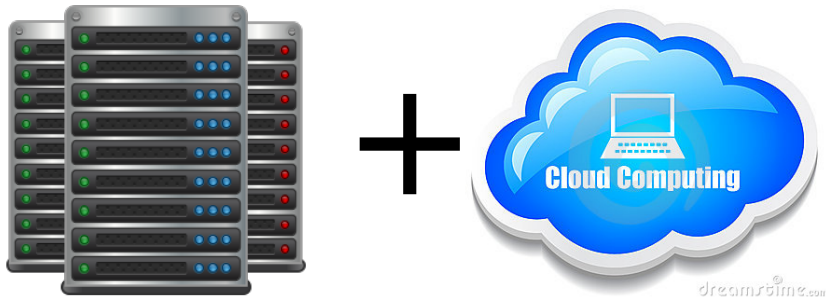
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

A Correlation-based Methodology to Infer Communication patterns between Cloud Virtual Machines

Claudia Canali
Riccardo Lancellotti
Dept of Engineering “Enzo Ferrari”
University of Modena and Reggio Emilia

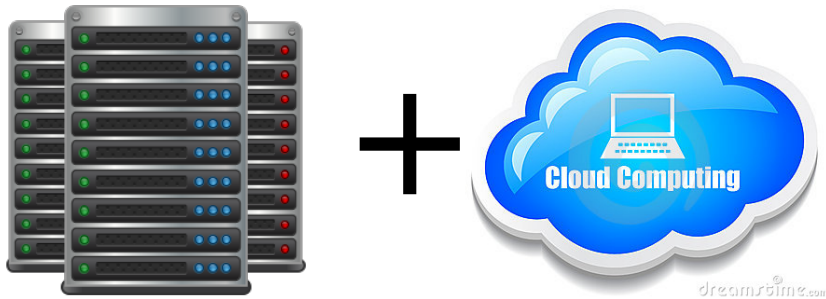
Motivation

- **The challenges of energy efficiency in Data Centers**
  - Multiple Heterogeneous VMs
  - Multiple Resources (CPU, Memory, Networking)
- **The challenges of Cloud Computing**
  - Dynamic environment
  - Complex SLA to meet

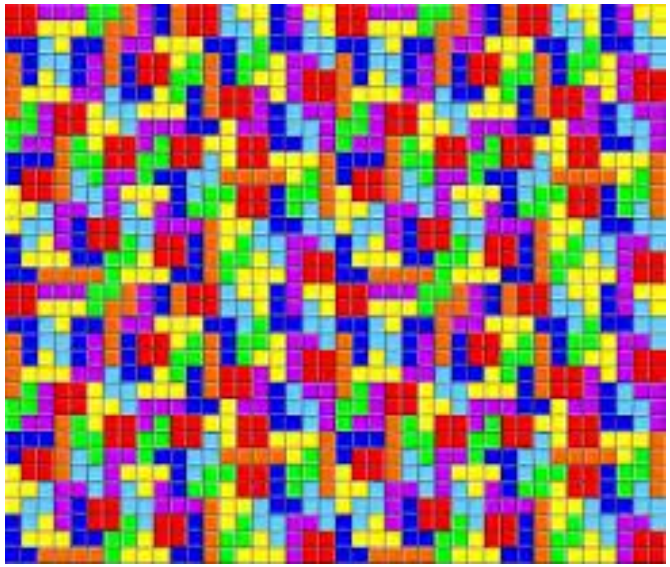


Motivation

- **The challenges of energy efficiency in Data Centers**
  - Multiple Heterogeneous VMs
  - Multiple Resources (CPU, Memory, Networking)
- **The challenges of Cloud Computing**
  - Dynamic environment
  - Complex SLA to meet
- **→ Data center management is a tough game!**



