Green Latency-aware Data Deployment in Data Centers: Balancing Latency, Energy in Networks and Servers

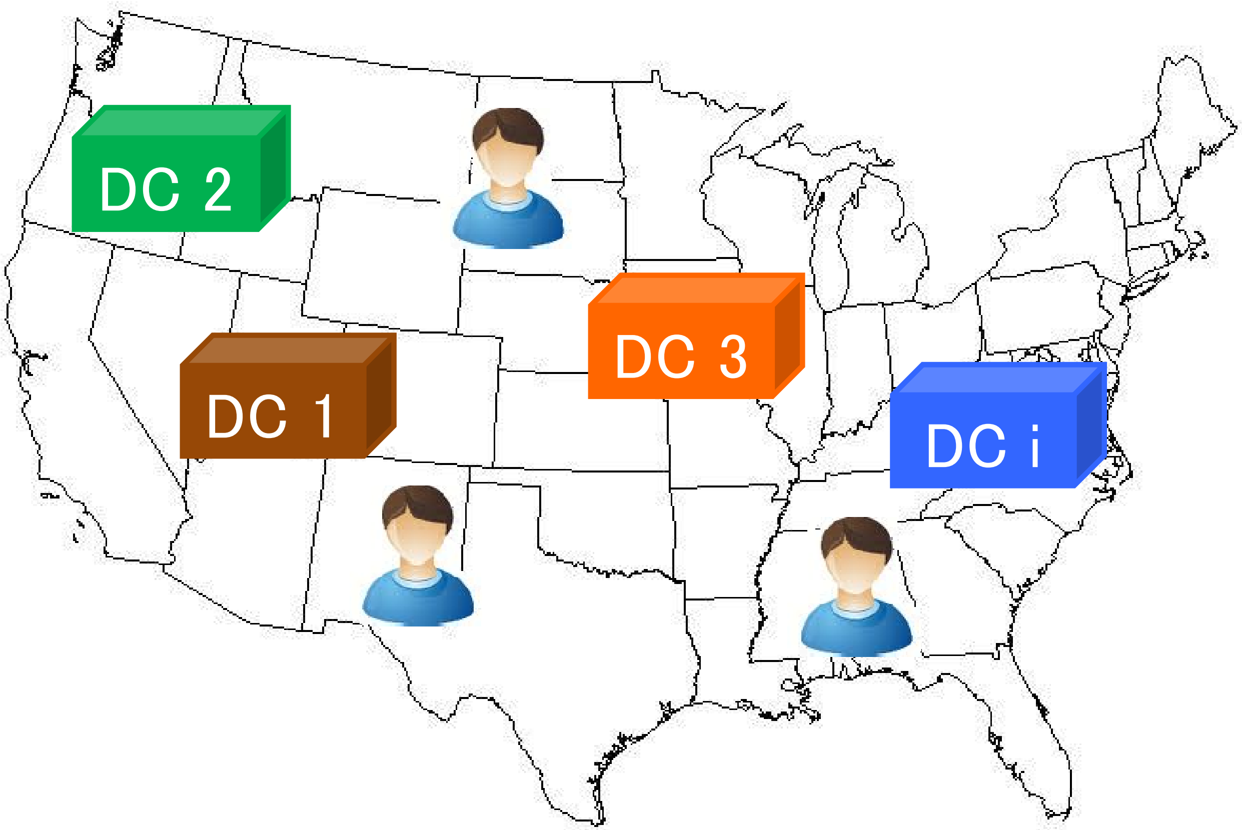
Yuqi Fan, Hongli Ding, Donghui Hu

School of Computer and Information

Hefei University of Technology

August 18, 2014

Model



Motivation

- Two concerns exist in service provisioning by data centers
  - Users require to experience low latency while accessing data from the data centers
  - Reduce the power consumed by network transport and servers in the data centers

