The Energy Optimization in Data Center Network

Reporter : 羅婧文

Advisor: Hsueh-Wen Tseng

Outline

- Introduction
- Paper1
- Paper2
- Paper3
- Conclusion

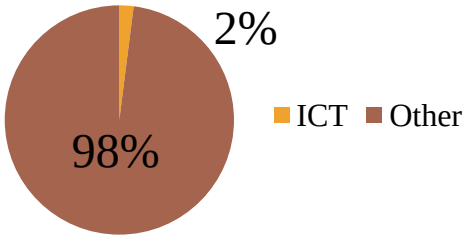
Reference

- **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

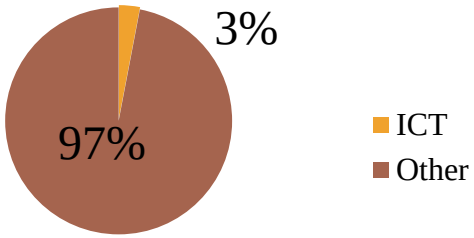
Introduction

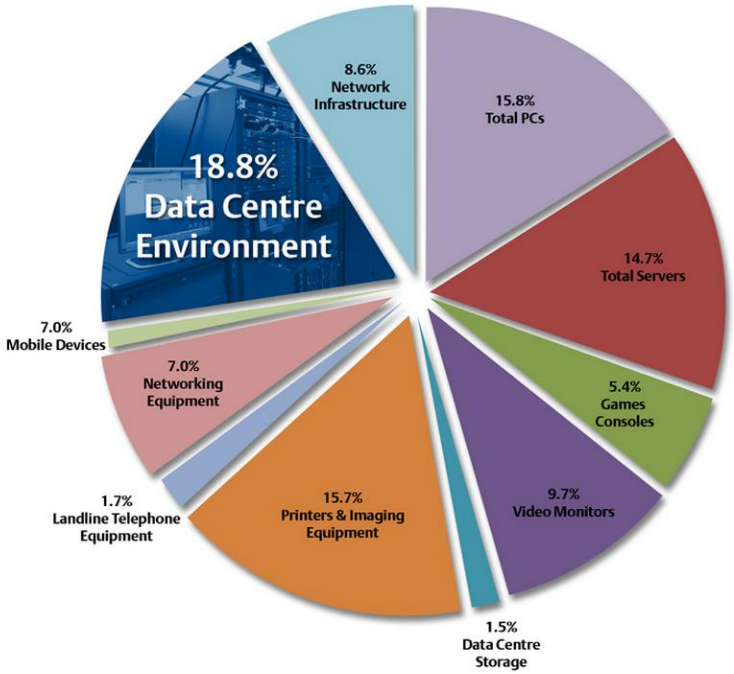
- Data centers should provide high availability and fault tolerant
  - Require high energy consumption