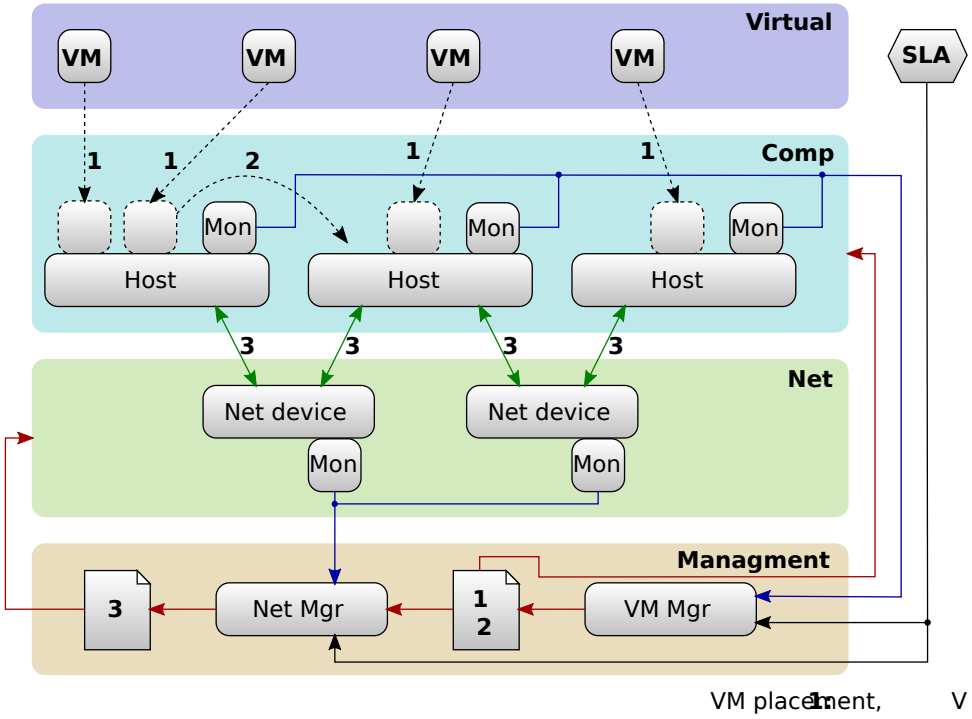
=



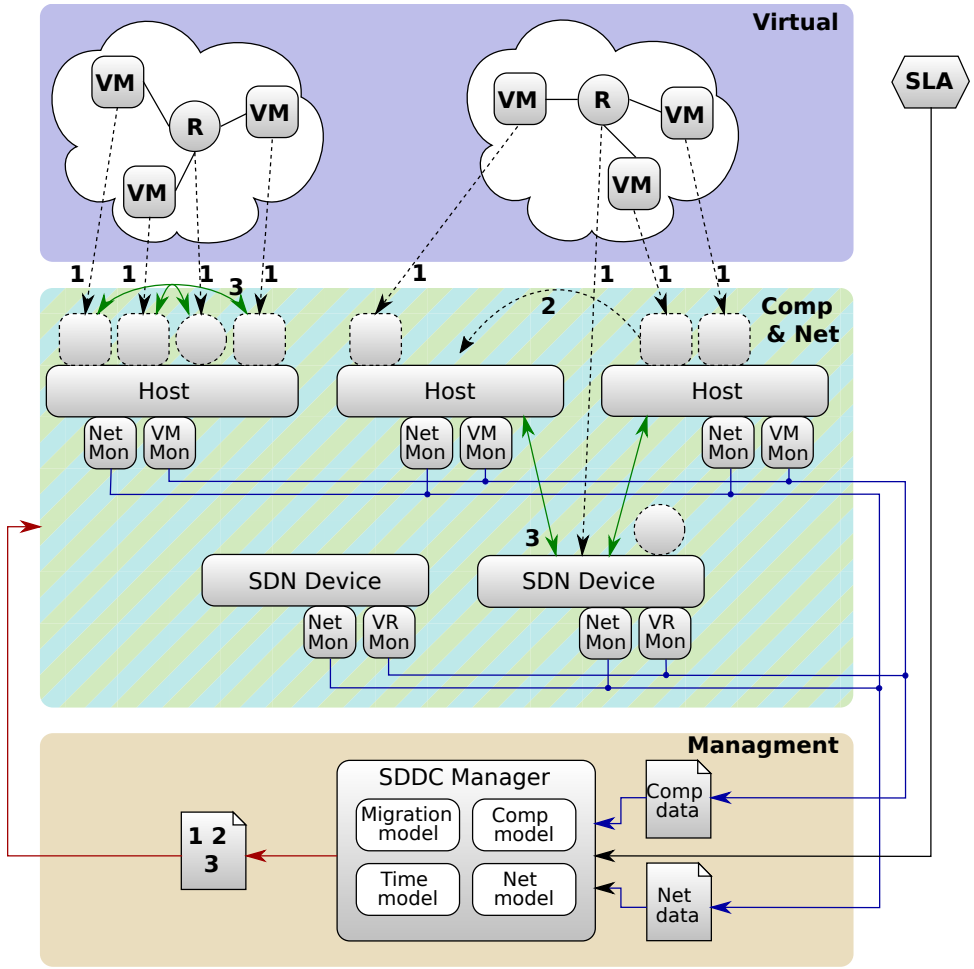
The critical role of networking

- **Typically not considered in existing energy models**
- **Interaction among VMs**
- **Impact of network patterns on:**
  - **Performance:** SLA satisfaction affected by latency
  - **Energy:** Network infrastructure consumes a non-negligible amount of energy
- **Evolution trend:**
  - Network importance is critical
  - Networking is going virtual: VR, NOS, SDN
  - → Introducing the SDDC