Problem

- We tackle the problem of green data deployment in the data centers, taking into account the three factors of latency, energy consumption of the data centers and the network transport
- The cost of deploying data on a server in a data center integrates the three factors above
  - each factor has a coefficient in the cost function

Objective Function

Minimize:

$$\lambda_1 \sum_{u_i, dc_j, s_m, d_k} rep(dc_j, s_m, d_k) p(u_i \mid d_k) l(u_i, dc_j)$$
$$+\lambda_2 \sum_{dc_j, s_m} rep(dc_j, s_m) e_S(dc_j, s_m)$$
$$+\lambda_3 \sum_{u_i, dc_j, s_m, d_k} s(d_k) rep(dc_j, s_m, d_k) p(u_i \mid d_k) e_I(u_i, dc_j)$$

Subject to:

$$rep(dc_j, s_m) = \sum_{d_k} rep(dc_j, s_m, d_k)$$

$$\sum_{dc_j, s_m} rep(dc_j, s_m, d_k) = 1$$

$$e_S(dc_j, s_m) = P_{s_m}^{dc_j} * PUE(dc_j)$$

$$\sum_{d_k} rep(dc_j, s_m, d_k) s(d_k) \leq C(s_m, dc_j), \forall s_m, dc_j$$

- λ_1 , λ_2 , and λ_3 are the weights of the sub-objectives of the latency, the energy consumption of the data centers and the network transport, respectively.
- rep(dc_j , s_m, d_k) indicates whether data d_k is deployed in server s_m in data center dc_j .
- p(u_i | d_k) is the probability that a given request is asking for data d_k and it comes from user group u_i.
- l(u_i, dc_j) is the latency between user group u_i and data center dc_j.
- rep(dc_j, s_m) is the indicator whether server s_m in data center dc_j has been deployed some data

- $e_S(dc_j , s_m)$ is the energy consumption of server s_m in data center dc_j.
- $s(d_k)$ is the size of data d_k.
- $e_I (u_i, dc_j)$ is the energy required to transport one bit from data center dc_j to user group u_i through the Internet.
- $PUE(dc_j)$ is the PUE of data center dc_j is the power of sever s_m in data center dc_j.
- $C(s_m, dc_j)$ is the capacity of server s_m in data center Dc_j.

GLDD (Green Latency-aware Data Deployment)

- When processing each data chunk d_k, GLDD searches the servers in all the data centers with the least cost to deploy data d_k.
- Each server in each data center is checked to obtain the cost to accommodate data d_k on the server if the server has enough capacity.
- The cost of deploying data d_k on server s_m in data center dc_j integrates the three factors of the latency, the power consumed by the servers and the network transport.

Algorithm 1 *GLDD Algorithm*

Input: Data Request Probability Matrix $P(u_i \mid d_k)$
Input: Network Latency Cost Matrix $L(u_i, dc_j)$
Input: Network Transport Energy Cost Matrix $E_I(u_i, dc_j)$
Input: Servers Power Cost Matrix $E_S(u_i, dc_j)$
Input: Data Size Queue $S(d_k)$
Output: $Rep(dc_j, s_m, d_k)$

```
Sort S(d_k) by non-ascending order of data size.
while Queue of S(d_k) not empty do
    get the head d_k from the Queue S(d_k)
    for each data center dc_j do
        for each server s_m in data center dc_j do
            if server s_m has enough capacity to accommodate
            data d_k then
                Calculate the cost to deploy data d_k on server
                s_m in data center dc_j;
            end if
        end for
    end for
    Obtain the server s_m in the data center dc_j that costs
    the least and has enough capacity to accommodate the
    data d_k.
    C(s_m, dc_j)=C(s_m, dc_j) - S(d_k)
    rep(dc_j, s_m, d_k)=true
end while
Return Rep(dc_j, s_m, d_k)
```

Result

We evaluate the performance of the algorithm GLDD by comparing GLDD with the algorithm FORTE proposed in SIGCOMM'12.