Predicting Service Metrics using Real-time Analytics

ROYAL INSTITUTE OF TECHNOLOGY

Rolf Stadler
School of Electrical Engineering
KTH Royal Institute of Technology, Sweden

April 25, 2016

AnNet 2016, Istanbul, Turkey

KTH VETENSKAP OCH KONST

ROYAL INSTITUTE OF TECHNOLOGY

The Success of Analytics in System Engineering

DARPA Grand Challenge (2004-2007)

- competition for autonomous vehicles

IBM’s Watson (2011)

- natural language processing and machine learning on unstructured data

Google Brain (2012)

- recognize higher-level concepts from unlabeled images

Google’s Alpha Go (2016)

- plays Go on level of best professionals

Cloud for Analytics vs. Analytics for Cloud

Cloud technologies in support of (big) data analytics

- enable virtualization, scaling, pay-as-you-go, multi-tenancy
- new computing paradigms, online algorithms, stream processing
- platforms

Analytics in support of engineering, operations of cloud technologies and services

→ this talk

ROYAL INSTITUTE OF TECHNOLOGY

The Role of Analytics in Systems Engineering and Operations

- What are the benefits and the costs of applying analytics methods?
- For which cases outperform analytics method traditional methods or provide new capabilities?
- How can we integrate analytics into an overall engineering methodology?

- Experience shows that both data science and domain knowledge needed.

  → Need to train engineers in data science.

Why Analytics for Clouds?

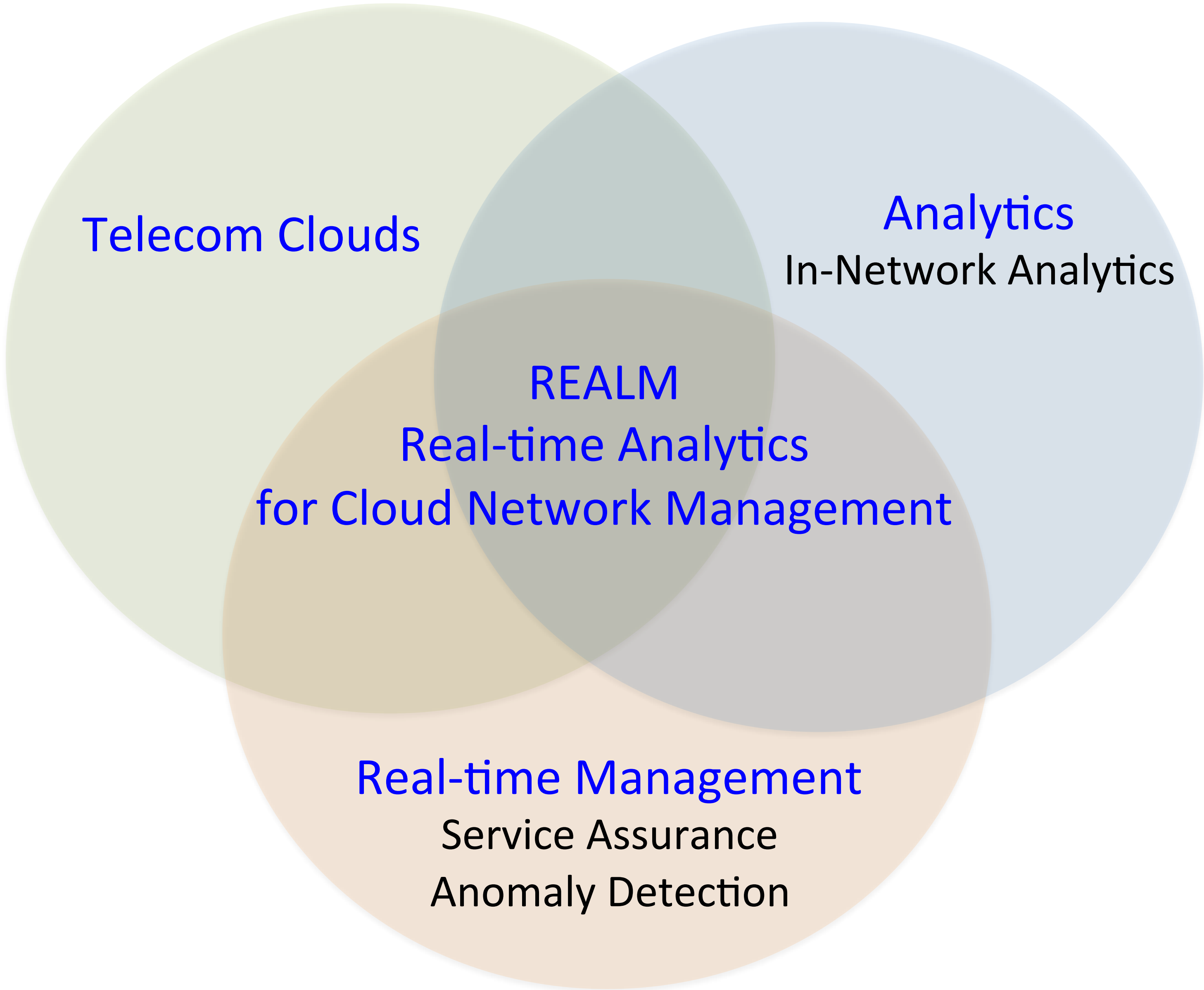
Enablers

- large amounts of counters, statistics, event streams
- technology has progressed to enable real-time storage, processing at source
- availability of platform technology

