Séminaire STORE 2019

Optimal Capacity of Fog Computing Infrastructures under Probabilistic Delay Guarantees

I. Stypsanelli - **O. Brun** - B.J. Prabhu - S. Medjiah

LAAS-CNRS, Toulouse, France

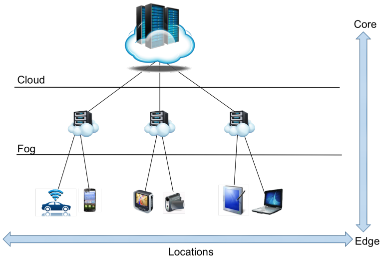
Outline

Fog Computing

Fog Computing: benefits and threats

Fog Computing

- ✔ Computing, networking and storage resources close to users.
- ✔ Connected vehicles, augmented reality, smart cities, etc.



Expected benefits

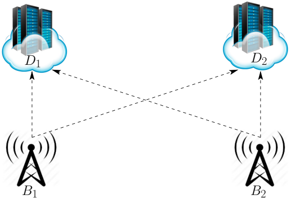
- ✔ Reduced latency, preservation of network resources, greater security, privacy and resilience, as well as easier scalability.

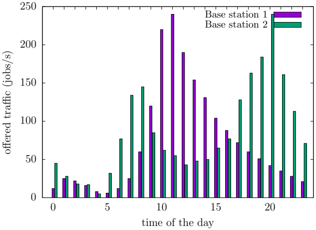
Threats

- ✔ Duplication of distributed resources may lead to an explosion of capacity, energy and operation costs

Geographic diversity vs data-centre sizes

Example





- ✔ Fully distributed solution: minimum latency, but provisioned for $240 + 240 = 480$ jobs/s.
- ✔ Centralized solution: higher latency, but provisioned only for 282 jobs/s.

Trade-off between geographic diversity and data-centre sizes

Capacity planning of micro data-centres

Decisions

- ✔ Where to place micro-datacentres? How big to make them?
- ✔ How user-generated requests are routed to these data-centres?

Objective

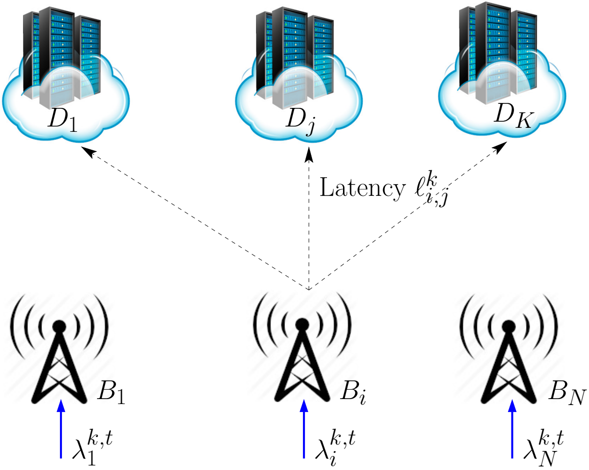
- ✔ Minimize infrastructure cost under probabilistic delay guarantees

Formulation as a Mixed Integer Linear Programming (MILP) problem

- ✔ Greenfield design or brownfield design

Mathematical model

Input Data

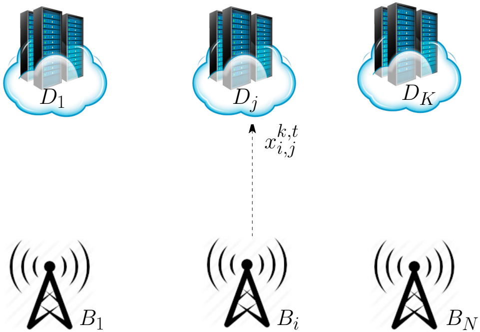

D_1
D_j
D_K
Latency $\ell_{i,j}^k$
B_1
B_i
B_N
$\lambda_1^{k,t}$
$\lambda_i^{k,t}$
$\lambda_N^{k,t}$

