Impact of theoretical performance models on the design of fog computing infrastructures

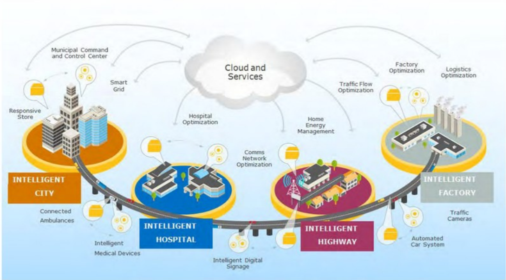
Claudia Canali,
Riccardo Lancellotti,
Stefano Rossi
DIEF, University of Modena and Reggio Emilia

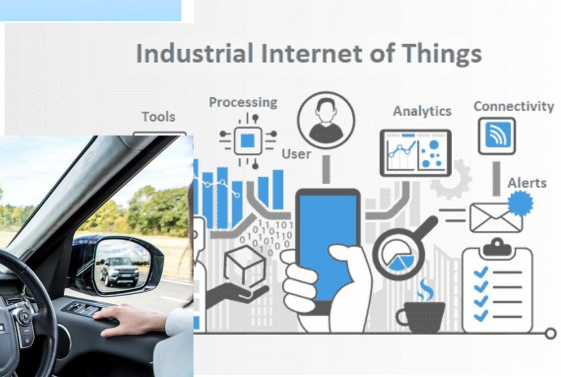


Motivating sceanrio

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

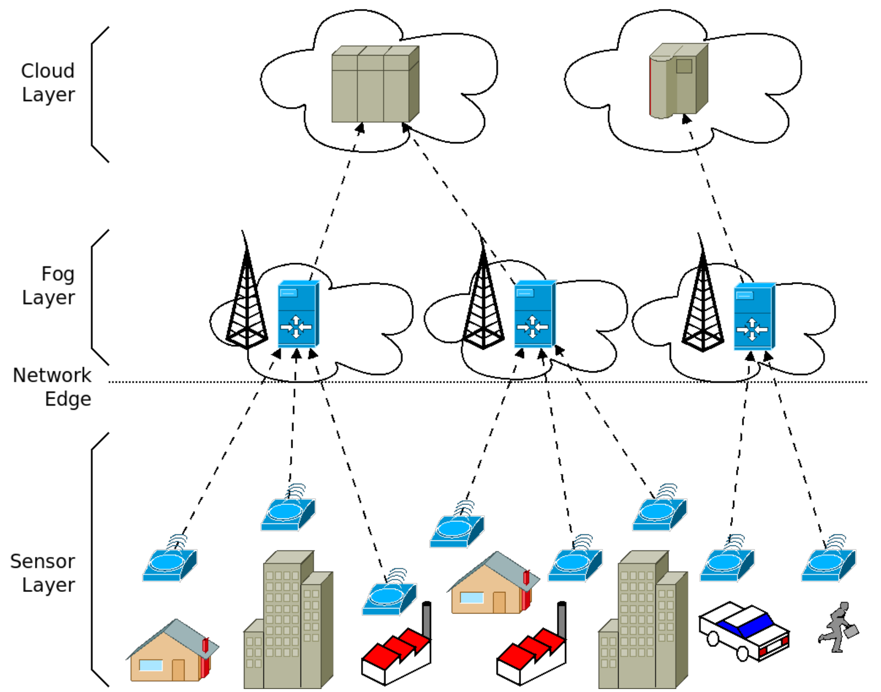
- Fog evolution motivated by new applications
- Several fields of application:
  - Smart cities
  - Industrial
  - Automotive
  - Healthcare
  - ...
- Data-intensive scenarios
- Distributed data sources
- Latency critical tasks





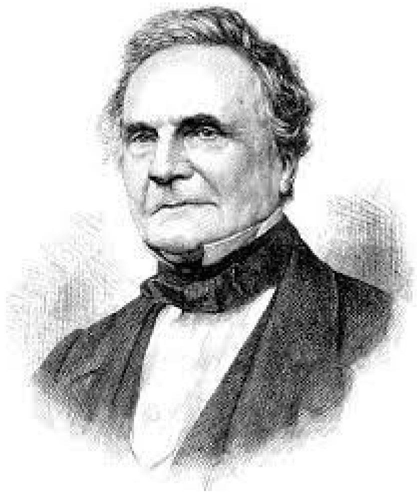
Fog infrastructure

- Cloud computing may not be suitable
  - High cost for data transfer → problem with huge data
  - High latency → not suitable for latency-bound applications
- Fog infrastructures
  - Close to end users
  - Distributed
- Can host
  - latency-critical tasks (e.g., autonomous driving)
  - Data aggregation and filtering (reduce data volume)



