Data centre localization for Internet services

HIBAT ALLAH OUNIFI[1], MUSTAPHA OUHIMMOU[1], MARC PAQUET[1], JULIO MOMTECINOS[1]

[1]ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
1100 RUE NOTRE-DAME OUEST, MONTRÉAL, QC H3C 1K3
HIBAT.ALLAH.OUNIFI@GMAIL.COM

***Abstract*- The study of data centre location is an active research area. Popular Internet services are hosted in geographically distributed data centres and their locations have a direct impact on Internet services response time and costs. The selection of data centre location takes into account several criteria including proximity to market areas, energy sources in the region, land costs, electricity, workforce and the average temperature in the region (related to cooling of equipment). Our objectives are to minimize data centre costs and to calculate the power usage effectiveness metric. Our project introduces an efficient model based on mixed integer linear programming (MILP) for the data centre localization problem. In addition, we apply the declining balance method in order to calculate the amortization costs of selected data centres. The proposed model is used to locate data centres under Internet services demands and, at the same time, ensure a low energy consumption of information technology (IT) equipment. Model validation and results are proven via experimentation example and numerical study.**

***Keywords*– Data centre, response time, latency delay, Internet services, Power usage effectiveness, declining balance method.**

1 INTRODUCTION

Today, most popular Internet companies such as Google ©, Yahoo © and Microsoft © offer a range of services to millions of users every day. These services are hosted in data centres. Also, the need for large storage memory, infrastructure and data processing is a necessity. Indeed, data centres have an axial point on which all the data and users applications are deployed. Also, the centralization of this data in one place makes the management and security of data easier.

In addition, in data centre studies, there are some points that are important to evaluate. Those points are: data centre topology, network connectivity, latency time, costs per megawatts (MW), data centre service response time and the use of clean energy. (Daim et al., 2012) discussed two essential purposes of a data centre: 1) ensure data security and 2) guarantee continuity of company operations. Similarly, they claim that the security of the data centre is manifested in two ways: 1) physical security of the data centre from natural and human disasters and 2) business continuity of data centre while planning the recovery of centres after disasters. In the latter case, we can speak about recovery data centre. Many companies such as IBM®, Amazon® and Microsoft® offer many services like database services and Web services to millions of users. These services are stored in data centres that are distributed geographically. The main needs of data centre users are high availability, low response time and good quality of service (QoS). To maintain these needs, we assume that each data centre has another data centre as a backup at a distinct location. In addition, (Covas et al., 2013) indicate that companies want to build data centres in locations that are not prone to natural disasters, terrorism and also offer less expensive resources such as electricity, network and easy access to transport. At the same time, they need a robust system that can operate 24/7/365 without interference and provides scalable computing power. Similarly, the response time, the availability and the costs depend strongly on the specific location of data centres.

IT data centres represent a huge investment and most companies have to set up their own data centres to run business operations, store data or provide services to clients. Indeed, the establishment of a new data centre requires a long-term effort, millions of dollars and months of planning. In the same context,(Greenberg et al., 2009) indicate that data centres are very expensive and the cost can mean putting out 12 million dollars per month for a data centre with 100,000 servers.
Today, site selection is also a key element in the data centre network design problem. The location of a data centre presents the strategic setting to determine the total cost of ownership and, to some extent, life span of a data centre. Thus, companies can't afford to make a bad decision when designing and implementing a new data centre because the planning must be done for around

20 years ahead of time. For this reason, the definition of data centre selection criteria is the most basic task to be done before setting up and choosing a site for a new data centre.

In this paper, we study Internet data centre location in the province of Quebec. First of all, we propose a mixed integer linear programming (MILP) model to select data centre locations and sizes. This model defines costs minimization objective with service response time, latency delay and power usage effectiveness (PUE) as main operational constraints. Our goals is to select data centre location when calculating total cost and calculate the power usage effectiveness (PUE) metric in order to validate the energy effectiveness of new data centre.

This paper is organized as follows. Section 2 presents some related works and gives some examples of data centre budgets and costs. Based on this literature review, we give an overview of data centre location criteria and highlight the advantages of the province of Quebec as a potential location to set up data centres. Section 3 explains in detail the mathematical model and proposed approaches. Section 4 gives the numerical results. Finally, Section 5 states our conclusions and research perspectives.

2 LITERATURE REVIEW

Several large companies in the world define common criteria for establishing a data centre.

Before setting up a data centre, all user needs should be met while ensuring high service availability, data security, good response time and a fast data transfer. These goals will be achieved while defining from the outset the right location of a data centre. Similarly, the definition of criteria for the selection of good data centre location differs according to business priorities. (Covas et al., 2013) present a multi-criteria decision analysis to identify the best location of a data centre, taking into account economic, social, technical and environmental dimensions. These dimensions will be evaluated as a problem of multi-criteria sorting. The problems will be analyzed by using the method of ranking "ELECTRE TRI". Indeed, the evaluation of the selection criteria to place data centre differs from one dimension to another: in fact, for the environmental evaluation, energy economy is measured in terms of quantity of CO2 emitted, while for the economic evaluation, energy saving is measured in monetary units.

