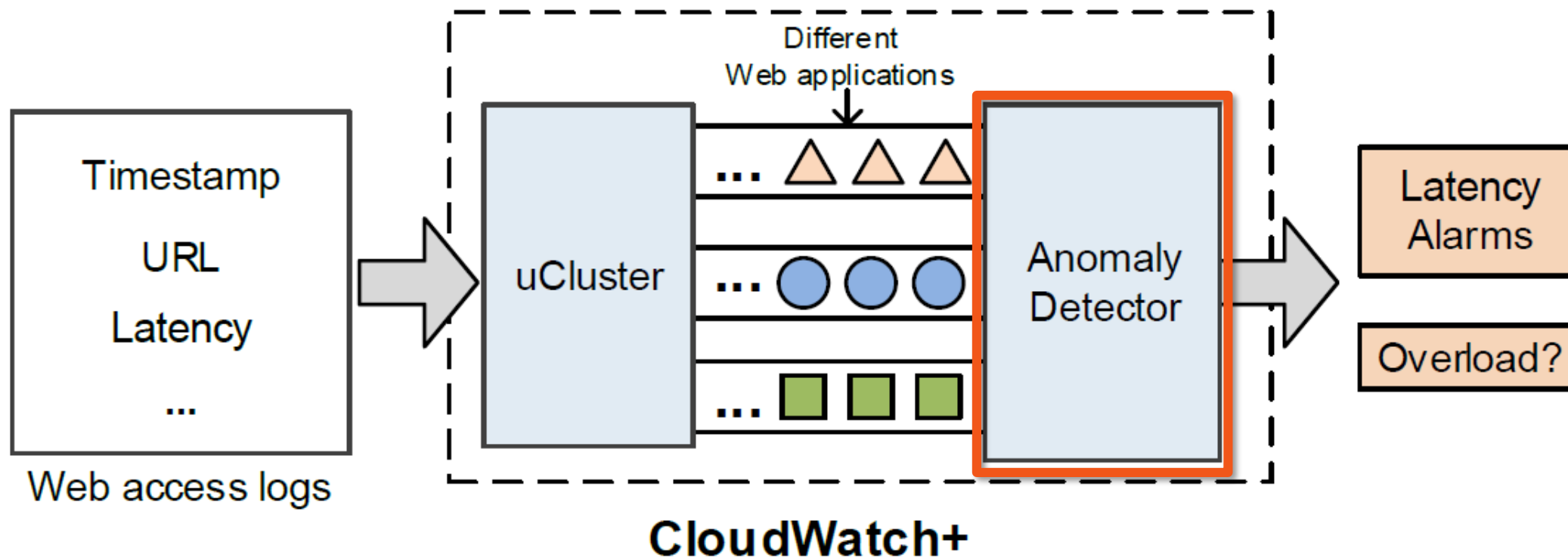
Arguments selection

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



(a) Tenant 1.

Architecture



Anomaly Detection



The response is slow!