Introducing SDDC



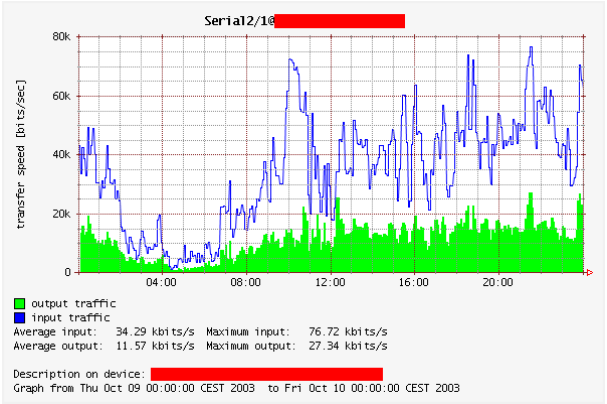
VM placement,

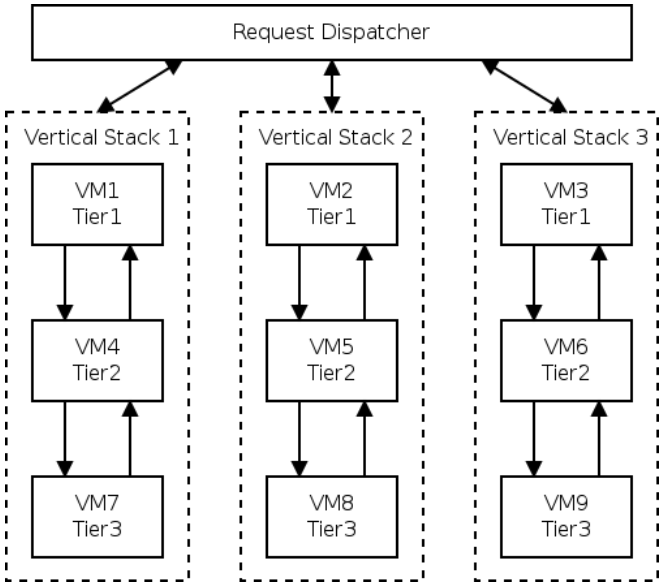


VM placement,

Knowing network patterns

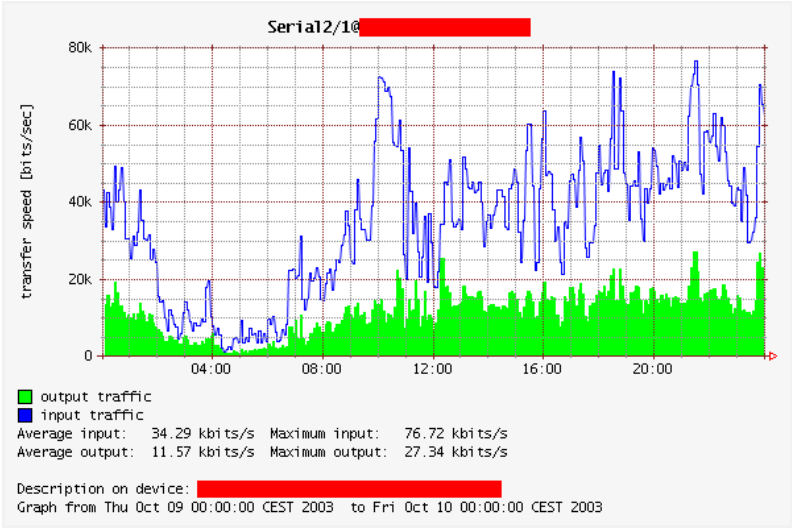
- **Management in SDDC requires knowledge of network patterns**
  - Which VMs exchange data?
  - Available information:
    - → Aggregate data
- **Horizontal replication:**
  - Multiple VMs have similar network patterns
- **→ Open challenge to address**





Goal

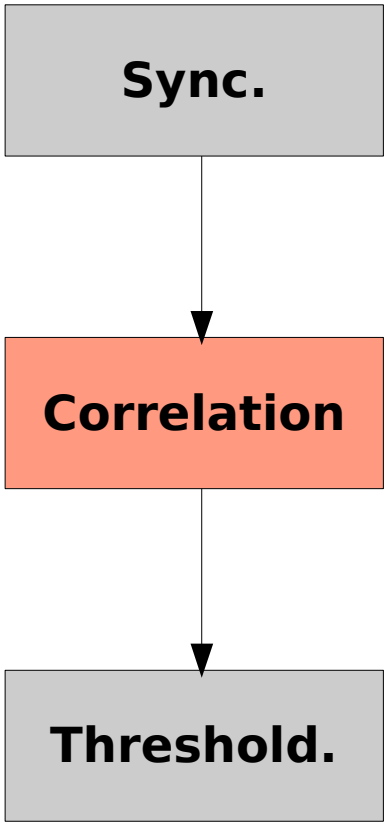
- **Input:**
  - Traffic pattern of each VM
  - Time series of pkt in/out
- **Output:**
  - VMs interaction matrix
- **Caveats:**
  - Presence of horizontal replication
  - Data samples may be not synchronized



	VM1	VM2	VM3	VM4	VM5	VM6	VM7	VM8	VM9
VM1	1	1	1	0	0	0	0	0	0
VM2	1	1	1	0	0	0	0	0	0
VM3	1	1	1	0	0	0	0	0	0
VM4	0	0	0	1	1	1	0	0	0
VM5	0	0	0	1	1	1	0	0	0
VM6	0	0	0	1	1	1	0	0	0
VM7	0	0	0	0	0	0	1	1	1
VM8	0	0	0	0	0	0	1	1	1
VM9	0	0	0	0	0	0	1	1	1

Methodology

- **Synchronization of time series**
  - Cubic interpolation of samples
- **Computation of correlation matrix**
  - Computes correlation matrix between all the (synchronized) time series
  - Multiple correlation indexes are considered
- **Identification of interacting VMs**
  - Use of threshold
  - More complex approaches may be used