Quest for the right model

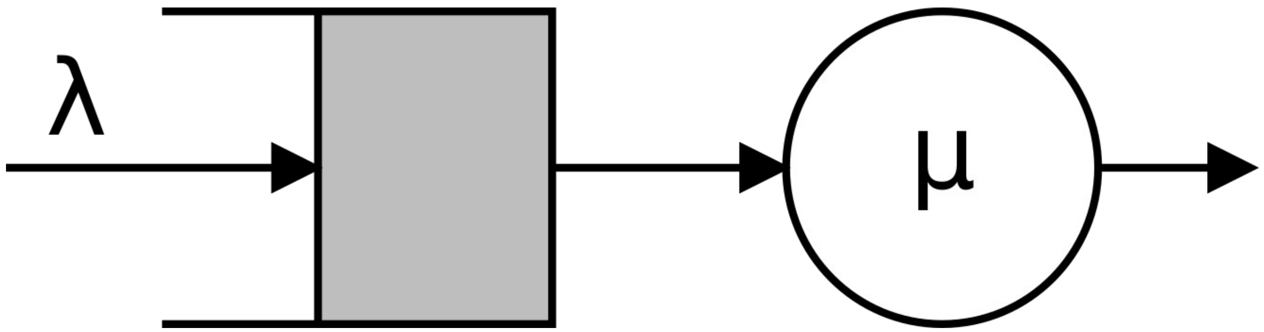
- Problem of performance evaluation in complex systems
  - Need accurate performance models
  - Inaccurate model leads to inaccurate evaluations
- Problem common to Fog and Cloud
- Actually the problem is much older…



“ I have been asked: ‘Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.”

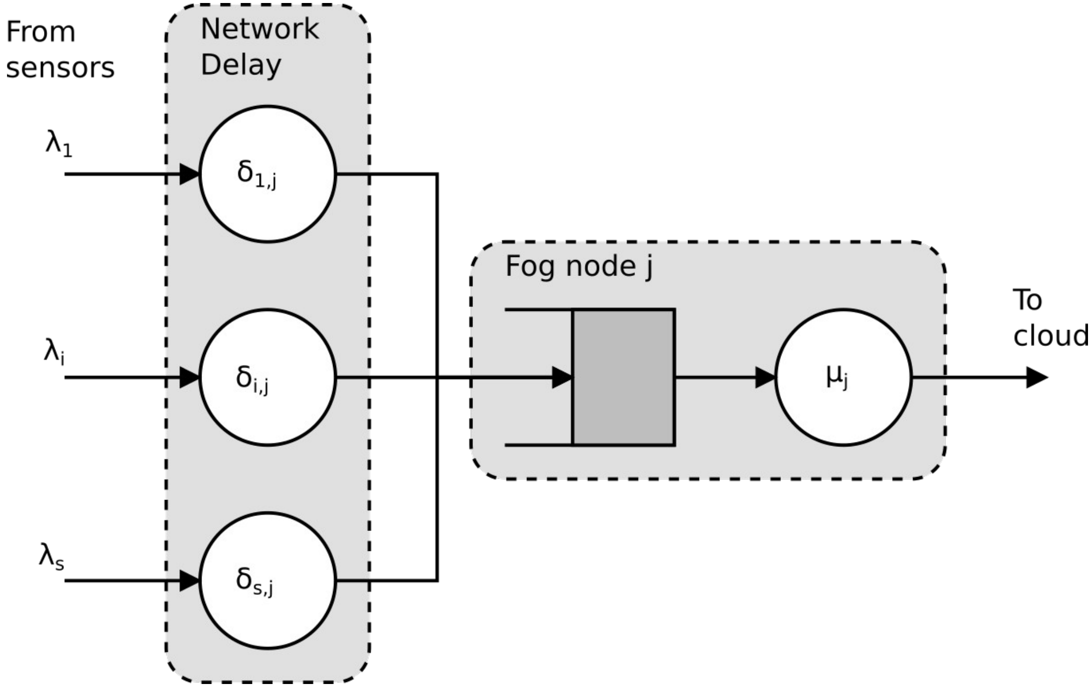
Quest for the right model

- Issues of Fog infrastructures
  - Significant network delays
  - Limited resources @ fog nodes
- Trade-off locality/load-balancing
- Typical approach (also in cloud computing)
  - Queuing theory
- Simplest model: M/M/1
  - Very simple: no parameters
  - Can be inaccurate!



