The Energy Optimization in Data Center Network

Reporter: 羅婧文
Advisor: Hsueh-Wen Tseng
Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Conclusion
Reference

- **Energy optimizations for data center network: Formulation and its solution**
  - Shuo Fang; Hui Li; Chuan Heng Foh; Yonggang Wen; Khin Mi Mi Aung
  - Global Communications Conference (GLOBECOM), 2012 IEEE

- **Limits of energy saving for the allocation of data center resources to networked applications**
  - Leon, X.; Navarro, L.
  - INFOCOM, 2011 Proceedings IEEE

- **HERO: Hierarchical energy optimization for data center networks**
  - Yan Zhang; Ansari, N.
  - Communications (ICC), 2012 IEEE International Conference on
Introduction

- Data centers should provide high availability and fault tolerant
  - Require high energy consumption

CO2 emissions

- 98% (ICT)
- 3% (Other)

Global energy expenditure

- 97% (ICT)
- 3% (Other)
Introduction

- Power consumption in a data center

- Nearly 30% of the total computing energy in a data center is consumed by the communication links, switching, and aggregation elements
Introduction

Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum utilization levels.
Energy optimizations for data center network: Formulation and its solution

Shuo Fang; Hui Li; Chuan Heng Foh; Yonggang Wen; Khin Mi Mi Aung; Global Communications Conference (GLOBECOM), 2012 IEEE
Paper 1

- Studies on data center traffic characteristics
  - Network is seldom utilized at its peak capacity
  - Idle state
    - Increase power consumption
  - Redundancies in network architecture
    - Increase power consumption

- Power saving in data centers
  - Minimize switch usage and adjust link rates of switch ports according to traffic loads
Paper 1

- Macro level
  - Switch
  - Reduce redundant energy usage incurred by network redundancies for load balancing

- Micro level
  - Port
  - Design algorithm to limit port rate in order to reduce unnecessary power consumption
Paper 1

Link rate ↑, Number of links ↑

Power consumption ↑

Table I: Switch power consumption of Dell PowerConnect 8024F.

<table>
<thead>
<tr>
<th>Number of links</th>
<th>100Mbps</th>
<th>1Gbps</th>
<th>10Gbps</th>
<th>Power consumption (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>115.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
<td>118.7</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td>119.7</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
<td>120.9</td>
</tr>
</tbody>
</table>

Fig. 3. Power consumption of Dell PowerConnect 8024F.
\[ \mathcal{R} = \{ \langle 0, r \rangle \mid r \in \{1, 2, \ldots, (m/2)^2\} \} \]
\[ \cup \{ \langle p, r \rangle \mid p \in \{1, 2, \ldots, m\}, r \in \{1, 2, \ldots, m\} \} \]

Fig. 1. Illustration of 4-ary Fat Tree topology.
Greedy approach

- The key idea in this solution is to utilize as few switches, switch links and switch link rates as possible.
- In the initial stage, network system begins with no active switches, switches are only enabled when packet arrives.
- Packets are automatically routed to a path on a spanning tree with the least link rate given the traffic load.
$T_{up} \rightarrow$ Upgrade threshold
$T_{down} \rightarrow$ Downgrade threshold
$\tau \rightarrow$ Time interval

Activate an adjacent port to support heavier traffic load

Port’s states: SLEEP, 100Mbps, 1Gbps and 10Gbps

Fig. 4. Port state transition.
Fig. 5. Switch state transition.
### Paper 1

<table>
<thead>
<tr>
<th>Number of flows</th>
<th>Power saving</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>62.9%</td>
</tr>
<tr>
<td>20</td>
<td>40.2%</td>
</tr>
<tr>
<td>50</td>
<td>45.3%</td>
</tr>
<tr>
<td>100</td>
<td>30.5%</td>
</tr>
<tr>
<td>200</td>
<td>22.9%</td>
</tr>
<tr>
<td>300</td>
<td>21.7%</td>
</tr>
</tbody>
</table>

#### TABLE V
**Energy usage comparison.**

<table>
<thead>
<tr>
<th>Number of flows</th>
<th>Energy usage (J)</th>
<th>Our solution</th>
<th>FT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>14009</td>
<td>37800</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>21844</td>
<td>36540</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>21786</td>
<td>36540</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>26250</td>
<td>37800</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>29119</td>
<td>37800</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>29585</td>
<td>37800</td>
<td></td>
</tr>
</tbody>
</table>
The mean delay among all hosts tends to be less as the number of flows increases.