Correlation indexes

- **Pearson** correlation coefficient

$$\rho(P_{j_1}^{*out}, P_{j_2}^{*in}) = \frac{E[(P_{j_1}^{*out} - \mu(P_{j_1}^{*out}))(P_{j_2}^{in} - \mu(P_{j_2}^{*in}))]}{\sigma(P_{j_1}^{*out})\sigma(P_{j_2}^{*in})}$$

- **Spearman** correlation coefficient

$$\rho_s(P_{j_1}^{*out}, P_{j_2}^{*in}) = 1 - \frac{6\sum_{i=0}^{T} r(P_{j_1}^{*out}(i)) - r(P_{j_2}^{*in}(i))}{T(T^2-1)}$$

basically we apply the Pearson correlation to the time series of *ranks* for each value in the original samples.

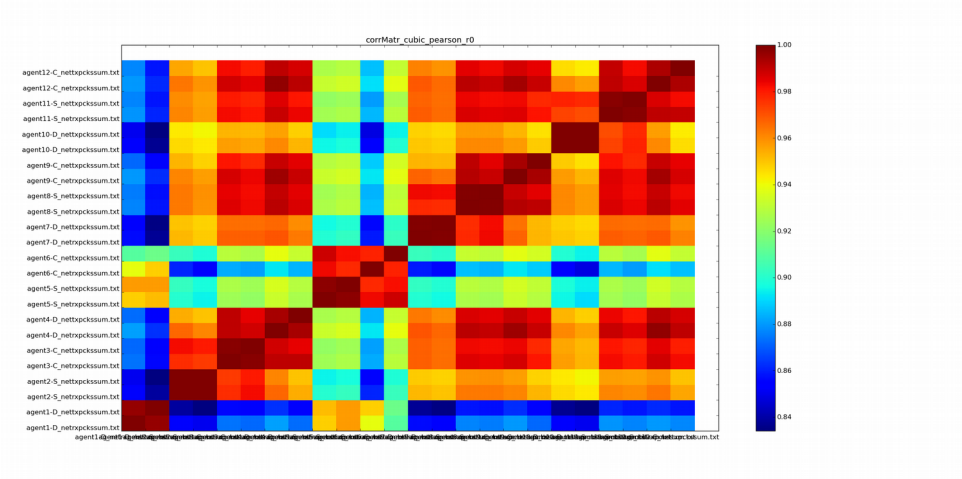
- **Spearman tends to amplify small oscillations around average value**

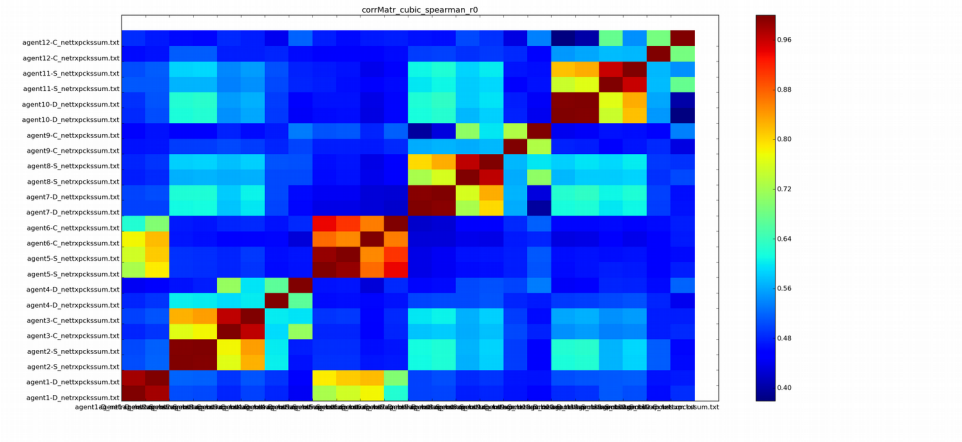
Experimental setup

- **Experiments on Amazon EC2**
  - Use of micro instances
- **Three-tier Web application benchmark: TPC-W**
  - 4 vertical stacks, 3VMs per stack
- **Data collection interval:**
  - 30 sec, 1 min, 2 min
- **Metrics of interest**
  - Precision (TP/TF+TP)
  - Recall (TP/TP+FN)
  - Accuracy (TP+TN/TP+TN+FP+FN)