Fog model overview

- Fog model overview
  - Main performance metric: response time
  - Two contributions: network delay and processing time
  - Still using queuing theory (not limited to M/M/1)



Optimization problem

- Two objective functions
  - Infrastructure cost
  - Response time
- SLA constraint $$T_{SLA} = K \cdot \frac{1}{\overline{\mu}} + \overline{\delta}$$
- Response time model
  - Network + Processing
  - M/M/1 or M/G/1 models

$$T_P = \frac{1}{\mu - \lambda}$$

$$T_P = \frac{1}{\mu}\left(1 + \frac{1 + \mathrm{CoV}^2}{2} \cdot \frac{\rho}{1-\rho}\right)$$

- Decision variable
  - Mapping of sensor→ fog data flows
  - Enabling of fog nodes
- Formalized problem

Minimize:

$$C = \sum_{j \in \mathcal{F}} c_j E_j$$

$$T_R = T_N + T_P$$

Subject to:

$$T_R \leq T_{SLA}$$

$$\lambda_j \leq E_j \mu_j, \quad \forall j \in \mathcal{F}$$

$$\sum_{j \in \mathcal{F}} x_{ij} = 1, \quad \forall i \in \mathcal{S},$$

$$E_j \in \{0,1\}, \quad \forall j \in \mathcal{F}$$

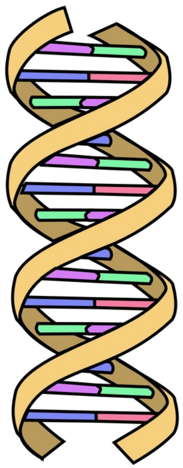
$$x_{ij} \in \{0,1\}, \quad \forall i \in \mathcal{S}, j \in \mathcal{F}$$

Infrastructure scaling

- Solving the problem
  - Estimation of minimum number of fog nodes N
  - Solution of problem with N nodes
  - In case of infeasibility, increase N and iterate
- Two approaches for N estimation (M/M/1 or M/G/1)

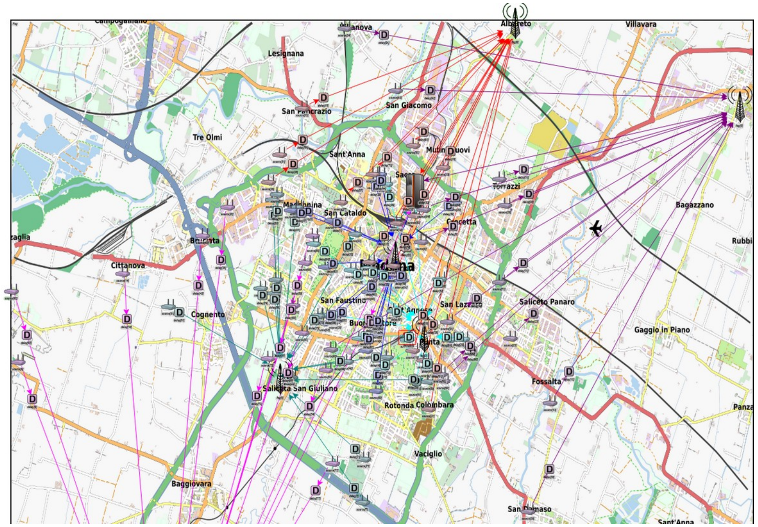
$$N = \sum_{j \in \mathcal{F}} E_j \geq \left\lceil \frac{\Lambda}{\mu} \cdot \frac{K-1}{K} \right\rceil \qquad N \geq \left\lceil \frac{\Lambda}{\mu} \cdot \frac{\mathrm{CoV}^2 - 2K - 1}{2K - 2} \right\rceil$$

- Solution based on heuristics
  - Genetic algorithm
  - Can use both M/M/1 or M/G/1 model for performance estimation



