Predicting service metrics for cluster-based services using real-time analytics

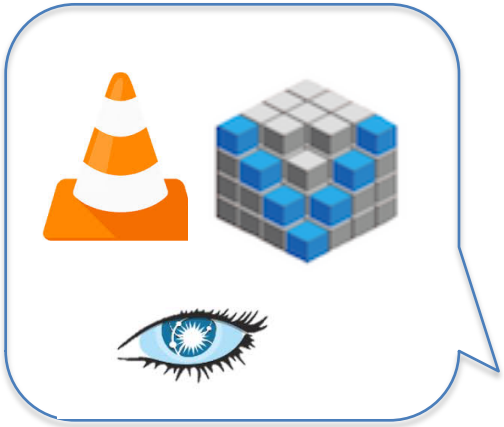
Rerngvit Yanggratoke[1] , Jawwad Ahmed[2], John Ardelius[3], Christofer Flinta[2], Andreas Johnsson[2], Daniel Gillblad[3], Rolf Stadler[1]

(1) KTH Royal Institute of Technology, Sweden
(2) Ericsson Research, Sweden
(3) Swedish Institute of Computer Science (SICS), Sweden

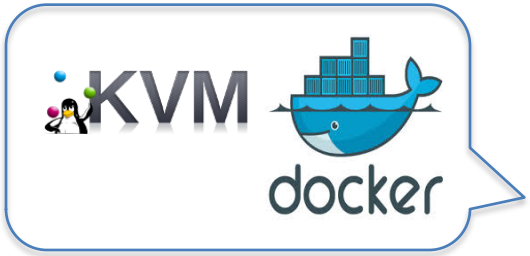
CNSM 2015, Barcelona

November 10, 2015

Overview



Services



....

Operating System

Hardware

Real-time service metrics Y

- Video frame rate, read latency, ..

Analytics

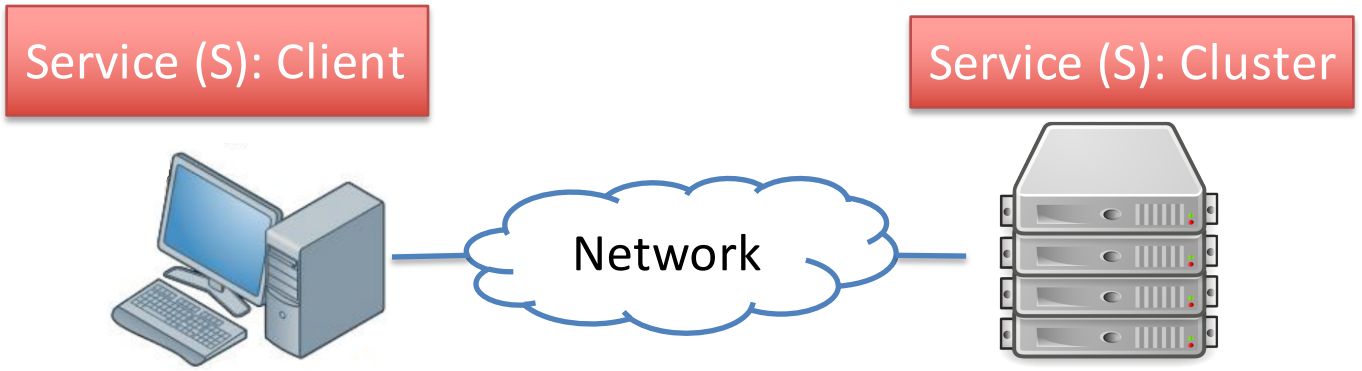
Device statistics X

- CPU load, memory load, #context switching, #processes, etc..
- We read raw data from /proc provided by Linux kernel

Outline

- Real-time prediction problem
- Service-agnostic approach
- Device statistics and service metrics
- Testbed and traces
- Real-time analytics engine
- Evaluation: batch, online, real-time
- Conclusions

Real-time prediction problem



Y: service-level metrics

- Video frame rate, audio buffer rate, network read rate
- Video streaming (VLC)