Qualitative analysis

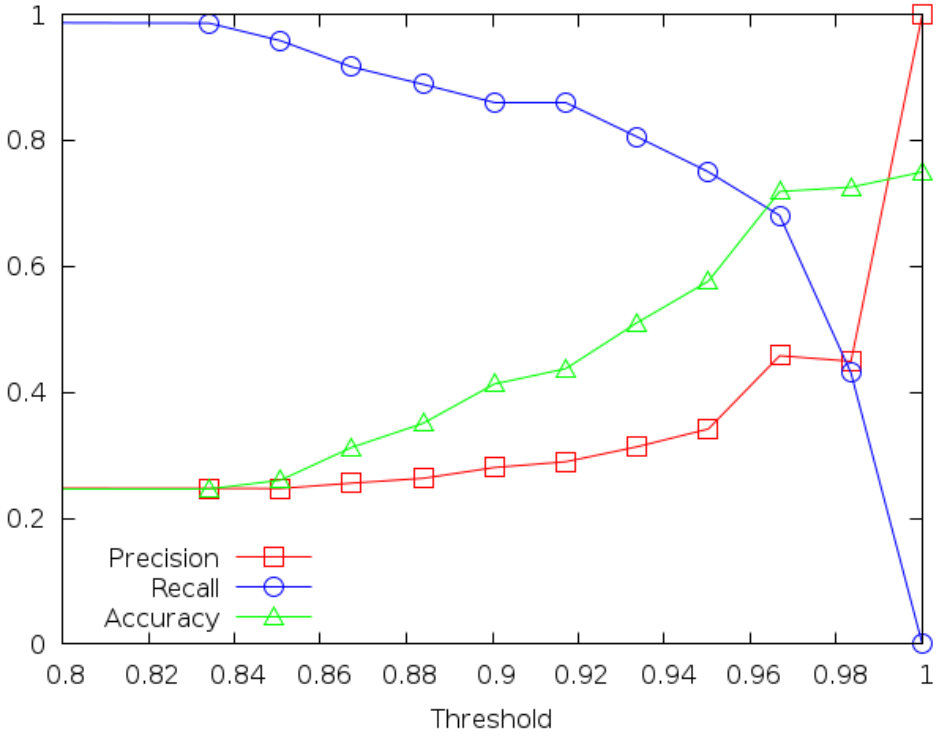
- **Use of heatmap**
- **Ideal result:**
  - Red boxes on diagonal
  - Blue everywhere else
- **Pearson coefficient**
  - Correlation always high
  - Large red halos
- **Spearman coefficient**
  - Seems to identify better the vertical stacks