In another paper, (Daim et al., 2012) define four factors to select data centre locations: geographical factor, cost factor, political factor and social factor. Indeed, the choice of site location should consider, according to Sun Microsystems, natural hazards (storms, earthquakes), human risks (industrial pollution, vibration) and availability of public services. In addition, (Rath, 2007) presents many other criteria to choose the potential location of data centres such as workforce qualification, land cost, green power and insurance. Also, (Covas et al., 2013, Lei et al., 2010) indicate that data centres are characterized by high energy consumption. When a company decides to open a data centre, the first decision variable is the energy cost. Similarly, data security and reliability of data centre are connected to the location of data centre. In our case, we choose the geographical factor and cost factor by selecting the province of Quebec as a strategic position to set up data centres. Thus, the province of Quebec has energy at competitive prices, renewable energy, and suitable climate for cooling, qualified workforce, robust telecommunications infrastructure and proximity to the US market.

Our aim is to define; first of all, site location criteria, then to minimize the total costs (including fixed and variable costs). In this context, (Goiri et al., 2011) present an optimization framework to choose the potential location of data centres by respecting response time and service availability constraints. But, there are some gaps: first, they do not specify the nature of the Internet services. As for Internet services, there are many services such as Web service, mailing service, computer processing service, Video conference service... For each service, there are different response times. In addition, this response time depends on the demands of users. In addition, they do not calculate the power usage effectiveness (PUE) metric.

Indeed, energy consumption is one of the most important criteria when companies plan to build data centres. Similarly, (Pakbaznia et Pedram, 2009) reported that the energy consumption of servers and data centres has doubled over the past five years and is expected to quadruple over the next five years to more than 100 billion kWh an annual cost of about $ 7.4 billion. In this case, (Raghavendra et al., 2008), (Yao et al., 2012),(Lei et al., 2011) and (Yuanxiong et al., 2011) present power management solutions in order to decrease the energy consumption of servers and electricity costs. In this context, (Uptime Institute 2012) mentions that the PUE is a metric used to define energy efficiency of data centres. Therefore, (Abbasi et al., 2012) show that calculating the power usage effectiveness (PUE) is important to define the energy used by IT equipment of data centres.

In addition, (Hong et Baochun, 2012) develop an optimization system that takes into consideration the use of bandwidth and the cost of electricity to solve the problems of data centre selection for cloud computing services. Cloud computing services are distributed in geographically distributed infrastructure. Also, data centres are located in different regions to provide more efficient services. For this reason, service providers need a method to redirect customers to the appropriate data centre. Indeed, (Hong et Baochun, 2012) showed that the traditional method for selecting data centre uses the DNS servers. Each geographic area has a different electricity cost. Similarly (Hong et Baochun, 2012) show that the geographic distribution of data centre has an advantage. Indeed, this distribution is related to a diversity of electricity prices. In this case, client requests will be redirected to data centres whose electricity cost is reduced. The authors indicate that other than the electricity cost, there is the bandwidth cost. The method used by (Hong et Baochun, 2012) is the multiplexing of bandwidth in order to reduce consumption of bandwidth and cost. In this context, (Hong et Baochun, 2012) have developed an optimization model cost. This model takes into consideration the cost of electricity and bandwidth. In addition, this model is the basis of an optimization algorithm that aims to select the optimal allocation of data centre nodes. The optimal allocation is based on the sub-gradient method. To solve the optimization problem defined as a linear program ((Hong et Baochun, 2012) using Lagrangian relaxation. Indeed, they show that the optimization function data centre "DC-OPT" can be decomposed into N data centres by maximizing problems. In this case, data centres are randomly selected and assigned a token with a total budget B. Indeed, the data centre marks the token and sends it to the next data centre which in turn updates the budget that remains. Similarly, the DC-OPT optimization function reduces the use of the bandwidth of most data centres 15 to 20%.

Also, (Lei et al., 2010) describe the electricity cost minimization problem for data centres offering Internet services. In fact, the minimization problem focuses on a multi-electrics market environment. The concept of multi-electrics market means that the cost of electricity varies in function of place and time. This creates a cost problem for data centres that should operate 365/7/24.

So, the objective of (Lei et al., 2010) is to minimize data centre costs while ensuring quality of service tailored to the diversity of electricity prices based on diversity of place and time of operation of data centres. The proposed model is a mixed linear program. Indeed, the total cost of localization of *N* data centres is given by the following function:

$$T_{total} = \sum_{i=1}^{N} m_i \mathrm{Pr}_i(t) Po_i$$

This function is the multiplication of the number of functional servers m_i by the energy needed for the operation of a server Po_i multiplied by the electricity cost $Pr_i(t)$ in the location i.

In addition, (Yuanxiong et al., 2011) present the same problem of electricity minimization cost proposed by (Lei et al., 2010), but the resolution of this problem is made through the optimization technique of "Lyapunov". Indeed, the proposed algorithm solves the problem of optimal distribution of traffic and loading/unloading of the battery energy data centres offering Internet services.

In the same context, (Jie et al., 2012) and (Yao et al., 2012) describe the cost minimization problem of electrical energy in data centres while providing a mathematical model of a mixed integer linear program (MILP). This solution ensures optimal management of the electricity demand of data centres. Moreover (Yao et al., 2012) present the management of energy by ensuring a good allocation of the energy demands (a distribution of applications on demands). Also, some related works have talked about data centre location, but in terms of operational cost and likelihood of natural disasters such as solutions of (Oley, 2010) and (Stansberr, 2006). However, those works haven't considered the optimization of all costs or relevant constraints. For example, none of them considers the energy efficiency of a data centre and the determination of the energy consumed by IT equipment.