X: device statistics

- CPU load, memory load, #context switching, #processes, …

Problem : **M: X→ $\hat{Y}$** predicts Y in real-time

Use case : Building block for real-time service assurance for service operator or infrastructure provider

Service-agnostic approach

Existing works

1. Apply formal models to model the system and the service
2. Statistical learning on few service-specific features (<10)

Design goal ➔ "Service-agnostic prediction"

Our approach

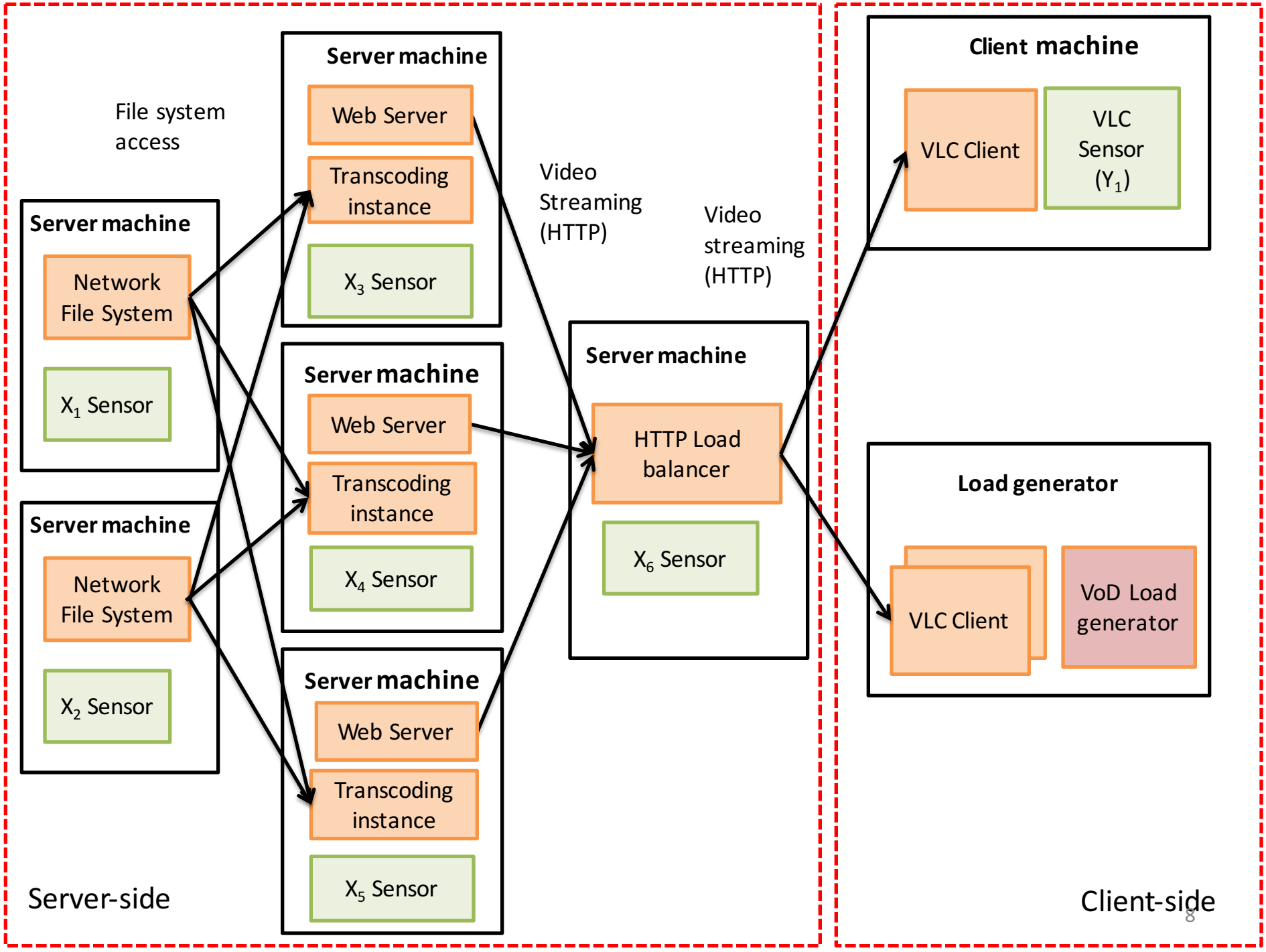
1. Take "all" available statistics (>= 4000 features)
2. Learn using low-level (OS-level) metrics