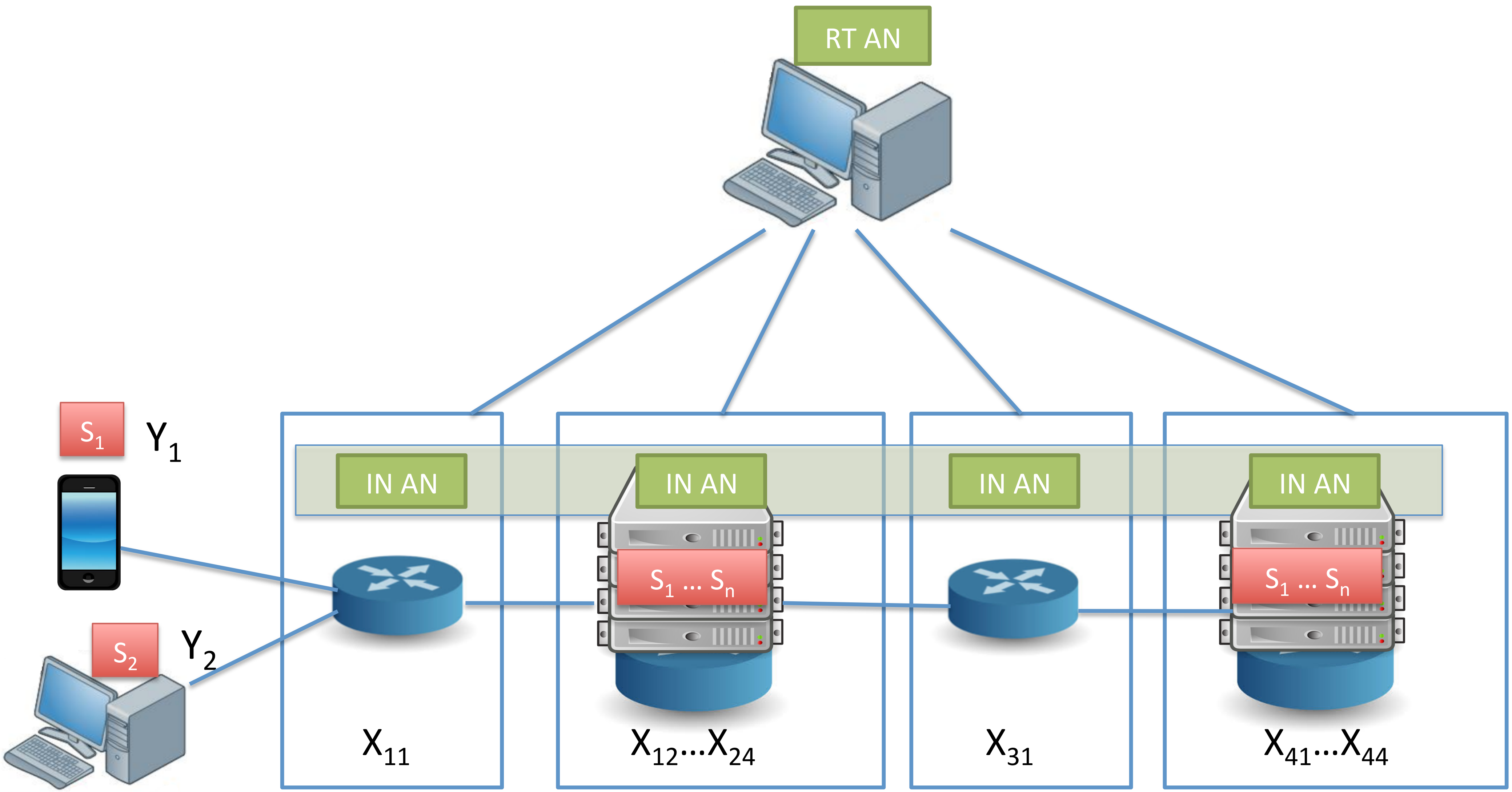
Need

- complexity makes traditional methods infeasible
  statistical learning creates a system model through observation
  without detailed knowledge of system architecture and its functional components

Example: Real-time Analytics for Network Management
Collaboration between Ericsson Research, KTH, SICS

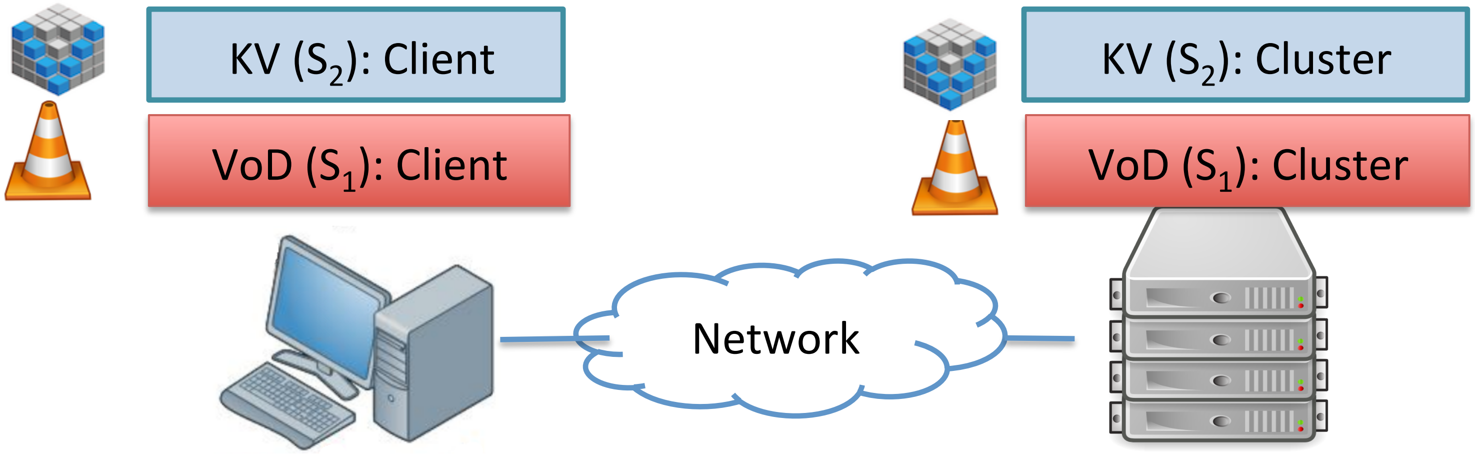


Real-time Prediction of Service Metrics



M: $X_{11} \ldots X_{44} \rightarrow \hat{Y}_1 \ldots \hat{Y}_n$: In-Network Computation

The Problem



Y: service-level metrics

X: device statistics

Video-on-demand (VoD)

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

KV-store

- response time

CPU load, memory load, #network active sockets, #context switching, #processes, etc..