Based on these related works, we follow an approach which combines two approaches: 1) minimize the total costs of data centre setup and 2) minimize the energy consumption. For the latter, especially the IT equipment energy such as servers by selecting the right size of data centre and assigning a specific number of hosted servers for each area zone (population centre). With the growing demand for Internet services in recent years, the energy consumed by data centres, which is related directly to the number of hosted servers, is experiencing a remarkable increase. In the same context, (Raghavendra et al., 2008) indicate that the total data centre power and cooling energy costs in the world was over $30 billion in year 2008. In another context, (Van Heddeghem et al., 2014) indicate that the total worldwide electricity consumption increased from 3.9% in 2007 to 4.6% in 2012.

To achieve the objectives stated previously, we should respect the Internet services response time and latency delay constraints. Our second contribution is the calculation of the energy efficiency (PUE) for each data centre and in each potential location, which makes our problem a complex combinatory problem.

3 MODEL FORMULATION

3.1 Model description

In this section, we summarize the model to capture the behaviour of data centre location for Internet services in the province of Quebec. It describes the mixed integer linear programming (MILP) model that considers the critical aspects for the design and selection of data centre locations and sizes while minimizing the total costs of setup, and determines the energy efficiency (PUE) metric. Indeed, companies take into account the value of the PUE while selecting locations of data centres. In this project, we chose in advance, eight regions in the province of Quebec to be the possible location to set up data centres. These regions are Montréal, Québec, Laval, Gatineau, Trois-Rivières, Longueuil, Saint-Jérôme and Terrebonne. Also, we selected the area zones to be served later by located data centres: Anjou, LaSalle, Verdun, Montréal-Nord, Plateau-Mont-Royal, Sherbrooke, Québec, Saguenay, Lévis, Laval, Gatineau, Outaouais, Trois-Rivières, Montérégie, Laurentides, and Lanaudière. These area zones are characterized by high population density (the largest cities in the province of Quebec.

Services response time: As previously mentioned, Internet services demands are increasing rapidly. To maintain the high availability of services, we should take into account, in the mathematical model, the services response time. In this case, we define h_{la}^{rep} parameter which presents the response time of a location *l* to an area zone (population centre) *a*. In more detail, for each Internet service, there is a different response time. Indeed, this parameter depends on the type of service (e.g., Web service needs less response time than the communication service).

Latency delay*:* One of the most important objectives of Internet data centre is to maintain the service availability for potential users. In this case, we think of the latency delay parameter that is defined when a data centre becomes unreachable or unavailable. That means that the active data centre becomes a backup centre for unavailable data centres. The $h_{ll'}$ parameter is calculated as a function of distance from the unavailable or unreachable data centre and a backup data centre. This parameter should be less or equal to the maximum latency delay value.

Power usage effectiveness (PUE): As mentioned above, Internet data centres consume a high amount of energy. With the increase of services demands, the numbers of hosted services will increase and consequently the energy consumed by servers. Thus, in selecting location, it is important to calculate the PUE value. In our mathematical model, we define the PUE as constraint to help us to choose the right size of data centre and affect later, for each population centre the required number of hosted servers and backup servers too. The value of PUE is calculated as follows:

$$\frac{electricity \times weighting\ factor + DistrictChilledWater \times weighting\ factor}{IT\ Energy}$$

3.2 Sets and indices

In this mathematical model, the following sets and indices are used:

$l \in L$	Data centre location.
$a \in A$	Population centre.
$d \in D$	Size of data centre.
$s \in S$	Provided Internet services.
$r \in R$	Set of resources that each server has, which include 1 = CPU, 2 = Network card, 3=RAM, 4 = Storage, 5 = FibreOptic card.

3.3 Parameters

To formulate the model, the following costs are required:

c_{dl}^{maint}	Maintenance cost of data centre having a size $d \in D$.
c_r^{server}	Server cost with resources $r \in R$.
c_d^{admin}	Administration cost of a data centre having a size $d \in D$ ($/administrator).
c_d^{erergy}	Energy cost for each data centre having a size $d \in D$.
c_{dl}^{land}	Land cost for a data centre in the location $l \in L$ and having a size $d \in D$.
$c_{dl}^{cooling}$	Cooling cost for a data centre in the location $l \in L$ and having a size $d \in D$.
c^{BW}	Cost of required bandwidth for each server providing a service $s \in S$.
c^{switch}	Switch cost ($/switch).
$c^{FiberOptic}$	Fibre-optic cost ($/ Km).

The following data are also required:

$d_{la}^{location}$	Distance between location $l \in L$ and population centre $a \in A$.
n_d	The number of administrators required to manage data centre having a size $d \in D$.
w_d	The amount of energy required to power a data centre having a size $d \in D$.
$h_{ll'}$	Latency between two potential locations $l \in L$ and $l' \in L$.
h_{la}^{rep}	Response time of service $s \in S$ for a population centre $a \in A$.
$f_d^{\max}$	Maximum capacity of a data centre in number of IT equipment (servers, Switches) for a data centre, which has a size $d \in D$.
δ_a	Number of hosted servers to serve a population centre $a \in A$.
k_{la}	Maximum response time of a location $l \in L$ for a population centre $a \in A$.
p^{switch}	Incoming power to serve a switch.
p^{server}	Incoming power to serve a server.
m	Maximum latency delay.
β	Weighting factor for water energy.
λ	Maximum power usage effectiveness (PUE)
α	Weighting factor for electric energy.
μ	Switch incoming ports for servers.
Qt_a^{bw}	The bandwidth quantity for a population centre $a \in A$.
ω	The depreciation coefficient.
t	Data centre and equipment lifetime (t=12 years for data centre lifetime, t=4 for IT equipment lifetime).