Device statistics X

- Interface: System Activity Report (SAR) X
  - SAR computes metrics from /proc over time interval
  - CPU core utilization, memory and swap space utilization, disk I/O statistics, …
  - About 840 features per machine
- SAR is based on /proc directory
  - Linux Kernel statistics
  - CPU core jiffies, current memory usage, virtual memory statistics, #processes, #blocked processes, …
- Use numerical features only for model predictions

Service-level metrics Y

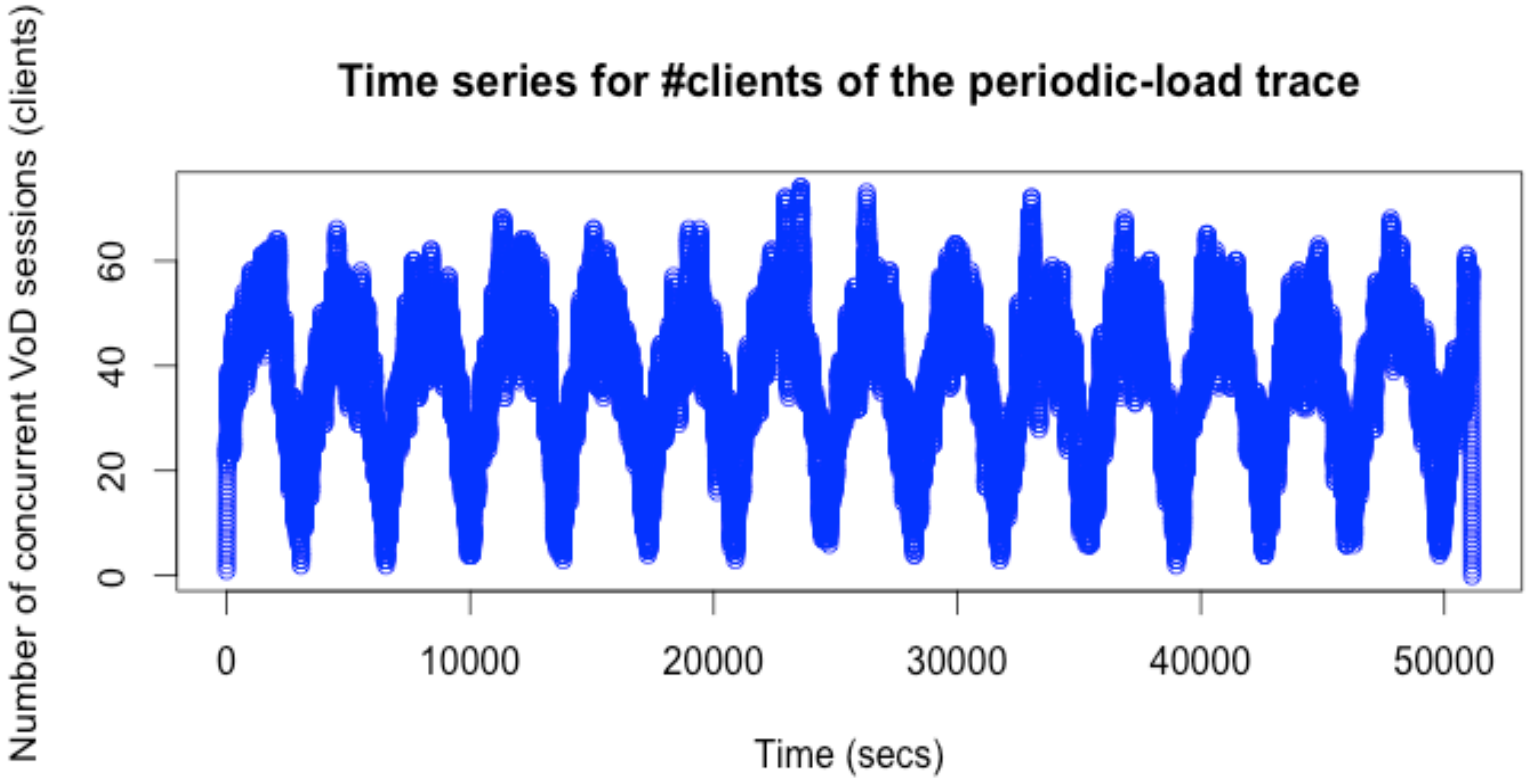
- Video streaming service based on VLC software
- Measured metrics
  - Video frame rate (frames/sec)
  - Audio buffer rate (buffers/sec)
  - Network read rate (operations/sec)
- Instrumented VLC software