Find M: X→ Ŷ that predicts Y in real-time.

Real-time Analytics for Management

Goal:

- Predicting Service Metrics from Device Statistics in real-time

Approach:

- Statistical learning, online methods,
  distributed learning on compute servers and network nodes;
  Experimentation on testbed

Benefits of approach:

- service-agnostic methodology, scalablity, …

Challenges:

- Large feature set (>1k features)
- Concept drift through changing load patterns and resource management functions

Device Statistics X

- Linux kernel statistics X_{proc}
  - Features extracted from /proc directory
  - CPU core jiffies, current memory usage, virtual memory statistics, #processes, #blocked processes, …
  - Some 4000 metrics
- System Activity Report (SAR) X_{sar}
  - SAR computes metrics from /proc over time interval
  - CPU core utilization, memory and swap space utilization, disk I/O statistics, …
  - Some 840 metrics
- X_{proc} contains many OS counters, while X_{sar} does not
- For model predictions, focus on numerical features
- Sensors read statistics 1-2 times per sec.

Service Metrics Y

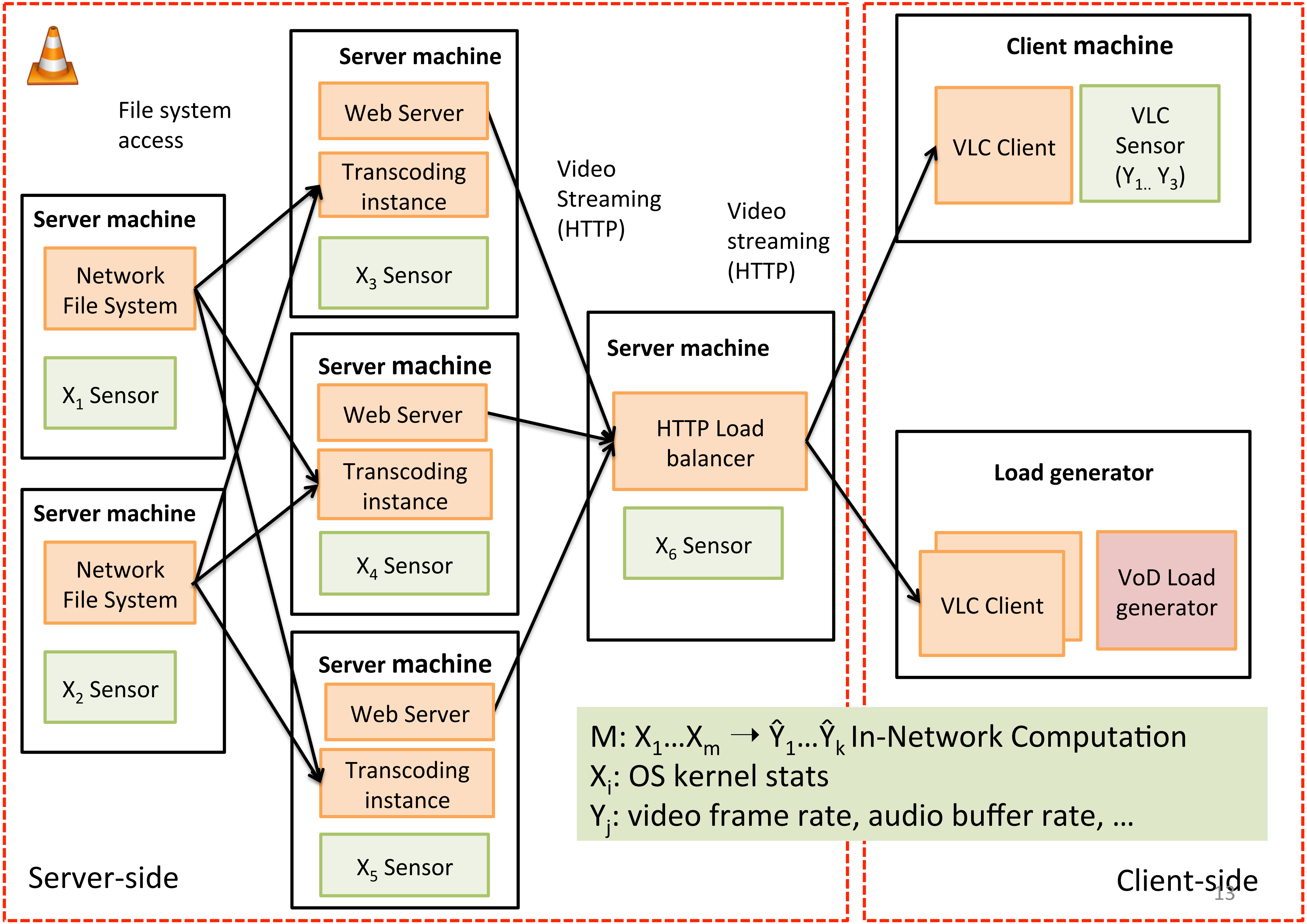
Video-on-Demand

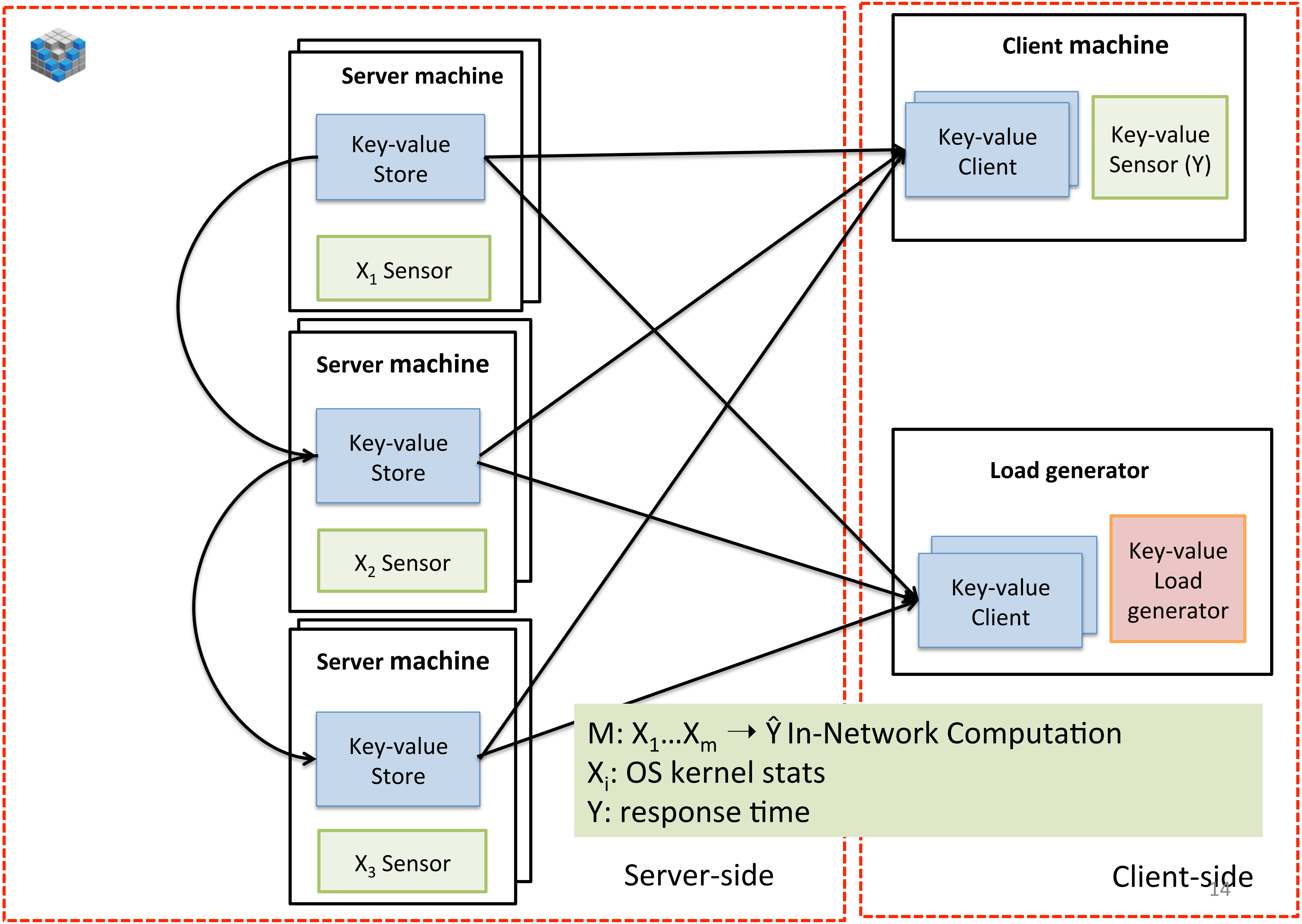
- Video streaming service based on VLC media player.
- We instrumented the VLC software to capture underlying events to compute the metrics.
- Metrics:
  video frame rate, audio buffer rate, RTP packet rate, …