3.4 Decision variable

In order to find the optimal data centre location in the province of Quebec, the following decision variables are required:

Y_l^{center}	Binary variable equals to 1 if a data centre is located in $l \in L$; 0 otherwise.
X_{la}^{backup}	Binary variable equals to 1 if a backup data centre is located in $l \in L$ to serve the population centre $a \in A$; 0 otherwise.
S_{la}^{backup}	The total number of servers of a backup data centre located in $l \in L$ and serving a population centre $a \in A$.
W_{dl}^{TI}	The IT energy consumption for a data centre of size $d \in D$ located in $l \in L$.
Y_{dl}^{size}	Binary variable equals to 1 if data centre located in $l \in L$ has as size $d \in D$; 0 otherwise.
X_{la}^{center}	Binary variable equals to 1 if data centre located in $l \in L$ serves the population centre $a \in A$; 0 otherwise.
S_{la}^{explot}	The total number of servers for a data centre located in $l \in L$ that serves the population centre $a \in A$.
N_{la}^{switch}	The number of switches to connect servers for a data centre located in $l \in L$ which serves the population centre $a \in A$.

3.5 Optimization model

3.5.1 Approach 1: Minimize total costs

Using these notations, a mixed integer linear programming (MILP) model can be formulated to select Internet data centre locations and sizes. The total data centre costs are segmented in six costs.

The fixed cost of a data centre (land cost and cooling cost) can be expressed as:

$$\sum_{d \in D} \sum_{l \in L} \left(c_{dl}^{land} + c_{dl}^{cooling} \right) Y_{dl}^{size} \tag{1}$$

The variable cost of a data centre (administration cost and maintenance cost) can be expressed as:

$$\sum_{d \in D} \sum_{l \in L} \left(n_d c_d^{admin} + c_d^{maint} \right) Y_{dl}^{size} \tag{2}$$

The Information technology equipment (IT) cost of a data centre (servers and switches) can be expressed as:

$$\sum_{l\in L}\sum_{a\in A} c_1^{server}(S_{la}^{\exp lot}+S_{la}^{backup})+c^{switch}N_{la}^{switch} \quad (3)$$

The total energy cost can be expressed as:

$$\sum_{d\in D}\sum_{l\in L} w_d\, c_d^{energy} Y_{dl}^{size} \quad (4)$$

The total bandwidth cost can be expressed as:

$$\sum_{l\in L}\sum_{a\in A} Qt_a^{bw}\, c^{BW} S_{la}^{\exp lot} \quad (5)$$

The interconnection of data centres to population centres cost can be expressed as:

$$c^{fiberOptic}\sum_{l\in L}\sum_{a\in A} d_{la}^{location}\; X_{la}^{center} \quad (6)$$

The objective-function (1) must meet the following constraints.

Each data centre serves at least one population centre:

$$\sum_{l\in L} X_{la}^{center}\geq 1, \forall a\in A \quad (7)$$

Defines if data centre is opened or not:

$$\sum_{d\in D} Y_{dl}^{size}\leq 1, \forall l\in L \quad (8)$$

Data centre must exist to serve a population centre:

$$X_{la}^{center}\leq \sum_{d\in D} Y_{dl}^{size}, \forall l\in L, a\in A \quad (9)$$

Each population centre is served by a backup data centre:

$$\sum_{l\in L} X_{la}^{backup}\geq 1, \forall a\in A \quad (10)$$

An informative constraint that contains an information variable Y_l^{center}:

$$\sum_{d\in D} Y_{dl}^{size} = Y_l^{center}, \forall l\in L \quad (11)$$

A backup data centre exists to serve a population centre; this backup data centre has an identical characteristic to a data centre:

$$\sum_{d\in D} Y_{dl}^{size}\geq X_{la}^{bakup}, \forall l\in L, a\in A \quad (12)$$

The data centre and its backup are not in the same location:

$$X_{la}^{center}+X_{la}^{backup}\leq 1, \forall l\in L, a\in A \quad (13)$$

Demand constraints:

$$\delta_a X_{la}^{center} = S_{la}^{\exp lot}, \forall l\in L, a\in A \quad (14)$$

$$\delta_a X_{la}^{backup}\leq S_{la}^{backup}, \forall l\in L, a\in A \quad (15)$$

$$\mu N_{la}^{switch}-1 < S_{la}^{\exp lot}, \forall l\in L, a\in A \quad (16)$$

$$\mu N_{la}^{switch}\geq S_{la}^{\exp lot}+S_{la}^{backup}, \forall l\in L, a\in A \quad (17)$$

Capacity constraints:

$$\sum_{a\in A}(S_{la}^{\exp lot}+S_{la}^{backup}+N_{la}^{switch})\leq \sum_{d\in D} f_d^{\max} Y_{dl}^{size}, \forall l\in L, a\in A \quad (18)$$

(0)

The data centre latency between two potential locations in order to determine the backup data centre:

$$h_{ll'}(X_{la}^{center}+X_{la}^{backup}-1)\leq m, \forall l\in L, l'\in L, a\in A, l\neq l' \quad (19)$$

The energy consumed by IT equipment with M as the maximum energy that can be consumed by IT equipment.

$$\sum_{a\in A}(p^{server} S_{la}^{\exp lot}+p^{switch} N_{la}^{switch}) = \sum_{d\in D} W_{dl}^{TI}, \forall l\in L \quad (20)$$

$$W_{dl}^{TI}\leq MY_{dl}^{size}, \forall d\in D, l\in L \quad (21)$$

Calculates the power usage effectiveness:

$$\alpha w_d Y_{dl}^{size}\leq \lambda\alpha w_{dl}^{TI}, \forall d\in D, l\in L \quad (22)$$