File system access
Server machine
Network File System
X_1 Sensor
Server machine
Network File System
X_2 Sensor
Server machine
Web Server
Transcoding instance
X_3 Sensor
Server machine
Web Server
Transcoding instance
X_4 Sensor
Server machine
Web Server
Transcoding instance
X_5 Sensor
Video Streaming (HTTP)
Server machine
HTTP Load balancer
X_6 Sensor
Video streaming (HTTP)
Client machine
VLC Client
VLC Sensor (Y_1)
Load generator
VLC Client
VoD Load generator
Server-side
Client-side

Dell PowerEdge R715 2U rack servers
CPU: two 12-core AMD Opteron processors
Memory: 64 GB RAM
Harddisk: 500 GB hard disk
NIC: 1 Gb network controller

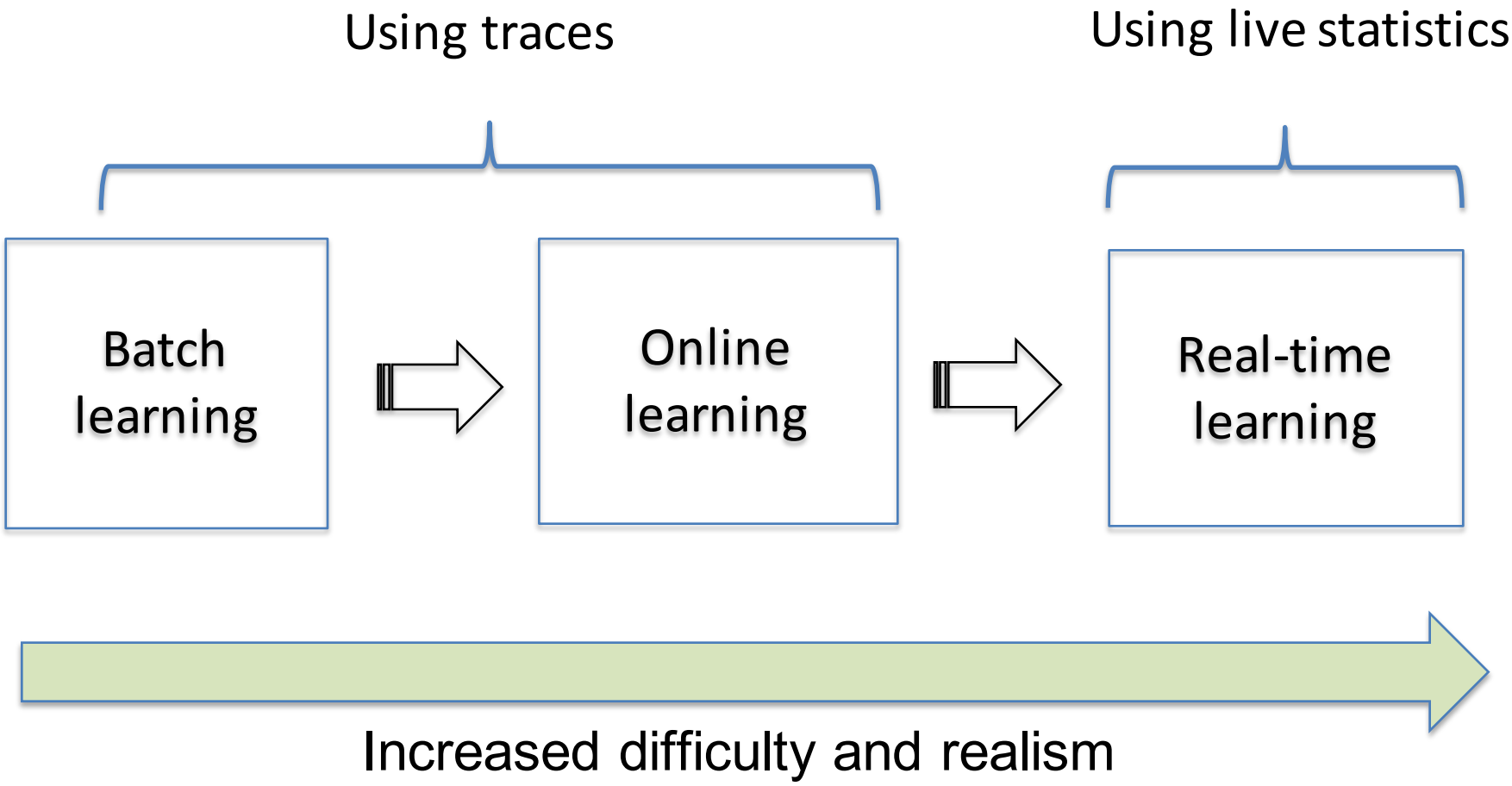
X-Y traces

Load patterns: Periodic-load, flashcrowd, poisson, ….