Fig. 9. Hosts delay statistics.
Limits of Energy Saving for the Allocation of Data Center Resources to Networked Applications

Leon, X. ; Navarro, L.

INFOCOM, 2011 Proceedings IEEE
Paper 2

- Power saving in data centers
  - Stackelberg leadership game
    - Leader
      - The infrastructure operator
      - Determining which resources to keep on and off
    - Follower
      - A set of strategic users buying resources as followers
      - Decide their optimal bidding vector of the resources
### Energy consumption model

The maximum amount of energy consumed by a device at full capacity

\[ f_e(s) = \sigma_e + \mu_e s^\alpha \]

- **Dynamic energy consumption**
- **CPU Load**

The fixed cost of maintaining a server power on and available
Paper 2

Resource allocation model
- The simplest and most appealing market-based mechanism for shared divisible resources is the **proportional share allocation mechanism**

\[ Y_j = \sum_{i=1}^{k} x_{ij} \]

- The price of the resource \( j \)
- The number of bids on resource \( j \)
- The bid of user \( i \) for resource \( j \)

\[ r_{ij} = \frac{x_{ij}}{Y_j} \]

\[ r_{ij} = q_j \frac{x_{ij}}{Y_j} \]

Resource state (on/off)
Paper 2

- **Stackelberg competition model**
  - **User model → Follower**
  - **Providers model → Leader**

**Linear payoff function**

\[ U_i(r_{i1}, \ldots, r_{im}) = q_{i1}w_{i1}r_{i1} + \ldots + q_{im}w_{im}r_{im} \]

- Resource state (on/off)
- The resource share obtained from resource k
- User i’s private preference for resource

\[
\max U_i(x_1, \ldots, x_m) = \sum_{j=1}^{m} q_{ij} \frac{x_{ij}}{x_{ij} + y_j}
\]
Fig. 1. User i’s best response algorithm

**Require:** $\phi \{\text{user i’s minimum share}\}$
**Require:** $X \{\text{user i’s budget}\}$
**Require:** $\{y_1, \ldots, y_m\} \{\text{list of resource prices}\}$
**Require:** $\{q_1, \ldots, q_m\} \{\text{list of resource states (on/off)}\}$

$M = \{y_j : q_j = 1\} \{\text{list of prices of on machines}\}$

Sort the set $M$ by $y_j$ in increasing order

Compute largest $k$ such that

$$\sum_{i=1}^{k} \frac{\sqrt{y_i}}{\sqrt{y_i}} (X + \sum_{i=1}^{k} y_i) - y_k \geq \frac{y_j \phi}{1 - \phi}$$

Set $x_j = 0$ for $j > k$, and for $1 \leq j \leq k$, set:

$$x_j = \frac{\sqrt{y_j}}{\sum_{i=1}^{k} \sqrt{y_i}} (X + \sum_{i=1}^{k} y_i) - y_j$$

**return** $(x_1, \ldots, x_m)$

$$x_j \geq \frac{y_j \phi_i}{1 - \phi_i}$$
 Providers model → Leader

\[
\begin{align*}
\text{Profit maximize} \\
\max_{q_1, \ldots, q_m} P(q_1, \ldots, q_m) &= \sum_{j=1}^{m} (q_j \sum_{i=1}^{n} x_{ij}) - \sum_{j=1}^{m} q_j c_j \\
\text{Resource state (on/off)} & \quad \text{A non-negative bid of user } i \\
\text{for resource } j & \quad \text{The actual cost of maintaining the infrastructure}
\end{align*}
\]
Paper 2

Fig. 2. Provider’s best response algorithm

Require: $\{\phi_1, \ldots, \phi_n\}$ {users minimum share}
Require: $\{\tau_1, \ldots, \tau_n\}$ {users minimum number of nodes}
Require: $\{X_i, \ldots, X_n\}$ {users budget}
Require: $M = \{c_1, \ldots, c_m\}$ {list of resource costs}
Sort the set $M$ by $c_j$ in increasing order
k=0
repeat
  $k \leftarrow k + 1$
  Set $q_j = 1$ for $j \leq k$
  Set $q_j = 0$ for $j > k$
  repeat
    for all user $i$ do
      UpdateBestResponse($\phi_i$, $\tau_i$, $X_i$)
    end for
  until convergence
until $\forall_i |S_i| \geq \tau_i$ OR $k = m$
return $(q_1, \ldots, q_m)$
Paper 2

The number \( m \) of nodes = 1000
The number \( n \) of users = 250

\[
P(q_1, \ldots, q_m) = \sum_{j=1}^{m} \left( q_j \sum_{i=1}^{n} x_{ij} \right) - \sum_{j=1}^{m} q_j c_j
\]

(a) Profit

(b) Energy Savings
The ratio between the number $k$ of satisfied users and the total number $n$ of users.
HERO: Hierarchical energy optimization for datacenter networks

Yan Zhang; Ansari, N.; Communications (ICC), 2012 IEEE International Conference on
Scalability problem
- As the data center networks become larger and larger, the complexity of solving this optimization problem increases.