KV-storage system

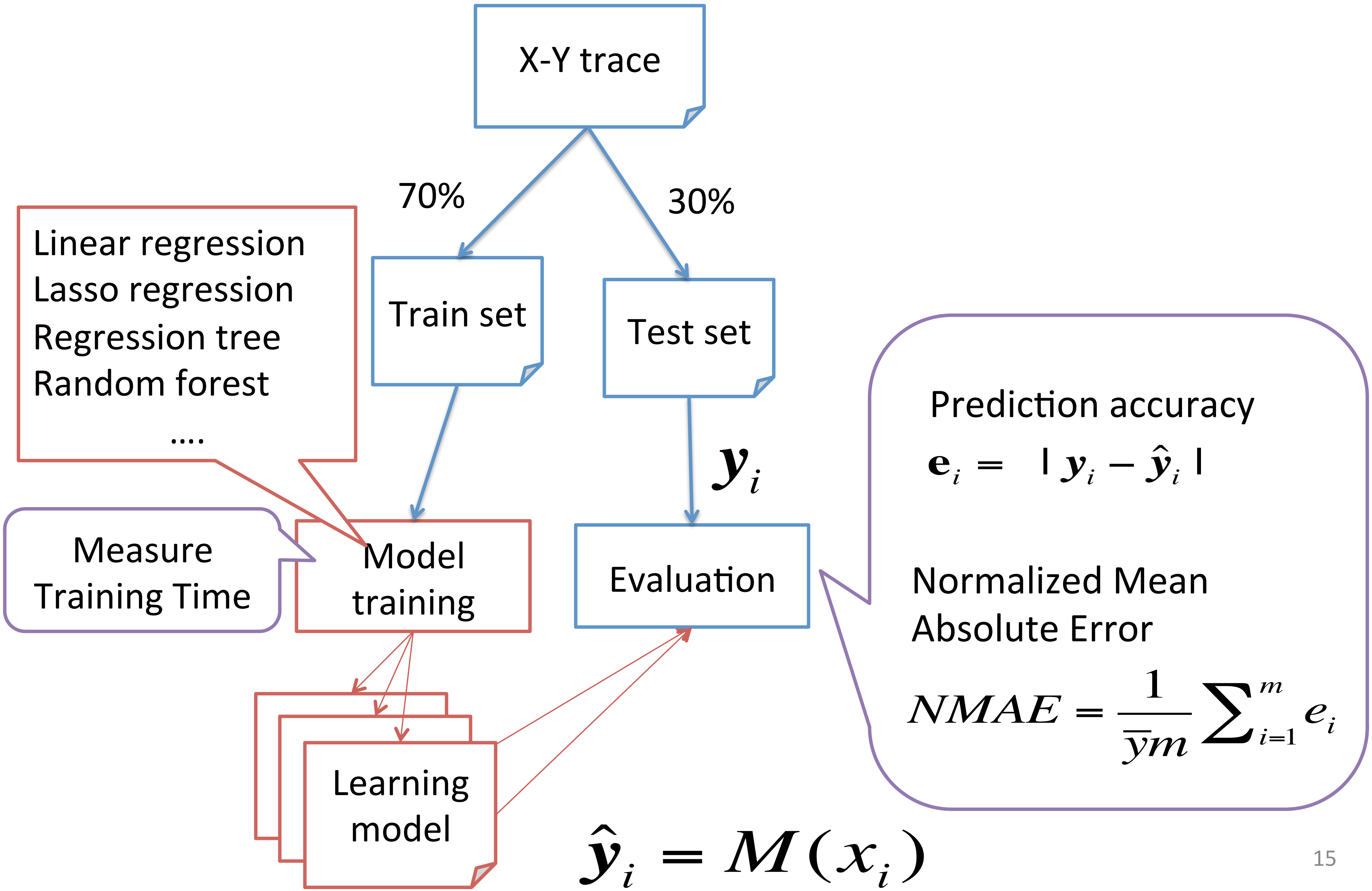
- Voldemort p2p system
- Metrics:
  response time

Metrics captured 1-2 times per sec.