CO2 emissions



Global energy expenditure

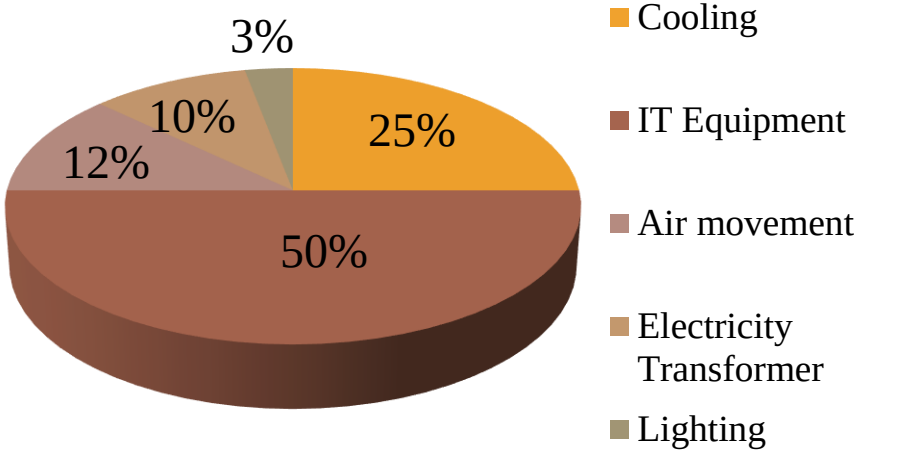


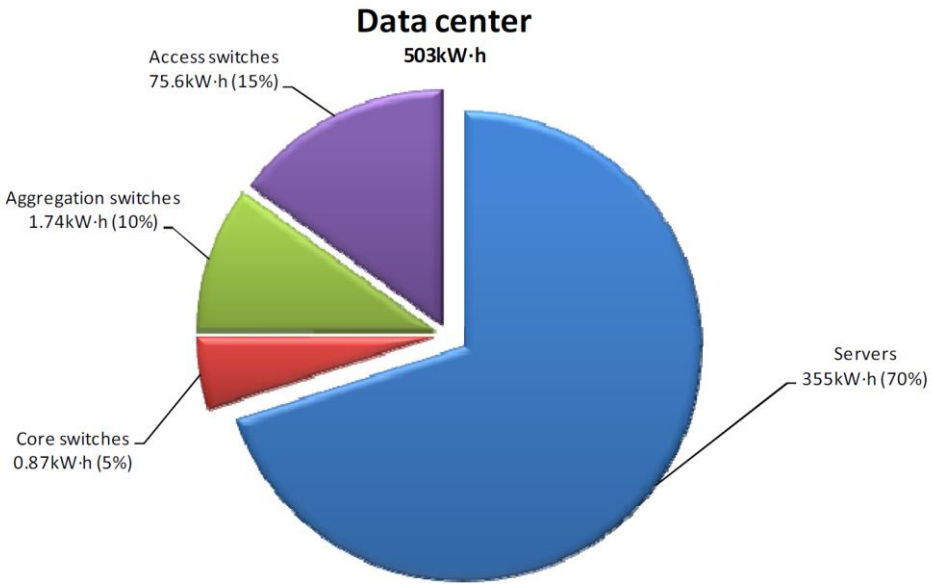


ICT Energy Consumption in Australia

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

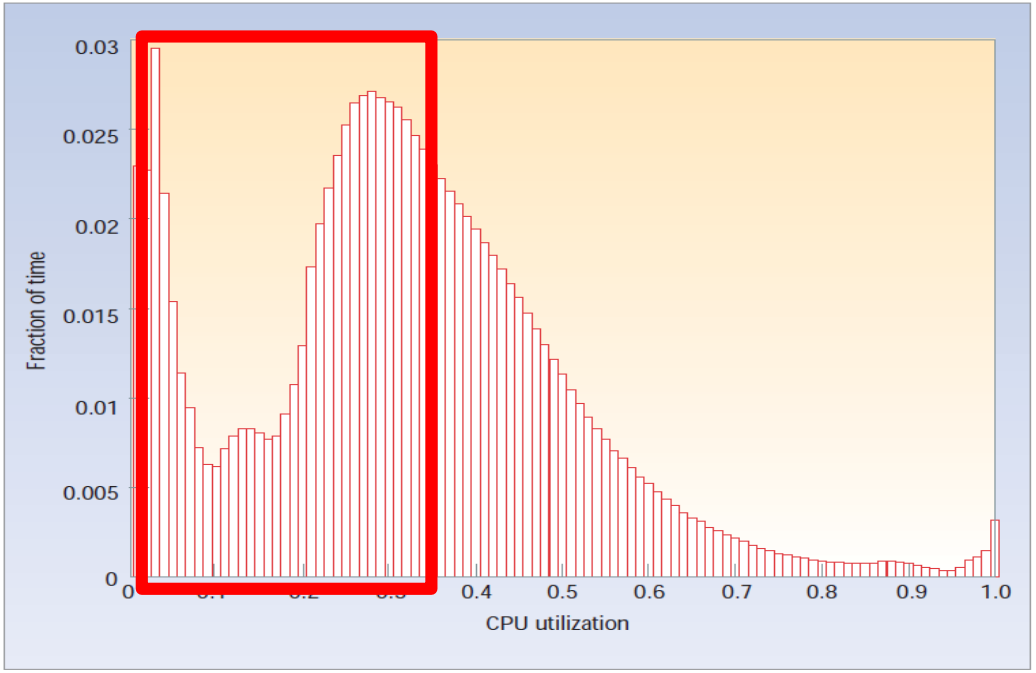
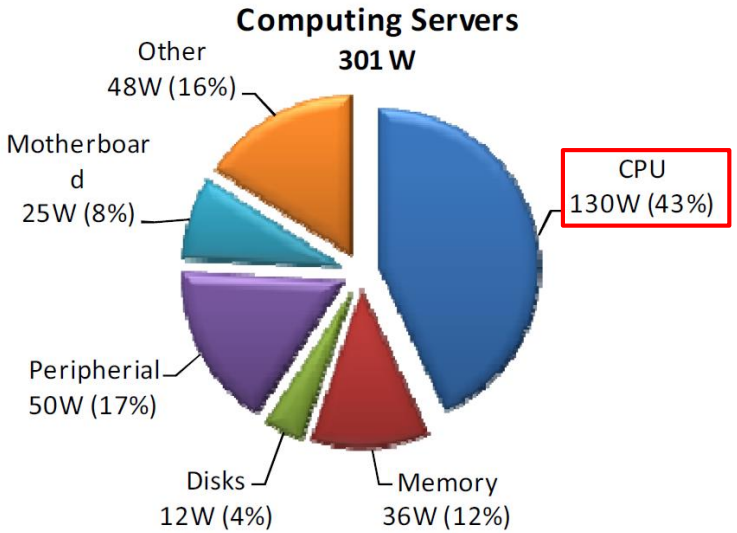


Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum utilization levels.