Service response time:

$$h_{la}^{rep} X_{la}^{center}\leq k_{la}, \forall l\in L, a\in A \quad (23)$$

Binary variables:

$$Y_{dl}^{size}\in\{0,1\}\ \forall d\in D, l\in L \quad (24)$$

$$Y_l^{center}\in\{0,1\}\ \forall l\in L \quad (25)$$

$$X_{la}^{backup}\in\{0,1\}\ \forall a\in A, l\in L \quad (26)$$

Integer and positive variables:

$$S_{la}^{explot}\geq 0 \text{ and Integer } \forall l\in L, a\in A \quad (27)$$

$$N_{la}^{switch}\geq 0 \text{ and Integer } \forall l\in L, a\in A \quad (28)$$

$$S_{la}^{backup}\geq 0 \text{ and Integer } \forall l\in L, a\in A \quad (29)$$

$$w_{dl}^{TI}\geq 0 \text{ and Integer } \forall d\in D, l\in L \quad (30)$$

Constraints description:

- Constraint (20) calculates the energy consumed by IT equipment in order to calculate in the next step, the power usage effectiveness. In addition, the backup servers could be included, but they are on stand-by and no energy is consumed.
- Constraint (21) ensures that W_{dl}^{TI} will indicate the selected size, with M as the maximum energy that can be consumed by IT equipment.
- Constraint (22) defines the power usage effectiveness, this constraint is a way to impose a minimum use of resources over the size, so it artificially increases the number of IT equipment.
- Constraint (23) defines the service response time that should be less than a maximum response time.

3.5.2 Approach 2: The declining balance method

Data centre costs assume a 12-year lifetime for data centres and a 4-year lifetime for servers. In this case, we calculate in the approach 2 the amortized cost. In this case, we admit that fixed and variable costs will be calculated for 12 years and energy, bandwidth, IT equipment and location costs will be calculated for only 4 years.

Indeed, the depreciation of a data centre begins from the commissioning of the property (when the data centre is working depending on the intended use).

In addition, the amortized cost consists of applying a tax coefficient at the linear rate (this coefficient was 1.25 for the amortization periods between 3 and 4 years, 1.75 for periods of 5 and 6 years and 2.25 for durations over 6 years). The additional depreciation resulting from the application of the declining balance must be accounted for as accelerated depreciation.

In our case, for IT equipment costs, bandwidth cost and energy cost this coefficient was 1.25. For fixed and variable costs, this coefficient was 2.25.

So the function F1 is a part of the objective function. This function calculates the total amortized cost for a data centre's lifetime.

$$
\begin{aligned}
Min\,F1 = {} & \omega\left(\frac{100}{t}\%\right)\sum_{d\in D}\sum_{l\in L}\left(c_{dl}^{land}+c_{dl}^{cooling}\right)Y_{dl}^{size} \\
& +\omega\left(\frac{100}{t}\%\right)\sum_{d\in D}\sum_{l\in L}\left(n_d c_d^{admin}+c_d^{maint}\right)Y_{dl}^{size} \\
& +\omega\left(\frac{100}{t}\%\right)\left(\sum_{l\in L}\sum_{a\in A} c_1^{server}\left(S_{la}^{explot}+S_{la}^{backup}\right)+c^{switch}N_{la}^{switch}\right) \\
& +\omega\left(\frac{100}{t}\%\right)\sum_{d\in D}\sum_{l\in L} w_d\, c_d^{energy}\, Y_{dl}^{size} \\
& +\omega\left(\frac{100}{t}\%\right)\sum_{l\in L}\sum_{a\in A} Qt_a^{bw}\, c^{BW}\, S_{la}^{explot} \\
& +\omega\left(\frac{100}{t}\%\right) c^{fiberOptic}\sum_{l\in L}\sum_{a\in A} d_{la}^{location}\, X_{la}^{center}
\end{aligned}
\qquad (31)
$$

4 EXPERIMENTATION and ANALYSIS

We present a solution method for the problem described in Section 3. In order to solve the problem, we approximate the problem by a linear programming formulation. Then, we convert the problem to a minimization cost problem.

The objective of this section is to validate the model and determine the numerical results. The integer programming model was tested with a real scenario and real data.

The mathematical model is applied for a case study: we select eight regions in the province of Quebec: Montréal, Québec, Laval, Gatineau, Trois-Rivières, Longueuil, Saint-Jérôme and Terrebonne. This selection is based on the number of population centres to be served. In addition, in our case study, we select five possible sizes that can be affected to opened data centres. We present in detail the possible sizes of data centres in (Table 1).

Table1. Data centre possible sizes

	d1	d2	d3	d4	d5
dimension(m²)	2600	2200	1700	1200	800

In addition, throughout our project, we make two assumptions. First, we choose to set up data centres that offer three different types of Internet services: Web service (Web server), processing information service (FTP Server) and communication service (Server of videoconferencing). Indeed, service availability varies from one service to another. Thus, in the second hypothesis, we assign a 10-port switch to interconnect servers with eight incoming ports for servers and 2 ports for Ethernet connection. The described model in Section 3 is solved by using *Lingo* optimizer version 14.0.

The optimal total cost is 141 507 130M$. This cost is the result of the setup of five data centres with different sizes: the first one is located in Montréal, the second one in Québec, the third in Laval, the fourth is located in Gatineau and the last one is located in Trois-Rivières. Each data centre serves area zones. The solution is detailed in (Figure 1).