Routing variables

- ✔ $x_{i,j}^{k,t}$ amounts of class-k traffic from BS i to DC j at time t

$$\sum_j x_{i,j}^{k,t} = \lambda_i^{k,t}, \quad x_{i,j}^{k,t} \geq 0$$

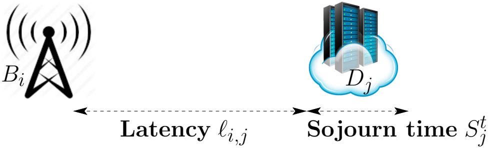
- ✔ Binary variables $a_{i,j}^{k,t} = 1$ if $x_{i,j}^{k,t} > 0$, and 0 otherwise



Other variables

- ✔ Choose whether site j is selected ($u_j = 1$) or not ($u_j = 0$)
- ✔ Choose the capacity c_j in DC j such that

$$\mathbb{P}\left(S_j^t + \ell_{i,j} \geq T\right) \leq \delta, \quad \forall t$$



Problem Formulation

$$\text{minimize} \sum_{j \in \mathcal{D}} (\beta_j \, u_j \; + \; g_j(c_j))$$

s.t

$$\mathbb{P}\left(S_j^{k,t} + \ell_{i,j}^{k} \geq T_k\right) \leq \delta_k,$$

$$\sum_{j \in \mathcal{D}} x_{ij}^{k,t} = \lambda_i^{k,t},$$

$$\ldots$$

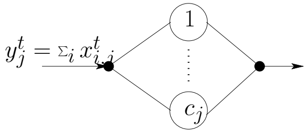
$$x_{ij}^{k,t} \geq 0,$$

$$u_j, a_{i,j}^{k,t} \in \{0, 1\},$$

Queueing model

- ✔ c_j parallel M/M/1 queues

$$\mathbb{P}\left(S_j^t \geq z\right) = e^{-(\mu - y_j^t/c_j)\, z}$$



- ✔ The latency constraint of jobs can be satisfied at site j iff

$$\ell_{i,j} a_{i,j}^t < T - \frac{\log(\frac{1}{\delta})}{\mu}, \quad i \in \mathcal{B}, t = 1, \ldots, \tau$$

- ✔ Optimal capacity at data center j

$$c_j \geq \frac{y_j^t}{\mu - d_{i,j}} - M\left(1 - a_{i,j}^t\right), \qquad (1)$$

$$c_j \geq 0, \qquad (2)$$

where M is a large constant and $d_{i,j} = \log(\frac{1}{\delta}) / \left[T - \ell_{i,j}\right]$.

Objective function

Linear objective function

$$\text{minimize} \sum_{j \in \mathcal{D}} (\beta_j\, u_j \;+\; \alpha_j\, c_j) \qquad \text{(CAPA-PL)}$$

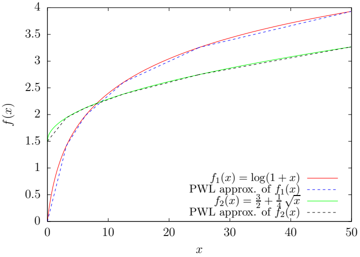
subject to previous linear constraints .

Concave objective function (economies of scale)

$$\text{minimize} \sum_{j \in \mathcal{D}} (\beta_j\, u_j \;+\; g_j(c_j))$$

s.t. linear constraints .

✔ Piecewise linear approximation



Experimental Results

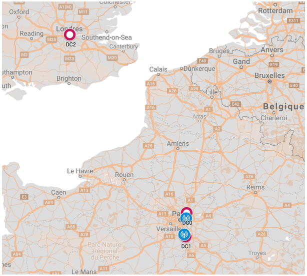
Experimental Results

Simple Scenario

- ✔ 2 private data centres , 1 big public cloud, and 2 base stations.