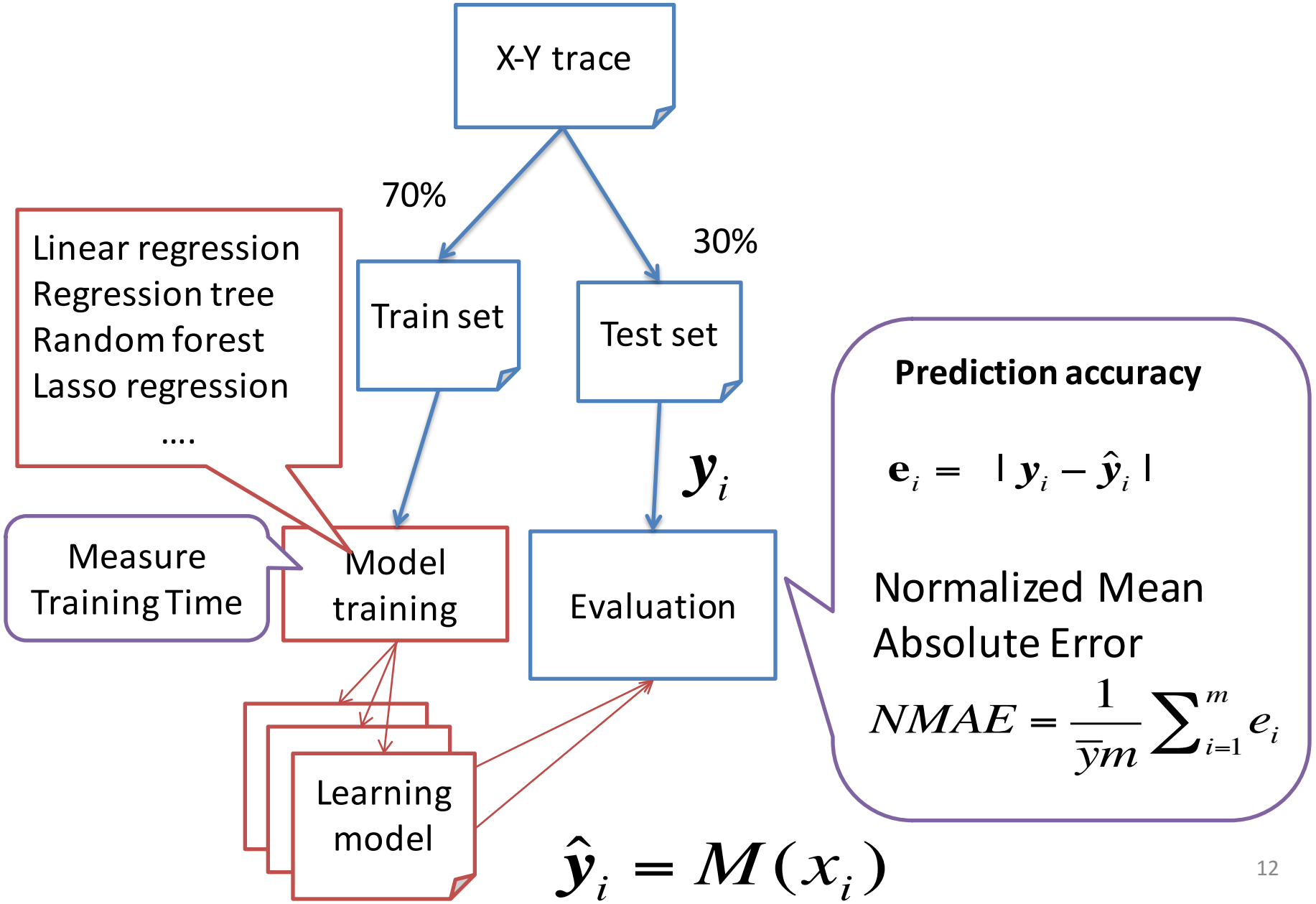


Traces published at **http://mldata.org**

Prediction methods



Batch learning on traces



Reduce feature set

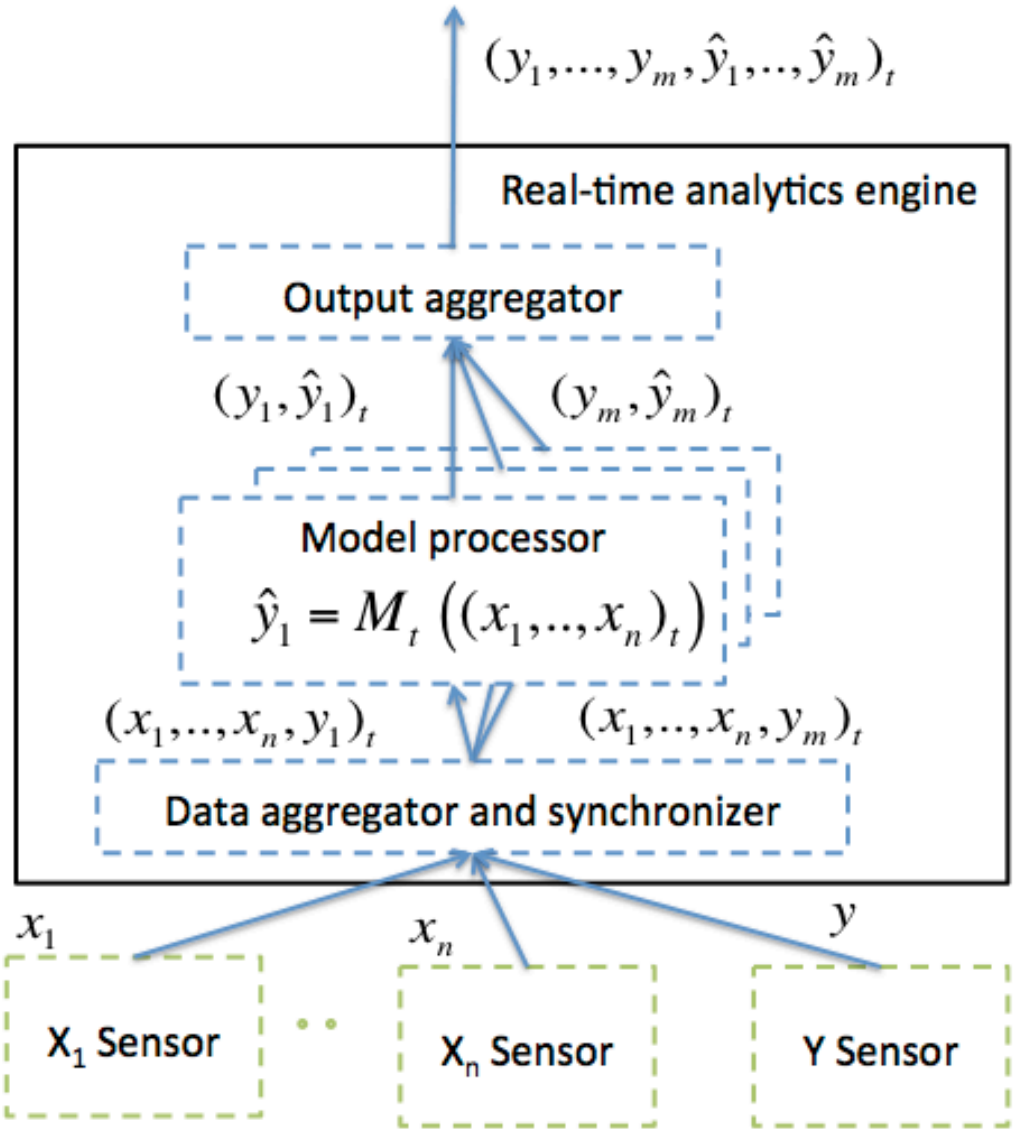
- Exhaustive search is infeasible
  - Requires $O(2^p)$ training executions (p = ~5000)
- Forward stepwise feature selection
  - Heuristic method $O(p^2)$ training executions
  - Incrementally grows the feature sets
- Reduce feature set from 5000 => 12 features