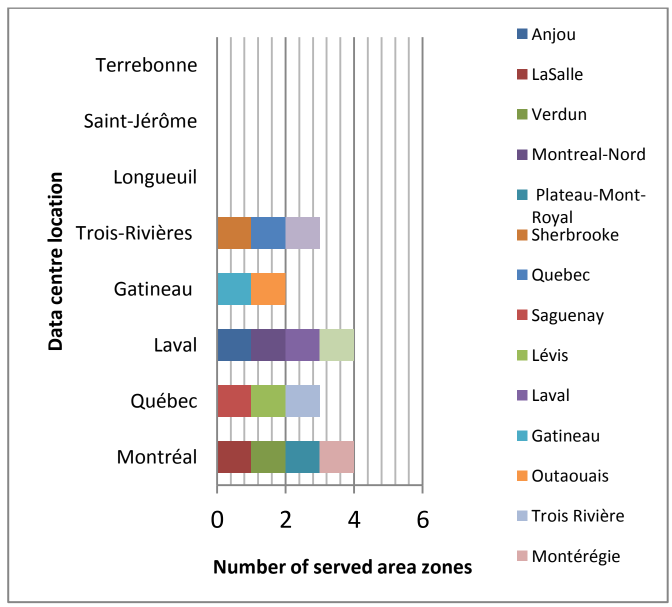

Figure 1. Data centre locations and assigned area zones

The (Figure 1) indicates that five opened data centres serve 16 population centres (e.g., Montréal serves LaSalle, Lévis, Plateau Mont-Royal and Montérégie). In addition, the minimized costs are formed mainly by five costs: IT equipment costs, BW costs, energy costs, variable costs and fixed costs. The percentages of the minimized cost is distributed as follows and presented in (Figure 2).

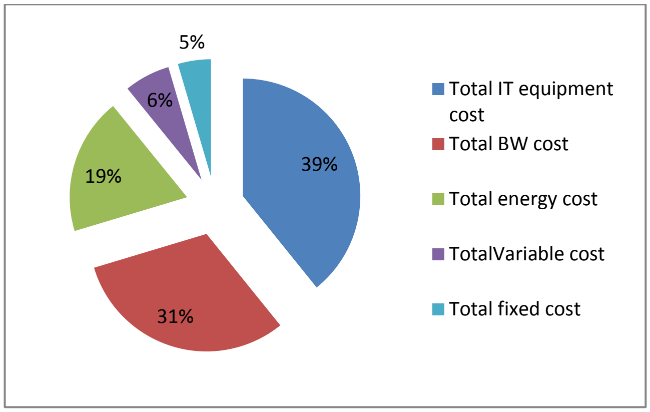

Figure 2. Data centre costs for the optimal solution of F

As indicated in the above figure, the total energy cost is 19% of the total costs. In fact, when a company decides to locate a data centre, the first decision variable is the cost of energy. In our case, the strategic position of the province of Quebec offers competitive energy prices. Therefore, there is a lower installation and implementation cost. The energy cost depends on the total amount of IT equipment and especially the number of servers affected for each area zone. In this case, we present in (Figure 3) the total number of servers assigned for each area zone.

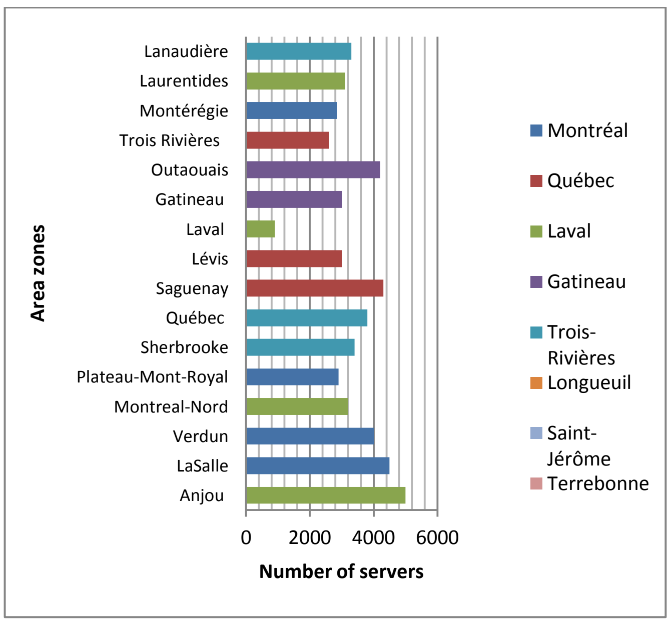

Figure 3. Total number of servers affected for each area zone

To obtain minimize costs of IT equipment, our MILP model ensures a good distribution of servers in each location in order to respond to customer demands. Thus, this distribution is performed by using the constraints (22). Indeed, this constraint is a way to impose a minimum use of resources over the size, so it artificially increases the amount of IT equipment.

In addition, throughout this project, we assume that the redundancy level of our data centre is Level I. To resolve the problem of data centre breakdown, we define a latency constraint that is detailed in constraint (19). If such problem occurs, a mirror data centre takes over the service in order to maintain high availability and continuity of services. Also, the backup capacity is shared between data centres. In addition, constraints (20) and (21) define the total energy consumed by IT equipment, thus, we calculate the power usage effectiveness (PUE) of located data centres. The PUE value requires a minimum number of servers in use. So, we don't take into account the total number of backup servers. In fact, these constraints force the opening of the data centres only if the data centres have sufficient equipment, while some centres take over hardware required for the operation and backup. This excess is contained in variable S_{la}^{backup}. (Figure 4) presents in detail the total number of servers in backup data centres and the distribution for population centres.

Note that that the backup servers are inactive until the appearance of a problem in some servers of another data centres.