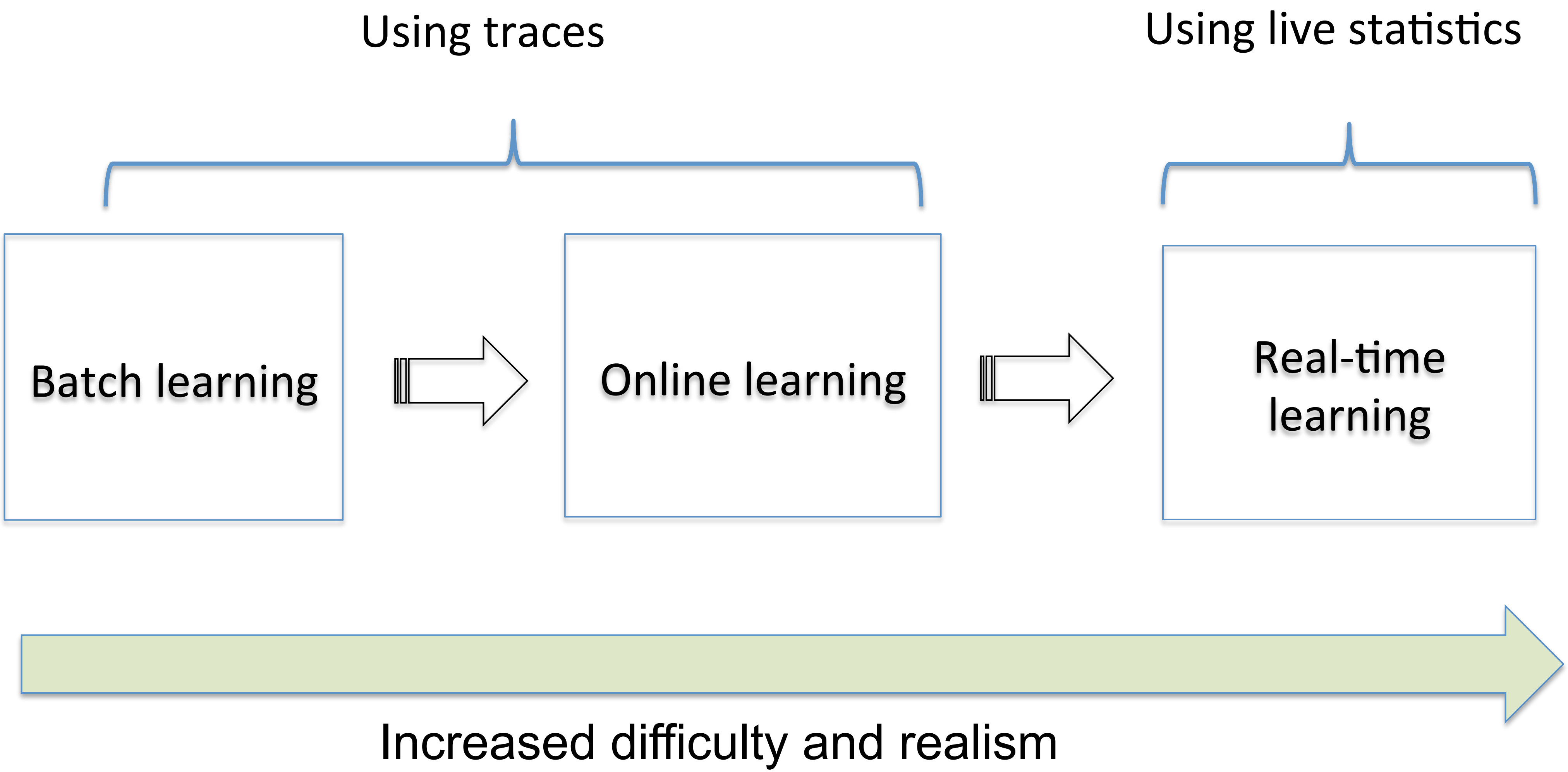

File system access
Server machine
Network File System
X_1 Sensor
Server machine
Network File System
X_2 Sensor
Server machine
Web Server
Transcoding instance
X_3 Sensor
Server machine
Web Server
Transcoding instance
X_4 Sensor
Server machine
Web Server
Transcoding instance
X_5 Sensor
Video Streaming (HTTP)
Video streaming (HTTP)
Server machine
HTTP Load balancer
X_6 Sensor
Client machine
VLC Client
VLC Sensor ($Y_{1..}$ Y_3)
Load generator
VLC Client
VoD Load generator
M: $X_1 \ldots X_m \rightarrow \hat{Y}_1 \ldots \hat{Y}_k$ In-Network Computation
X_i: OS kernel stats
Y_j: video frame rate, audio buffer rate, …
Server-side
Client-side


Server machine
Key-value Store
X_1 Sensor
Server machine
Key-value Store
X_2 Sensor
Server machine
Key-value Store
X_3 Sensor
Client machine
Key-value Client