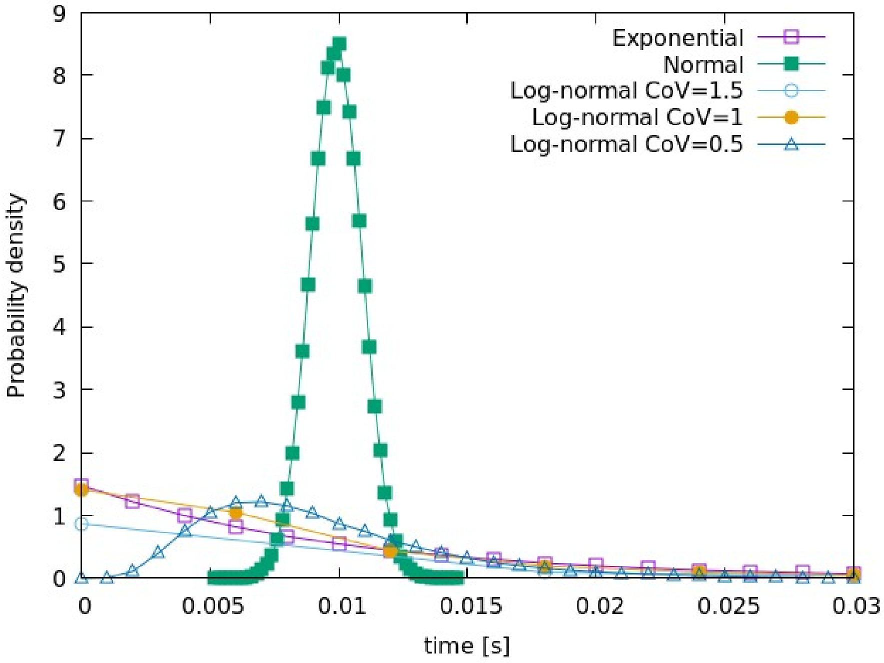
Simulation model

- Simulation based on Omnet++ simulation framework
- Integration with OSM
  - Fog nodes are geo-referenced
- Smart-city application scenario
  - Network delay increase with sensor-fog distance
  - LoRa-WAN model
  - Network delay comparable with processing time
- Simplified assumption
  - Homogeneous nodes
  - Easy to extend



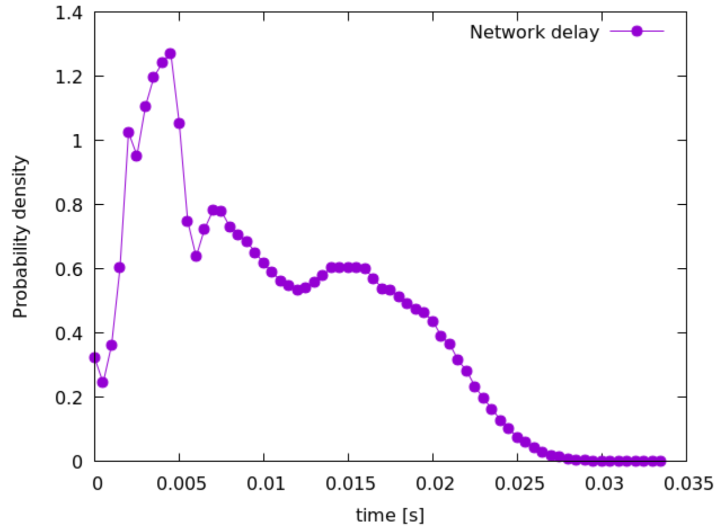
Simulation model

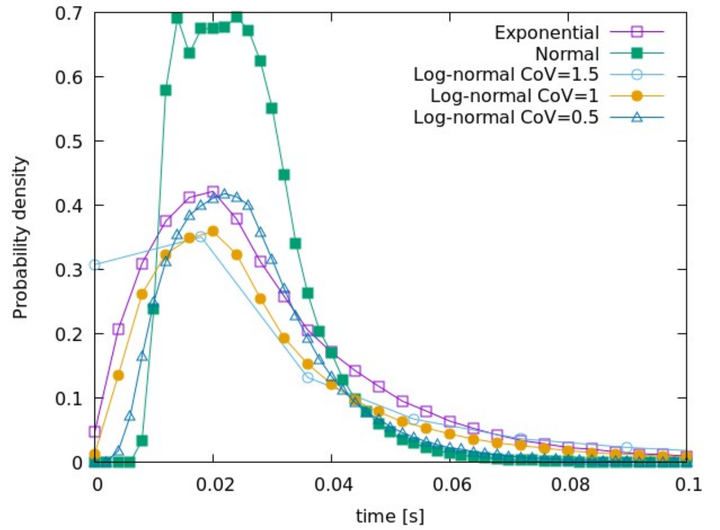
- Several models for service time
  - Exponential
  - Normal (Gaussian)
  - Log-normal
    CoV = [0.5, 1.0, 1.5]
- Medium-high load scenario
  - Same number of fog nodes
  - Load ρ=0.8
- Use of GA to map data flows
  - Load balancing
  - Locality of access