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

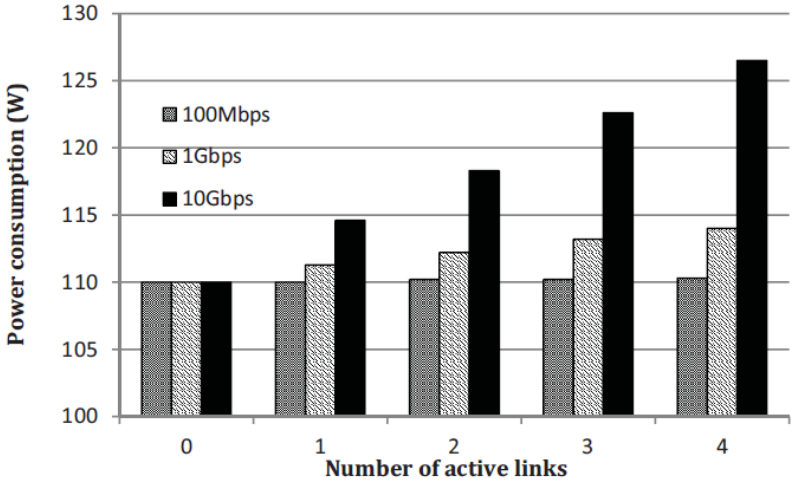


Fig. 3. Power consumption of Dell PowerConnect 8024F.

TABLE I
SWITCH POWER CONSUMPTION OF DELL POWERCONNECT 8024F.

Number of links			Power consumption
100Mbps	1Gbps	10Gbps	(W)
0	1	1	115.6
1	0	2	118.7
0	1	2	119.7
0	2	2	120.9

Link rate↑、Number of links ↑

Power consumption ↑

Paper 1

core switch

$$\mathcal{R} = \{\langle 0, r\rangle \,|r \in \{1, 2, ..., (m/2)^2\}\}$$
$$\bigcup \{\langle p, r\rangle \,|p \in \{1, 2, ..., m\}, r \in \{1, 2, ..., m\}\}$$