Key-value Sensor (Y)
Load generator
Key-value Client
Key-value Load generator
M: $X_1 \ldots X_m \rightarrow \hat{Y}$ In-Network Computation
X_i: OS kernel stats
Y: response time
Server-side
Client-side

Batch Learning on Traces



Prediction Methods



Feature Set Reduction

- Exhaustive search is infeasible
  - Requires $O(2^p)$ training executions $(p \approx 5000)$
- Option: forward stepwise feature selection
  - Heuristic method $O(p^2)$ training executions
  - Incrementally grows the feature sets
- Reduces feature set from 5000 to 12 features

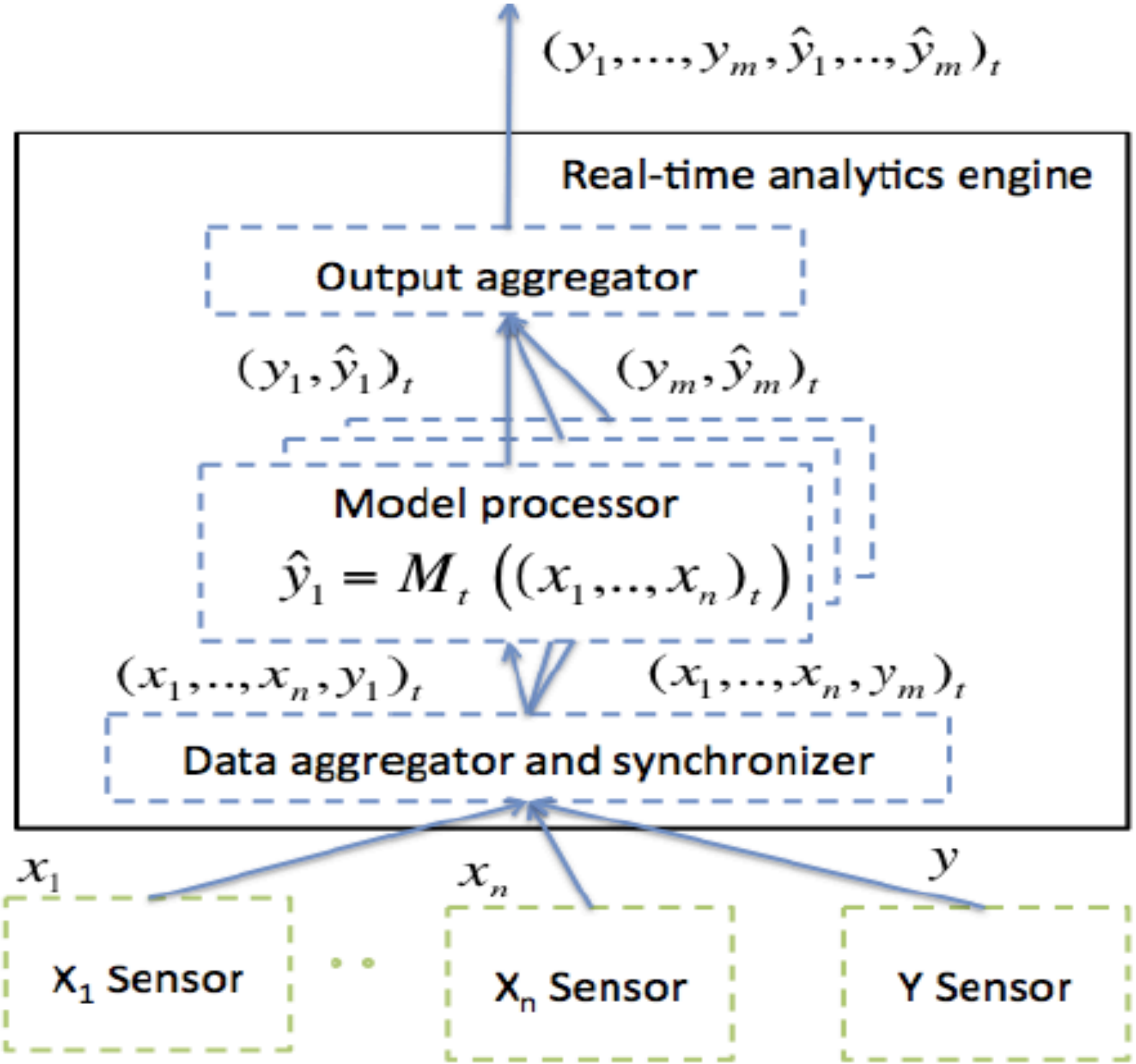
Effect of Feature Set Reduction

Load pattern	Feature set	Video		Audio	
		NMAE (%)	Train (sec)	NMAE(%)	Train (sec)
Periodic	Full	12	> 59000	32	> 70000
	"Minimal"	**6**	862	**22**	1600
Flash	Full	8	> 55000	21	> 75000
	"Minimal"	**4**	778	**15**	1750