Code bugs

User's problems

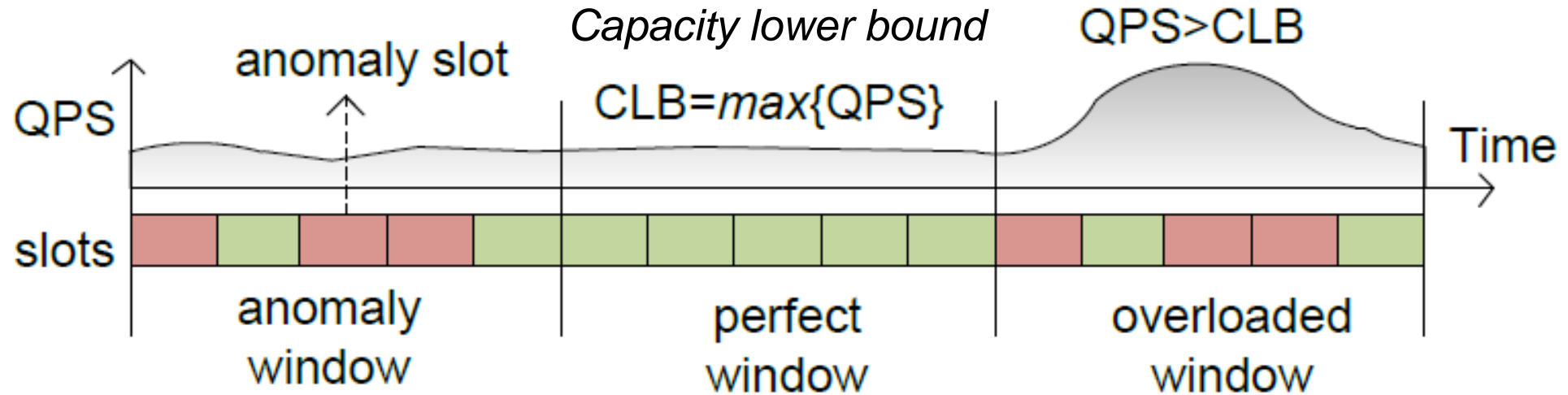
Cloud failures

Overload

Useful information for
deciding elastic scaling

Anomaly Detection

- For any application



Detection window length $W=5$

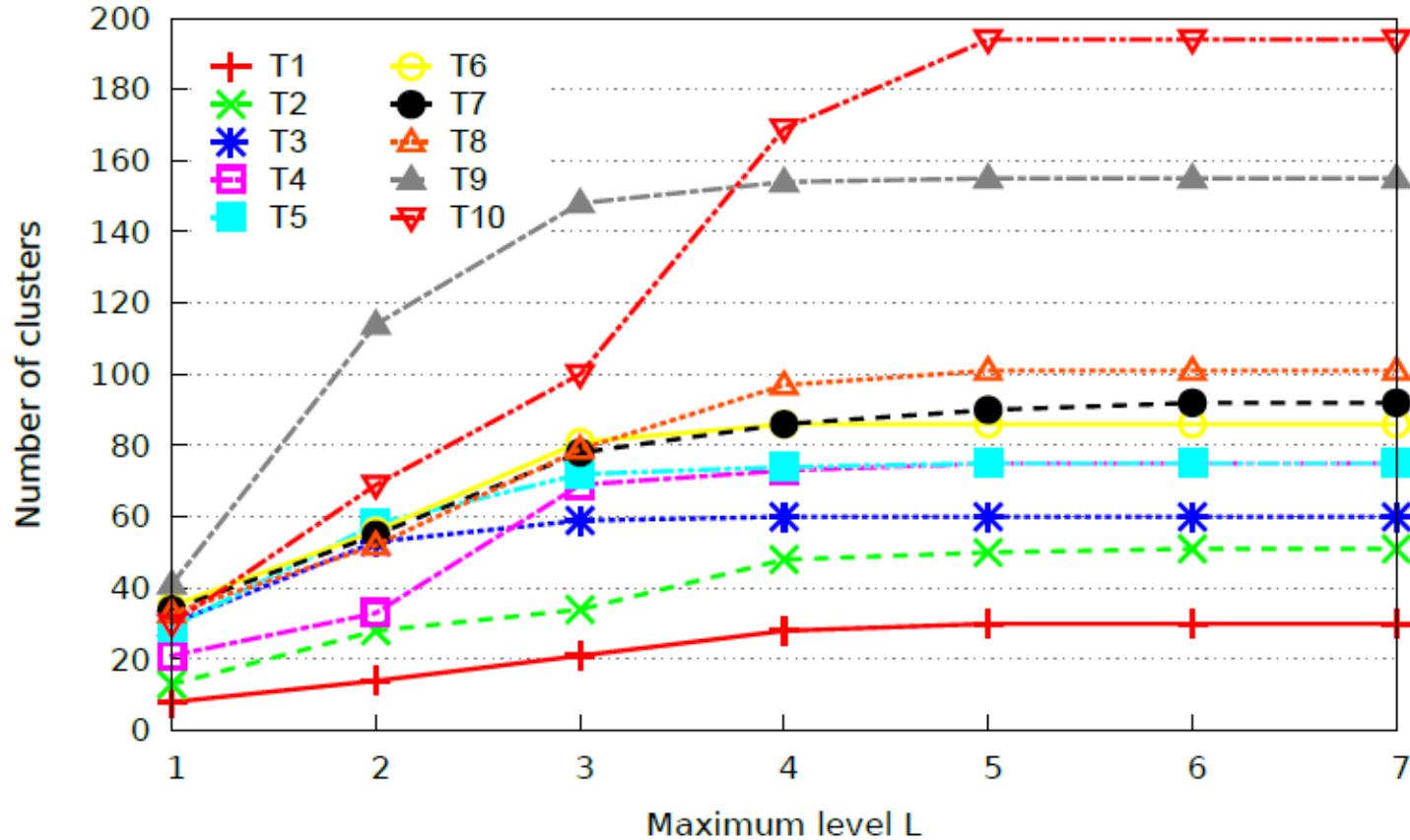
Alarm threshold $n=3$

Outline

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

Clustering Results

- For the top 10 most visited tenants



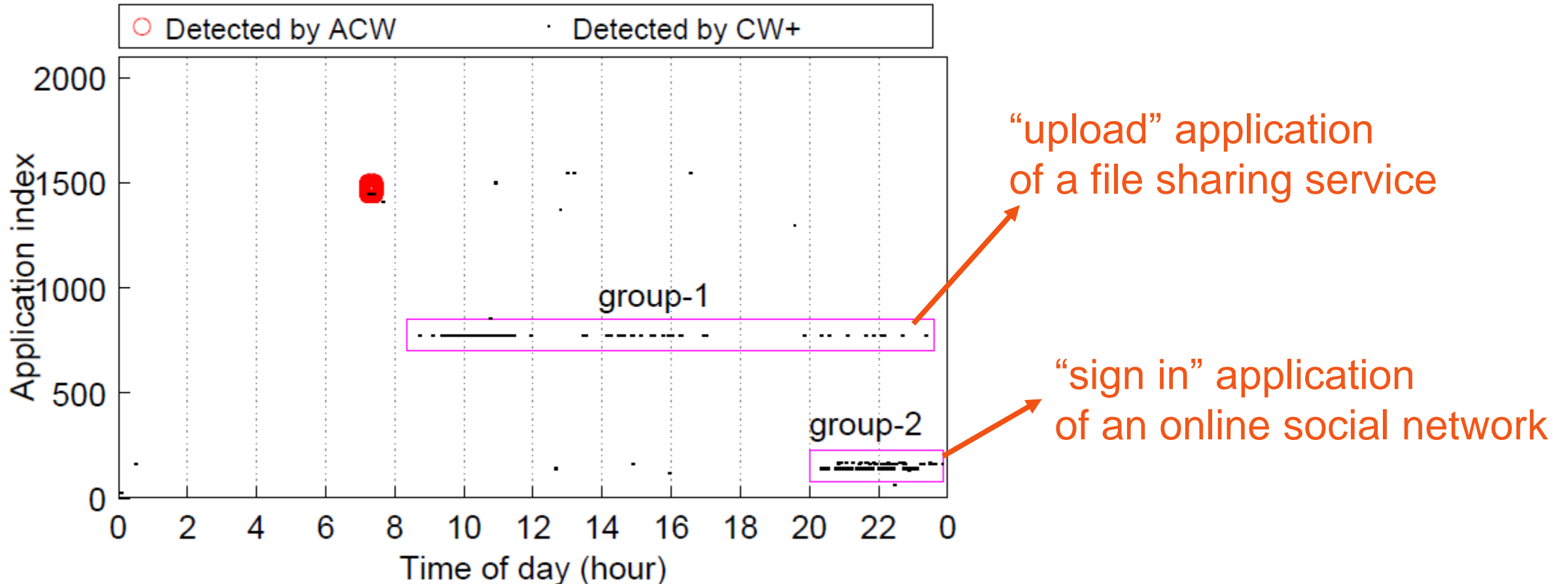
160,000- 6,000,000 Unique URLs



10-200 Clusters (applications)

Detection Results

- For the top 64 most visited tenants



The logo for CloudWatch+ features a grey, stylized cloud on the left. An orange rectangular box is positioned horizontally across the middle of the cloud, containing the text "CloudWatch+" in white, sans-serif font.

CloudWatch+

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