Power saving in data centers
- Establish a two-level power optimization model
  - Hierarchical energy optimization (HERO) model
    - Switching off network switches and links
    - Guarantee full connectivity and QoS
Traffic in DCNs can be categorized into five classes:
- Intra-edge switch traffic
- Inter-edge but intra-pod traffic
- Inter-pod traffic
- Incoming traffic
- Outgoing traffic

Fig. 1: Data center network topologies and traffic patterns.
The power optimization of datacenters can be divided into two levels

- **Core-level**
  - To determine the core switches that must stay active to flow the outgoing traffic
  - To determine the aggregation switches which serve the out-pod traffic in each pod

- **Pod-level**
  - To determine the aggregation switches that must be powered to flow the intra-pod traffic
The potential benefit of hierarchical energy optimization is to simplify energy optimization problem by reducing optimization variables greatly.
Hierarchical energy optimization algorithm

- Find the minimum power network subset to meet performance and fault tolerance goals by powering off the unneeded switches and links

Problem Formulation

- CMCF (Capacitated Multi-commodity Minimum Cost Flow) problems
  - Core-level
  - Pod-level
  - NP-hard
    - Heuristic algorithm
The power consumption of core-level and pod-level

Minimize

\[ P^{c}_{total} = \sum_{i=1}^{N^c} \sum_{j=1}^{N^c} x_{ij} P^{L}_{ij} + \sum_{i=1}^{N^c} y_i P^N_i \]

The node state

The power consumption of node \( i \)

The link state

The power consumption of the link between node \( i \) and node \( j \)

Minimize

\[ P^{pm}_{total} = \sum_{i=1}^{N^p_m} \sum_{j=1}^{N^p_m} x_{ij} P^{L}_{ij} + \sum_{i=1}^{N^p_m} y_i P^N_i \]
Determine the state of switches and links

Algorithm 1 Hierarchical Energy Optimization Algorithm

Stage 1: Determine in descending order of need to be powered on according to the traffic matrix $T$.

Stage 2: Solve the core-level CMCF optimization problem.
Stage 2.1: The power status of core switches and core-level links connecting the aggregation switches and the core switches is decided by solving the core-level CMCF optimization problem.
Stage 2.2: The aggregation switches serving the out-pod traffic in each pod are selected with the power status of the core-level links, and the selected aggregation switches are powered on.

Stage 3: Solve the pod-level CMCF optimization problem.
for $i = 1$ to $N^p$ do
    Determine the power status of the aggregation switches and the pod-level links connecting the edge switches and the aggregation switches by solving the pod-level optimization problem.
end for

Stage 4: In order to provision the whole network connectivity and to meet QoS goals, a merging process is performed.
Network connectivity

All the traffic flows in a traffic matrix can be classified into intra-edge traffic or inter-edge but intra-pod traffic

- Core switches stay in the idle state
- At least one core switch is powered on
  - Random select
- At least one aggregation switch that can connect to one active core switch must be turned on in each pod
The ratio of the total number of variables decreases with the increase of parameter K with the same number of flows.

**K-ary fat-tree topology**

Fig. 5: The power consumption results of different heuristic algorithms.
The power consumptions of HERO and the non-hierarchical model are almost the same under different traffic loads.

**Fig. 3:** The power consumption of 4-ary Fat-tree data center networks with different number of traffic flows.

**Fig. 4:** The power consumption of a 4-ary Fat-tree data center network with all-to-all traffic under different traffic load.
Conclusion

- Turn on/off switch and adjust link rates [1][3]
- Game theory[2]
- Power saving
  - Cooling
  - IT equipment
  - Location
## Conclusion

<table>
<thead>
<tr>
<th>Power Optimization</th>
<th>Resource Allocation</th>
<th>Load Balance</th>
<th>QoS</th>
<th>Fault Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>[2]</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>[3]</td>
<td>●</td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>