"Minimal" feature set
- improves prediction accuracy over full set, feature set selected by experts
- reduces training time

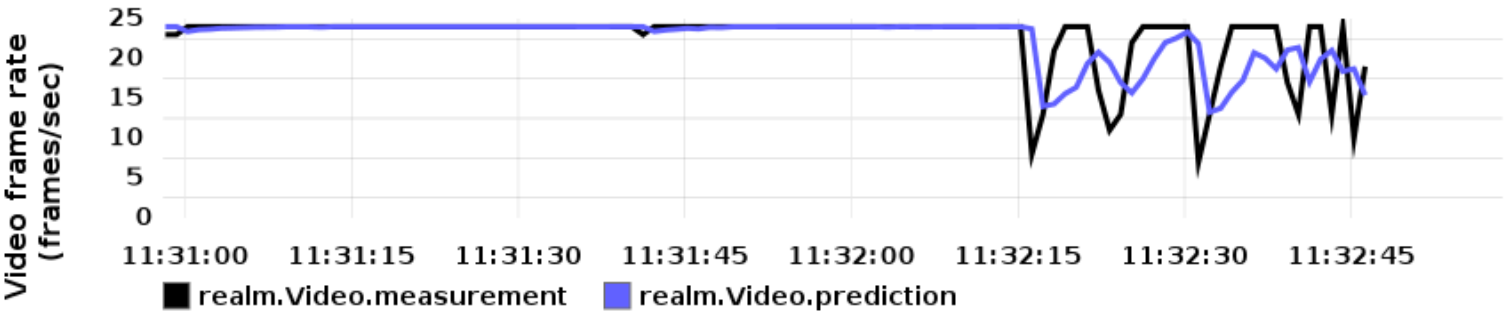
$$NMAE = \frac{1}{\bar{y}}\left(\frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right|\right)$$

Real-time Model Computation and Evaluation



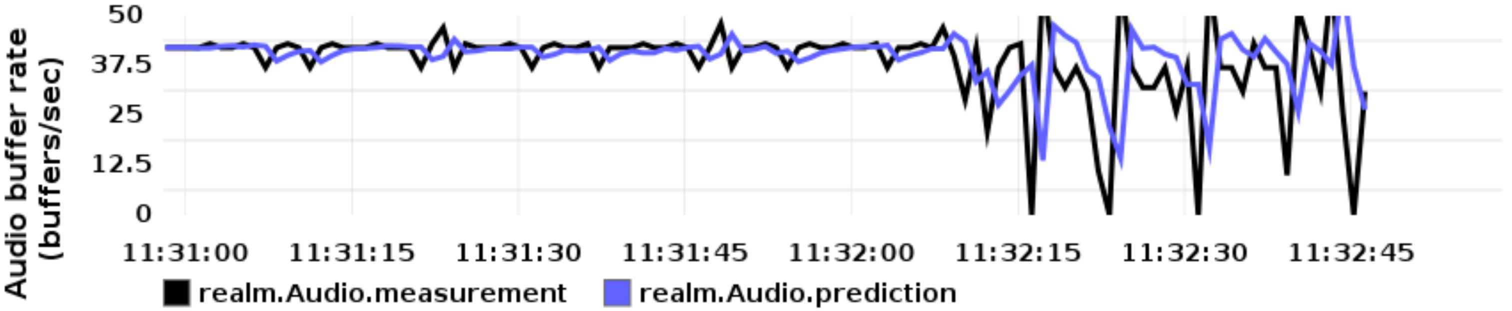
Visualizing Output from Analytics Engine

Real-time Predictions of Service Metrics from Device Statistics



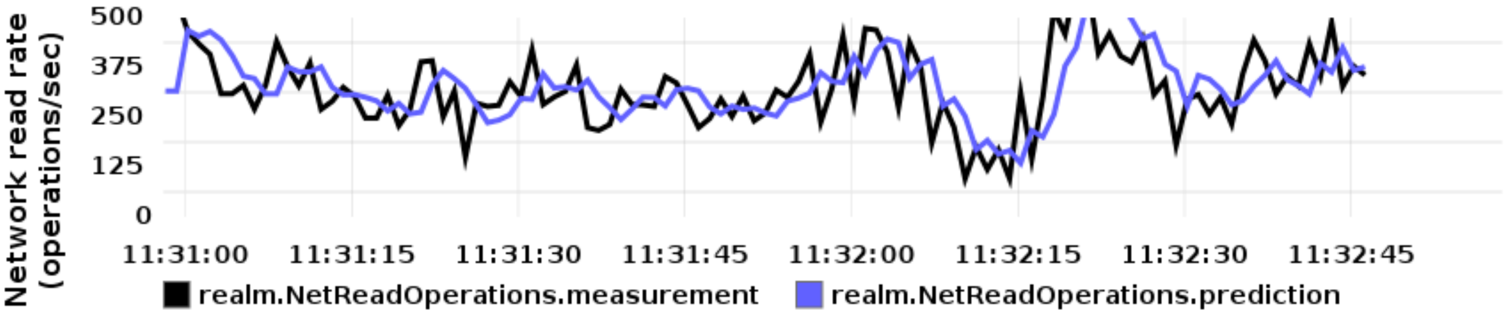
Normalized Mean Absolute Error (last 5 minutes)