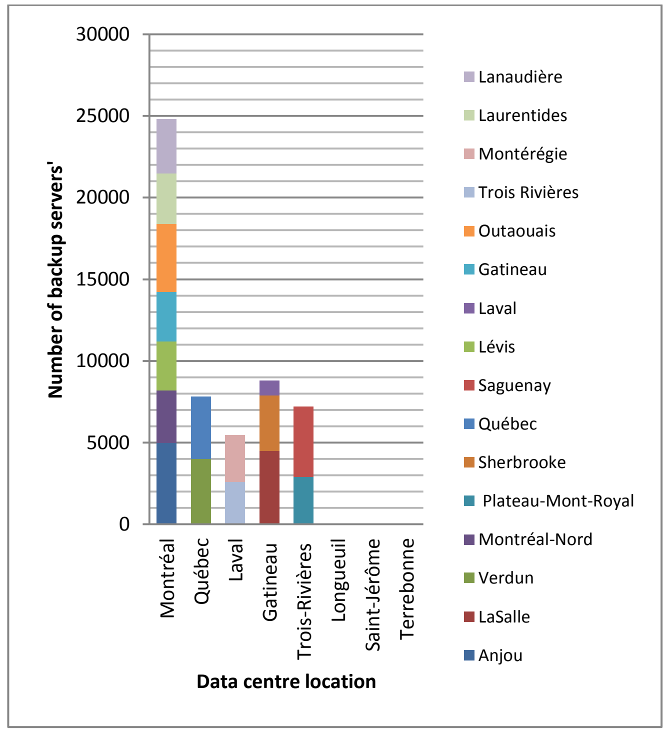

Figure 4. Total number of servers affected for each area zone in backup data centres

As is calculated in constraint (22), the PUE values are given in Fig.5. This value depends on location temperature and the total number of functional servers in each data centre. As the located data centres have different sizes (size d2= 2200 m² and size d5= 800 m²), we have, as is indicated on (Figure 3) and (Figure 4), different numbers of servers in data centres and those used as backup. In addition, we calculate the total energy consumed by the IT equipment W_{dl}^{TI}. Those values are detailed in (Figure 5).

W_{dl}^{TI} (MW)	Montreal	Quebec	Laval	Gatineau	Trois-Rivière	Longueuil	Saint-Jérôme	Terrebonne
d1	0	0	0	0	0	0	0	0
d2	6188	0	0	0	0	0	0	0
d3	0	0	0	0	0	0	0	0
d4	0	0	0	0	0	0	0	0
d5	0	4181	5101	3080	4421	0	0	0

Figure 5. Energy consumed by IT equipment in MW

Therefore, the power usage effectiveness (PUE) values are respectively 1.5, 1.34, 1.33, 1.28 and 1.28.

These values respect the maximum power usage effectiveness (PUE) value which is fixed according to (Uptime Institute 2012) to 1.65.

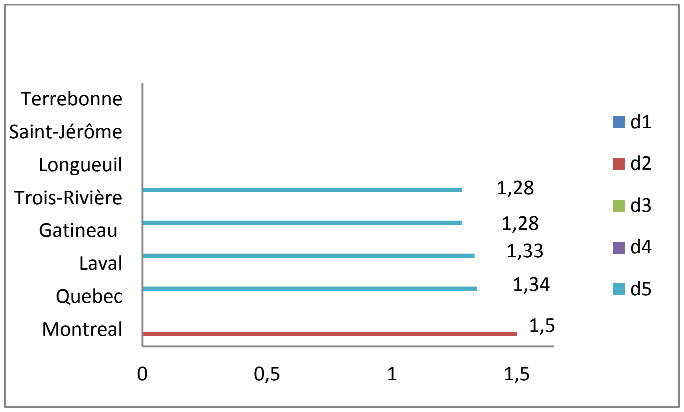

Figure 5. Power usage effectiveness (PUE) values

These PUE values indicate that the numbers of hosted servers are distributed in function of data centre sizes and the population centres to serve. Based on these values, we can state that the opened data centres consume the right amount of energy. Thus, the constraint (22) indicated that we cannot open a data centre if we do not need it. In addition, this constraint ensures selecting just one data centre size. In addition, when our model selects data centre sizes, it should take into account the total number of servers so as to not use more energy. In this case, the model is highly combinatorial.

Finally, we observe that the solution runtime takes 90 seconds. *Lingo* supports Special Ordered Sets (SOS). In this case, we use the SOS1 and SOS3 functions respectively in constraints (10) and (11). It defines a set of binary variables and places restrictions on their collective values. For SOS1, at most, only one variable belonging to an SOS1 set will be greater than 0. For SOS3, exactly one variable from a given SOS3 set will be equal to 1.

As result, the elapsed runtime of our solution is just 11 seconds. All details are presented in (Table 2).

Table 2. Running time of solution approach

	Generator memory (k)	Runtime (s)
Pure MILP	3537	90
MILP with SOS sets	427	11

In addition, the execution of the function F1 (31) gives us the amortization cost. The value of this cost is 55 241 307 $.

The amortized cost is detailed in (Figure 6). This Figure indicates that variable and fixed costs are negligible compared to bandwidth, energy and IT equipment costs.