Effect of feature set reduction

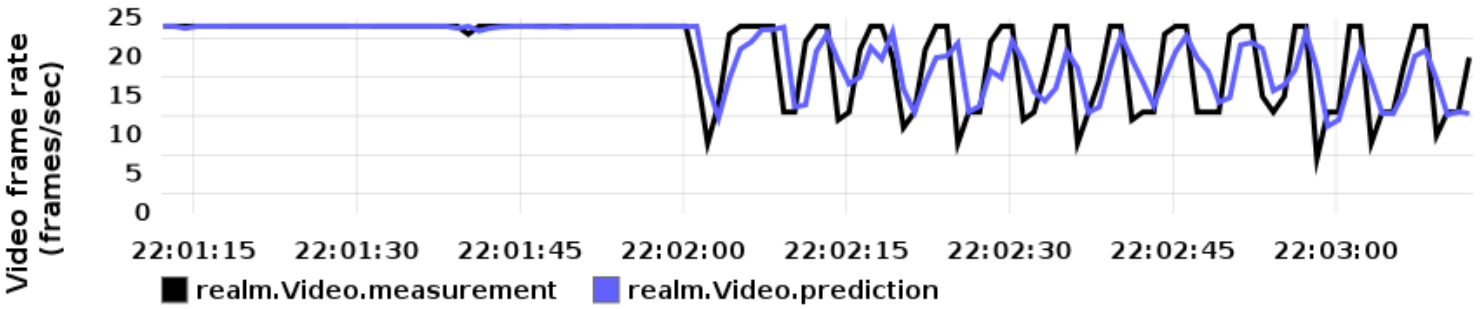
Trace	Feature set	Video		Audio	
		NMAE (%)	Training (secs)	NMAE(%)	Training (secs)
Periodic-load	Full	12	> 59000	32	> 70000
	"Minimal"	**6**	862	**22**	1600
Flash-load	Full	8	> 55000	21	> 75000
	"Minimal"	**4**	778	**15**	1750