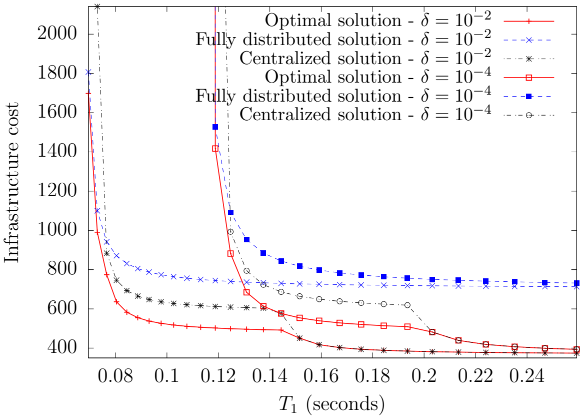
$$100 \times (u_1 + u_2) + c_1 + c_2 + \frac{3}{4} \times c_3$$

- ✔ Real-time jobs (variable offered traffic) and best-effort jobs (constant offered traffic)



Experimental Results

Simple Scenario – Results



Experimental Results

Simple Scenario – Results

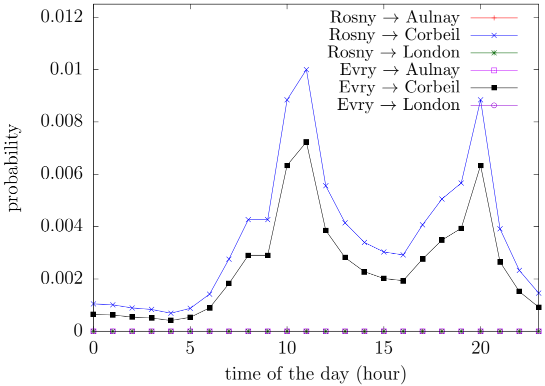
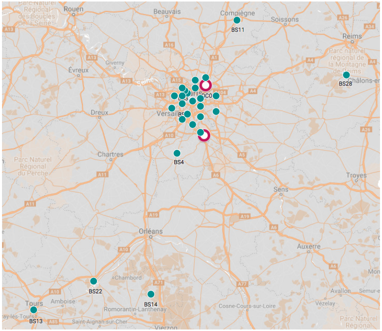


Figure: Probability that the end-to-end delay in the optimal solution be greater than $T = 100$ ms when $\delta = 0.01$.

Experimental Results

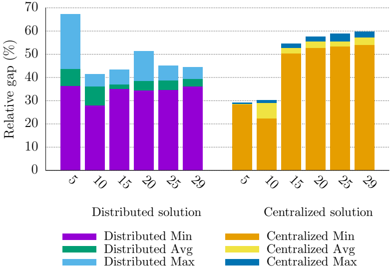
Larger number of base stations

- ✔ Same potential data centres, but 29 base stations.
- ✔ Real-time jobs with $T_1 = 105$ ms and $\delta_1 = 0.01$ and best-effort jobs
- ✔ 1^{st} scenario = 5 first base stations, 2^{nd} scenario = 10 first base stations, etc.
- ✔ 16 randomly generated problem instances for each scenario using a spatio-temporal traffic model

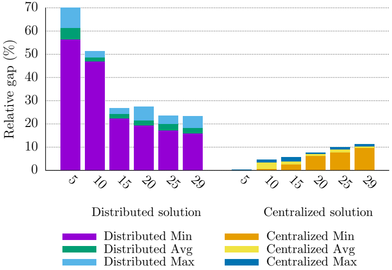


Experimental Results

Larger number of base stations



Linear objective function



Logarithmic objective function

Conclusion

Conclusion

Optimal capacity-planning of micro data centres as a MILP problem

- ✔ Can be solved efficiently even for large-size problem instances
- ✔ Significant cost savings can be obtained w.r.t. heuristic solutions

Future work

- ✔ Resource sharing between job classes (e.g., strict priority mechanism),
- ✔ General distribution of job service times (analytical approximations),
- ✔ Advanced load-balancing policies (e.g., Power of Two Choices or Join the Shortest Queue).

Questions ?