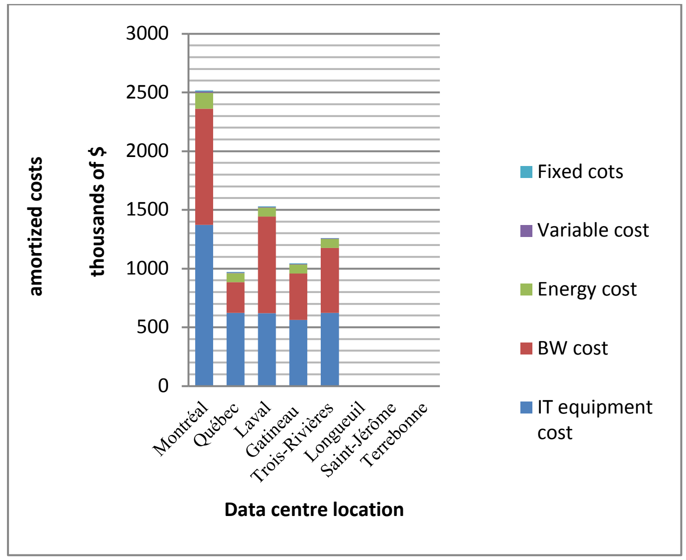

Figure 6. Total amortized costs

5 CONCLUSIONS AND PERSPECTIVES

Throughout this paper, we presented a location model in an MILP framework. This model aims to minimize the data centre costs including variable costs, fixed costs, energy costs, bandwidth cost and IT equipment cost. This objective is obtained by respecting three constraints: service response time, data centre latency time and the power usage effectiveness metric. By calculating the energy consumption by IT equipment, we determined the PUE values for each location in order to verify the energy efficiency of located data centres. In addition, Data centre costs assume a 12-year lifetime for data centres and a 4-year lifetime for IT equipment. In this case, we apply the declining balance method by calculating the total amortized cost.

As future work, we will apply our mathematical model to a real life size example. Indeed, the quality of results returned by our approaches depends primarily on the quality and accuracy of the data in question. In addition, data centre location is a well-known theoretical problem. However, treatment of this problem has focused on algorithmic aspects and theoretical implications, in the absence of real location data.

Also, we will consider minimizing and increasing energy efficiency by integrating “green” servers.

6 REFERENCES

Abbasi, Z., Varsamopoulos, G.,Gupta, S. K. S., (2012) TACOMA: Server and workload management in internet data centers considering cooling-computing power trade-off and energy proportionality. *ACM Trans. Archit. Code Optim.,* (9)**,** pp.1-37.

Covas, M. T., Silva, C. A.,Dias, L. C., (2013) Multicriteria decision analysis for sustainable data centers location. *International Transactions in Operational Research,* (20)**,** pp.269-299.

Daim, T. U., Bhatla, A.,Mansour, M., (2012) Site selection for a data centre - a multi-criteria decision-making model. *International Journal of Sustainable Engineering,* (6)**,** pp.10-22.

Goiri, I., Kien, L., Guitart, J., Torres, J.,Bianchini, R., (2011) Intelligent Placement of Datacenters for Internet Services. Distributed Computing Systems (ICDCS), 31st International Conference on, 20-24 June 2011, pp.131-142.

Greenberg, A., Hamilton, J. R., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D. A., Patel, P.,Sengupta, S., (2009) VL2: a scalable and flexible data center network. *Proceedings of the ACM SIGCOMM 2009 conference on Data communication.* Barcelona, Spain: ACM.

Hong, X.,Baochun, L., (2012) Cost efficient datacenter selection for cloud services. Communications in China (ICCC), 1st IEEE International Conference on, 15-17 Aug. 2012, pp.51-56.

Jie, L., Zuyi, L., Kui, R.,Xue, L., (2012) Towards Optimal Electric Demand Management for Internet Data Centers. *Smart Grid, IEEE Transactions on,* (3)**,** pp.183-192.

Lei, R., Xue, L., Le, X.,Wenyu, L., (2010) Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi-Electricity-Market Environment. INFOCOM, Proceedings IEEE, 14-19 March 2010. pp.1-9.

Lei, R., Xue, L., Le, X.,Zhan, P., (2011) Hedging Against Uncertainty: A Tale of Internet Data Center Operations Under Smart Grid Environment. *Smart Grid, IEEE Transactions on,* (2)**,** pp.555-563.

Oley, B. 2010. *Where's the Best Place to Build a Data Center?* [Online]. Data Center Journal.

Pakbaznia, E.,Pedram, M., (2009) Minimizing data center cooling and server power costs. *Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design.* San Fancisco, CA, USA: ACM.

Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z.,Zhu, X., (2008) No "power" struggles: coordinated multi-level power management for the data center. *SIGPLAN Not.,* (43)**,** pp.48-59.

Rath, J., (2007) Data center site selection, pp.37.

Stansberr, M., (2006) Data Center Locations Ranked by Operating Cost.

Uptime Institute , L., (2012) Data center site infrastructure Tier Standard Available: http:// uptimeinstitute.com.

Van Heddeghem, W., Lambert, S., Lannoo, B., Colle, D., Pickavet, M.,Demeester, P., (2014) Trends in worldwide ICT electricity consumption from 2007 to 2012. *Computer Communications,* (50)**,** pp.64-76.

Yao, J., Xue, L., Wenbo, H.,Rahman, A., (2012) Dynamic Control of Electricity Cost with Power Demand Smoothing and Peak Shaving for Distributed Internet Data Centers. Distributed Computing Systems (ICDCS), IEEE 32nd International Conference on, 18-21 June 2012, pp.416-424.

Yuanxiong, G., Zongrui, D., Yuguang, F., (2012) Dapeng, W. Cutting Down Electricity Cost in Internet Data Centers by Using Energy Storage. Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, 5-9 Dec.2011, pp.1-5.