=> Minimal feature set
improves prediction accuracy
reduces training time

Real-time analytics engine

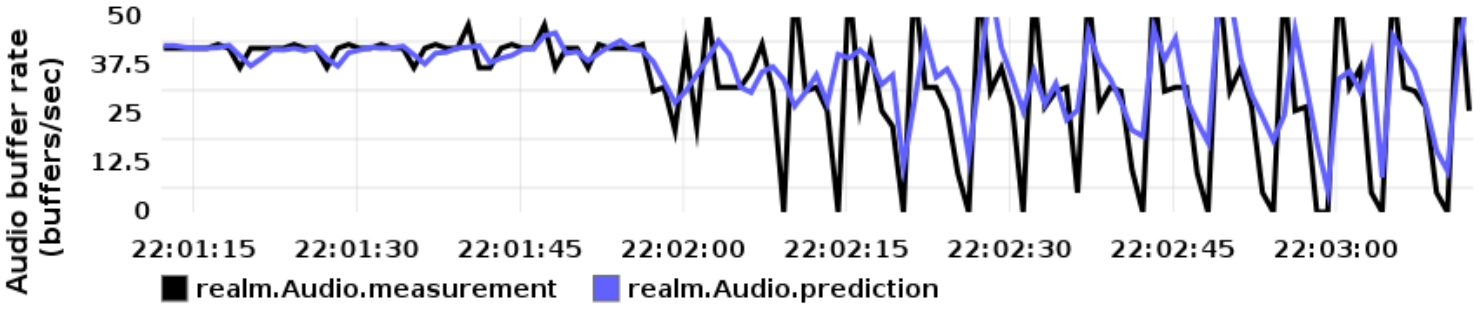


Real-time Predictions of Service Metrics from Device Statistics



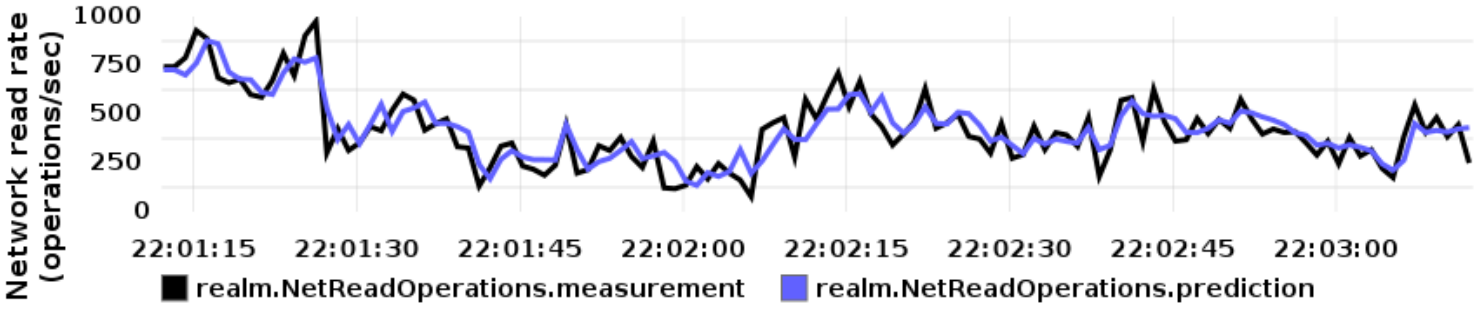
Normalized Mean Absolute Error (last 5 minutes)

4.43 %



Normalized Mean Absolute Error (last 5 minutes)

13.73 %



Normalized Mean Absolute Error (last 5 minutes)

18.77 %

Real-time learning results

Real-time load pattern	NMAE(%)		
	Video	Audio	Network
Periodic-load pattern	3.6	14	28.5
Flash-load pattern	5.6	11	28

Discussion

- It is feasible to predict real-time service metrics from device statistics
- Feature set reduction is critical for real-time prediction
- Random forest on our testbed is the best performing method
- The key strength of this approach is that it is service agnostic