Response time

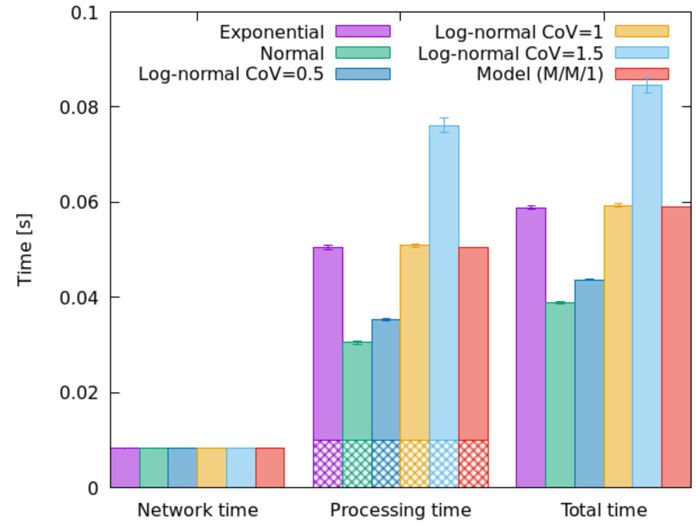
- Probability Density Functions
  - Use of histograms
- Network delay
  - Depends only on sensor→fog mapping
  - Same for every scenario
- Response time
  - Depends on service time model
  - In the following main focus on average values

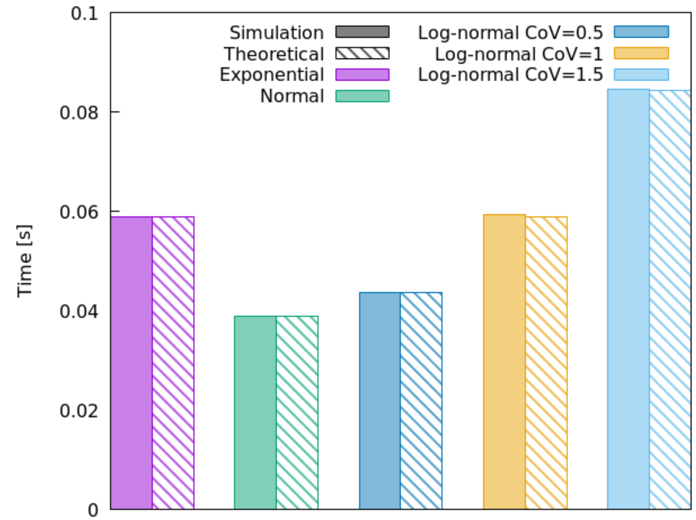




Average response time

- Comparison of average response time
  - Breakdown of components
  - Same time for network and service
  - Impact of queuing time
  - Depends on service time variance
  - Poor fitting of M/M/1 model
- Pollaczek Khinchin formula to predict response time
  - M/G/1 provides good fitting