2.95 %



Normalized Mean Absolute Error (last 5 minutes)

10.46 %



Normalized Mean Absolute Error (last 5 minutes)

22.31 %

Evaluation for Real-time Learning

Load pattern	Video-on-demand (VoD)		KV-store
	Video frame rate	Audio buffer rate	Response time
Periodic	3.6%	14%	7%
Flashcrowd	5.6%	11%	6%
Periodic + Flashcrowd	8%	29%	11%

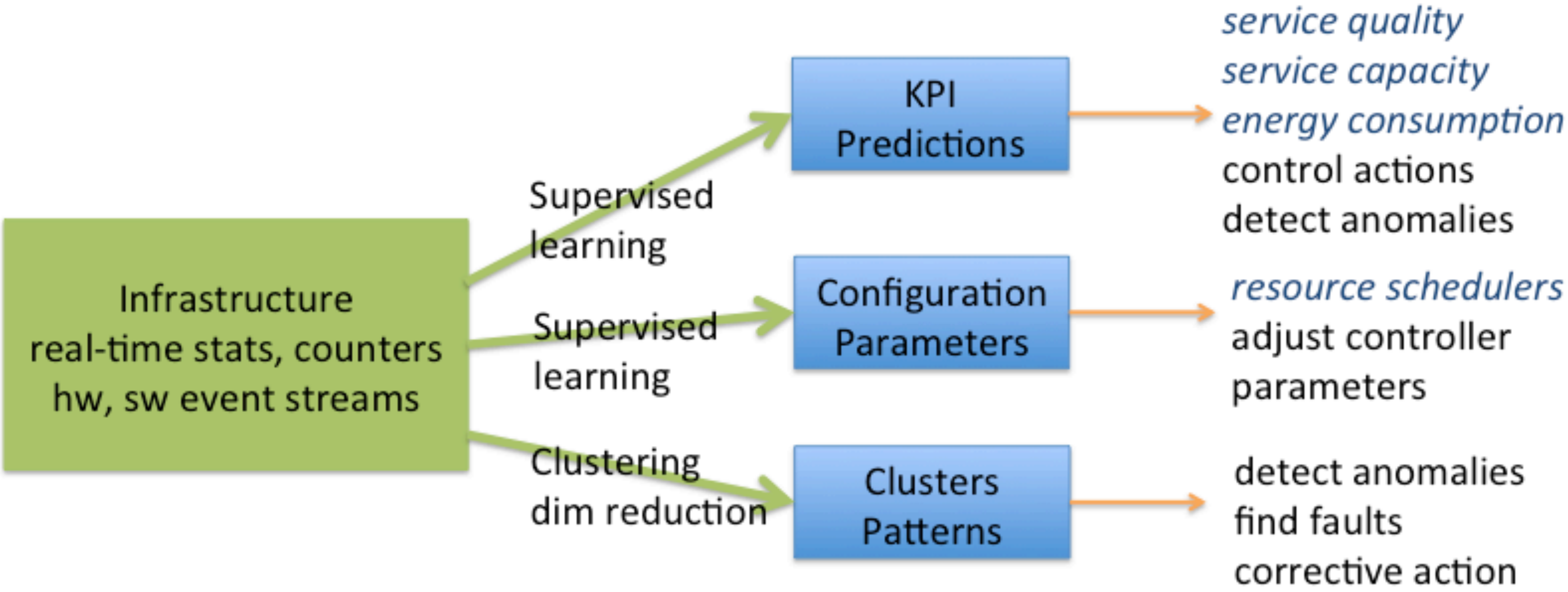
With virtualized infrastructure: $\approx$ same results

End-to-end service along network path: +10%

Results to Date

- Predicting service metrics for cluster-based services is feasible:
  - video streaming, key-value store
    (NMAE below 14% for video, audio frame rates, etc.)
- Feature set reduction on X_{sar} reduces model computation time and improves accuracy.
- Real-time analytics engine
  - allows to observe effect of system perturbation on service quality;
  - serves as building block for service quality assurance system, anomaly detection system.

Real-time Analytics Functions for Clouds



KTH VETENSKAP OCH KONST

ROYAL INSTITUTE
OF TECHNOLOGY

Conclusions (1)

The availability of operational and historical data and recent technology advancement make real-time analytics for clouds possible.

Analytics methods create models from observations, without knowing detailed architectural and functional model of a system.

Training of engineers is key.

Conclusions (2)

We demonstrated the feasibility of estimating service metrics in real-time.

Promising application of real-time analytics for engineering and operation of cloud services:

- Estimation of KPIs and control parameters;
- Quality assurance and anomaly detection/root-cause analysis.

Major challenges remain:

- scalability, in-network computation
- integrating analytics into an overall engineering methodology.

Publications

1. R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad, and R. Stadler, “A service-agnostic method for predicting service metrics in real-time,” submitted for publication.

2. R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad, and R. Stadler, “Predicting service metrics for cluster-based services using real-time analytics,” International Conference on Network and Service Management (CNSM 2015), Barcelona, Spain, November 2015.

3. R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad, and R. Stadler, “Predicting Real-time Service-level Metrics from Device Statistics,” International Symposium on Integrated Network Management (IM 2015), Ottawa, CA, May 2015.

4. R. Yanggratoke, J. Ahmed, J. Ardelius, C. Flinta, A. Johnsson, D. Gillblad, and R. Stadler, “A platform for predicting real-time service-level metrics from device statistics,” International Symposium on Integrated Network Management (IM 2015), Demonstration Program, Ottawa, CA, May 2015.

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