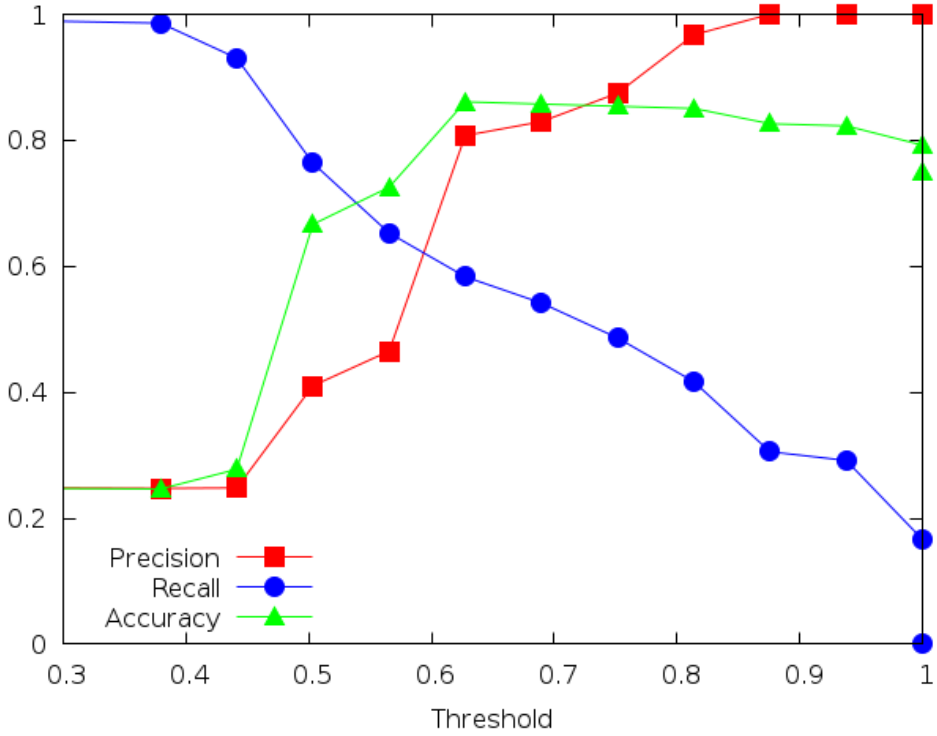


Experimental results

Pearson

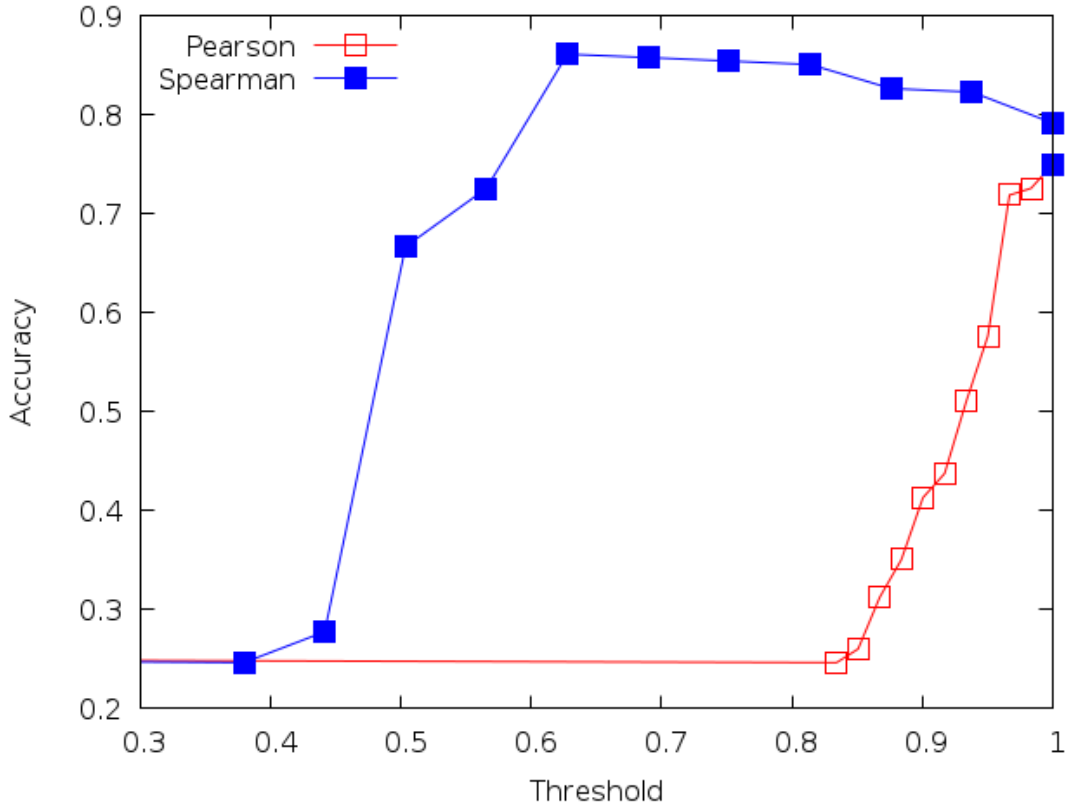


Spearman



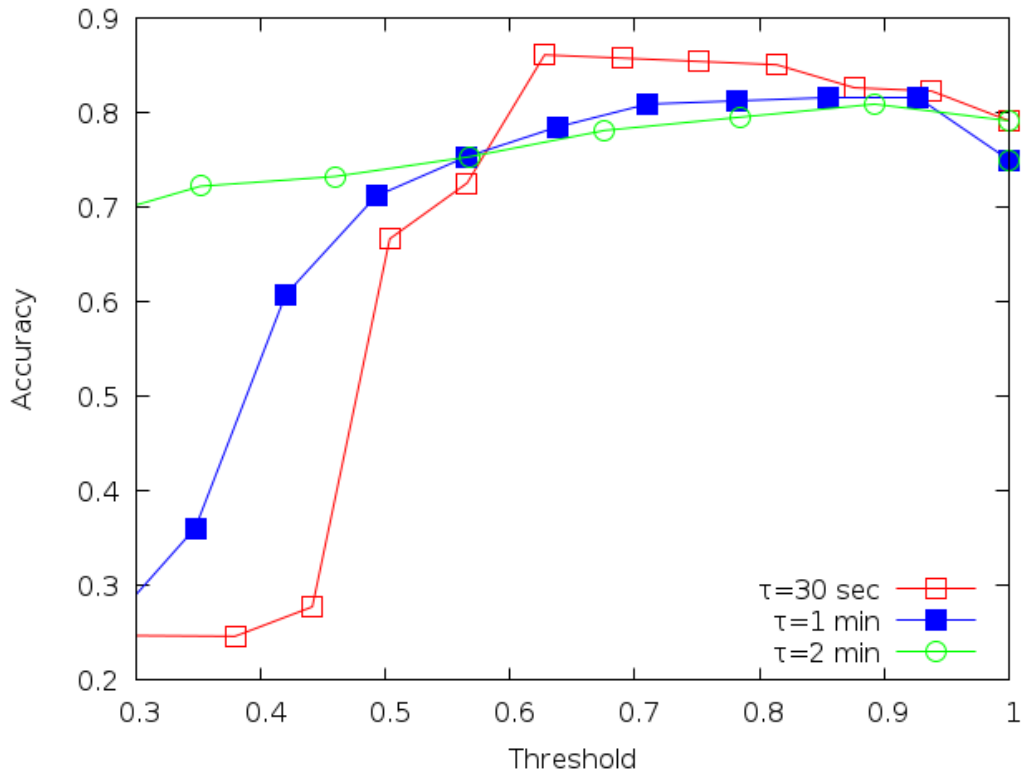
- Precision, Recall, Accuracy
- **Poor precision** for **Pearson** correlation

Comparison



- **Spearman is a clear winner**
  - Higher accuracy
  - Better stability w.r.t. Threshold

Sensitivity to sampling period



- **Smoothing effect of sampling frequency**
  - Reduced maximum accuracy
  - Increased stability w.r.t. Threshold

Conclusions

- **Energy management in cloud data centers**
  - Need to consider network interactions
  - No per-destination/per source breakdown of traffic
- **Proposal of a novel methodology**
  - Interacting VMs from aggregated network data
  - Horizontal replication + traces not synchronized
- **Experiments on a cloud infrastructure**
  - Comparison of correlation indexes
  - Sensitivity to sampling frequency

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

A Correlation-based Methodology to Infer Communication patterns between Cloud Virtual Machines

Claudia Canali
Riccardo Lancellotti
Dept of Engineering “Enzo Ferrari”
University of Modena and Reggio Emilia