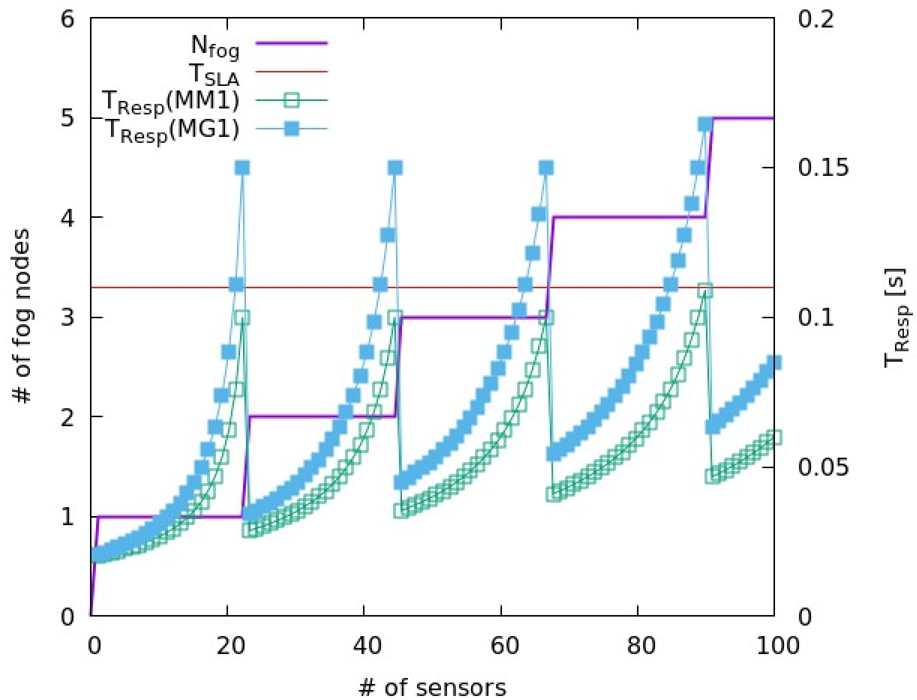


Infrastructure scaling

- Use of wrong model can affect infrastructure scaling
- Number of fog nodes based on sensors (purple line)
  - Use of M/M/1 model

$$N = \sum_{j \in \mathcal{F}} E_j \geq \left\lceil \frac{\Lambda}{\overline{\mu}} \cdot \frac{K-1}{K} \right\rceil$$

- Response time based on M/M/1 model (green line)
  - No SLA violations
- Service time is more skewed M/G/1 with σ>1(blue line)
  - SLA violations occur

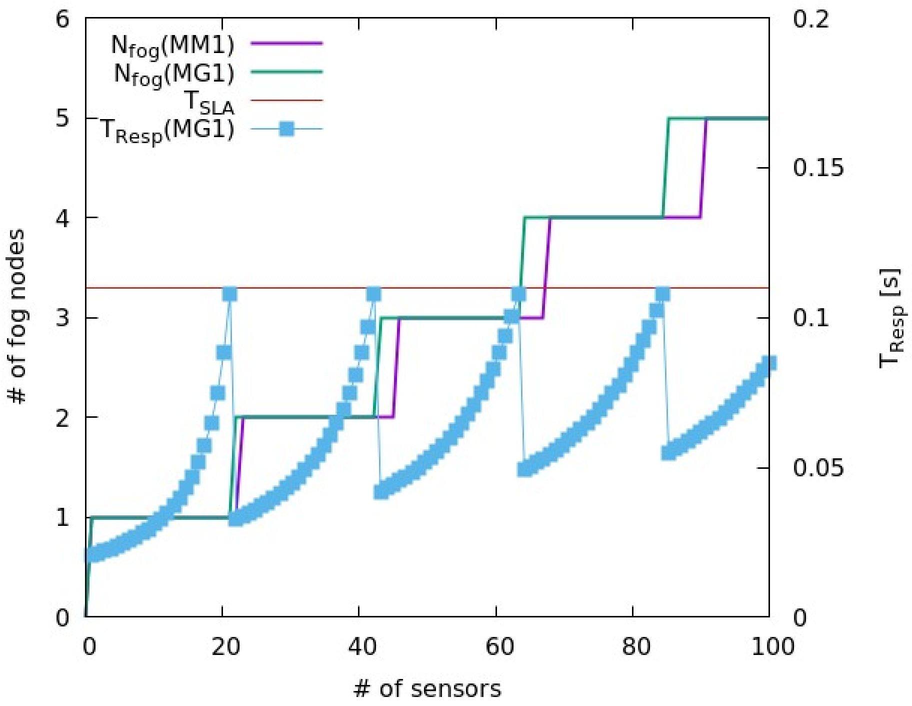


Applying the right model

- Adoption of the right model
- Infrastructure scaling based on M/G/1 model

$$N \geq \left\lceil \frac{\Lambda}{\overline{\mu}} \cdot \frac{\mathrm{CoV}^2 - 2K - 1}{2K - 2} \right\rceil$$

- Aggressive increase in number of fog nodes
  - Compare green and purple line
- No more SLA violations



Conclusions

- Key role of Fog computing in modern applications
- The need for accurate performance models in Fog systems
- Case study based on a smart-city application
  - Theoretical models (queuing networks)
  - Simulation
- Impact of over-simplifying:
  - Error in response time estimation (up to 50%)
  - Wrong infrastructure scaling (SLA violations)

Further questions to...

Claudia Canali
claudia.canali@unimore.it

Riccardo Lancellotti
riccardo.lancellotti@unimore.it

Stefano Rossi
stefano.rossi@unimore.it