pods

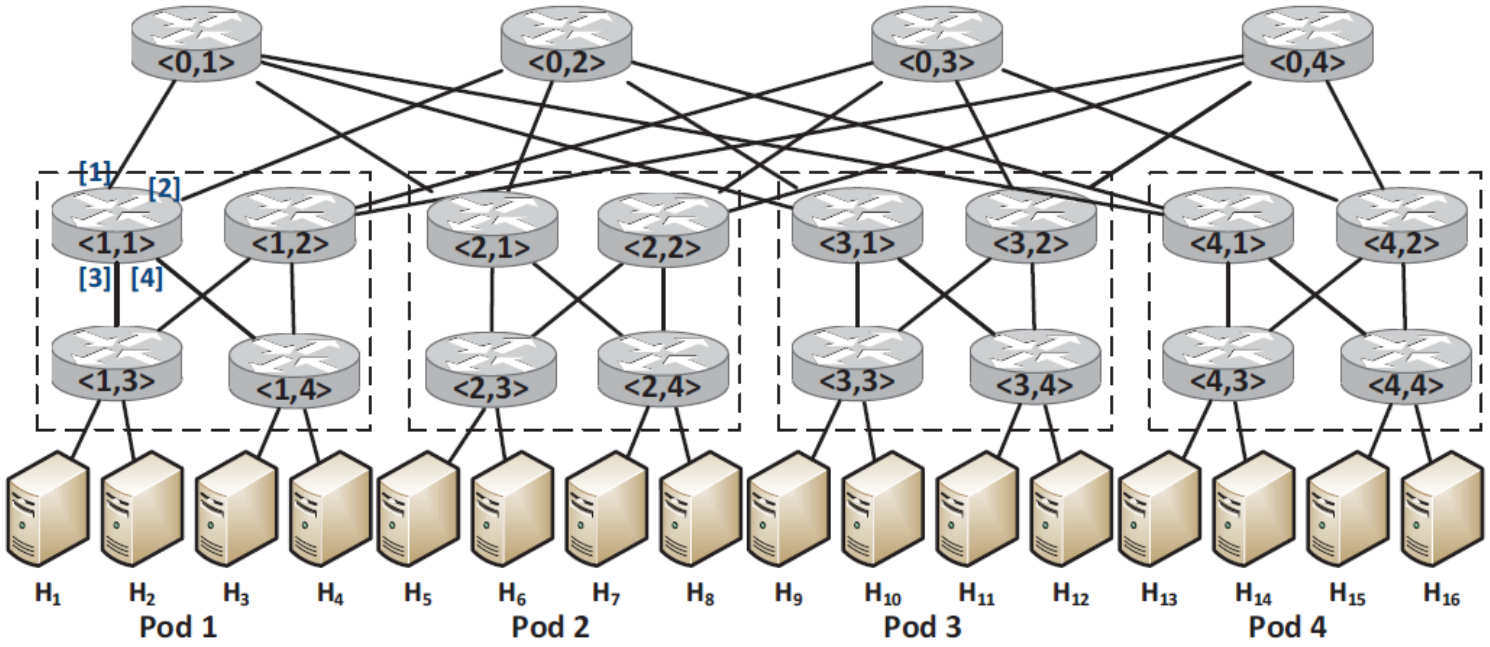


Fig. 1. Illustration of 4-ary Fat Tree topology.

$$\min \sum_{\langle p,r\rangle} P\Big(\sum_i l_i^{\langle p,r\rangle}\Big)$$

Paper 1

- 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

Paper 1

T_{up} → Upgrade threshold
T_{down} → Downgrade threshold
τ → Time interval

Activate an adjacent port to support heavier traffic load

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

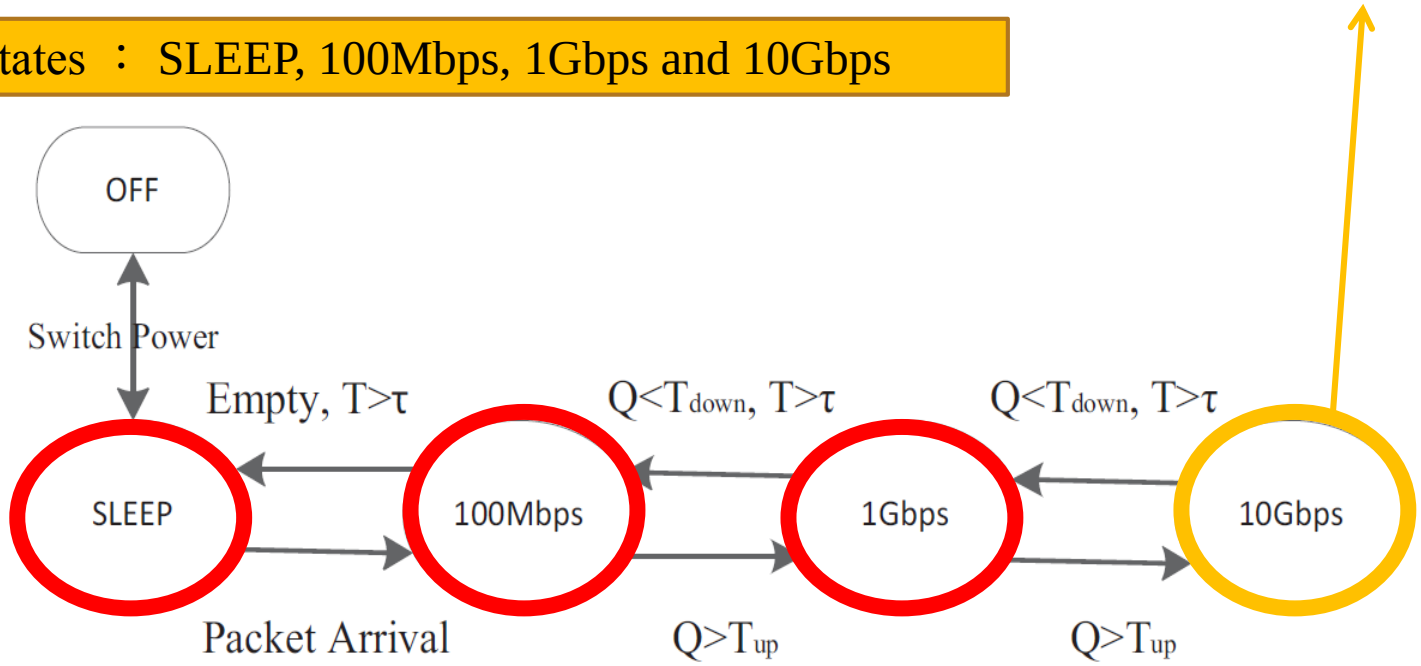


Fig. 4. Port state transition.

Paper 1

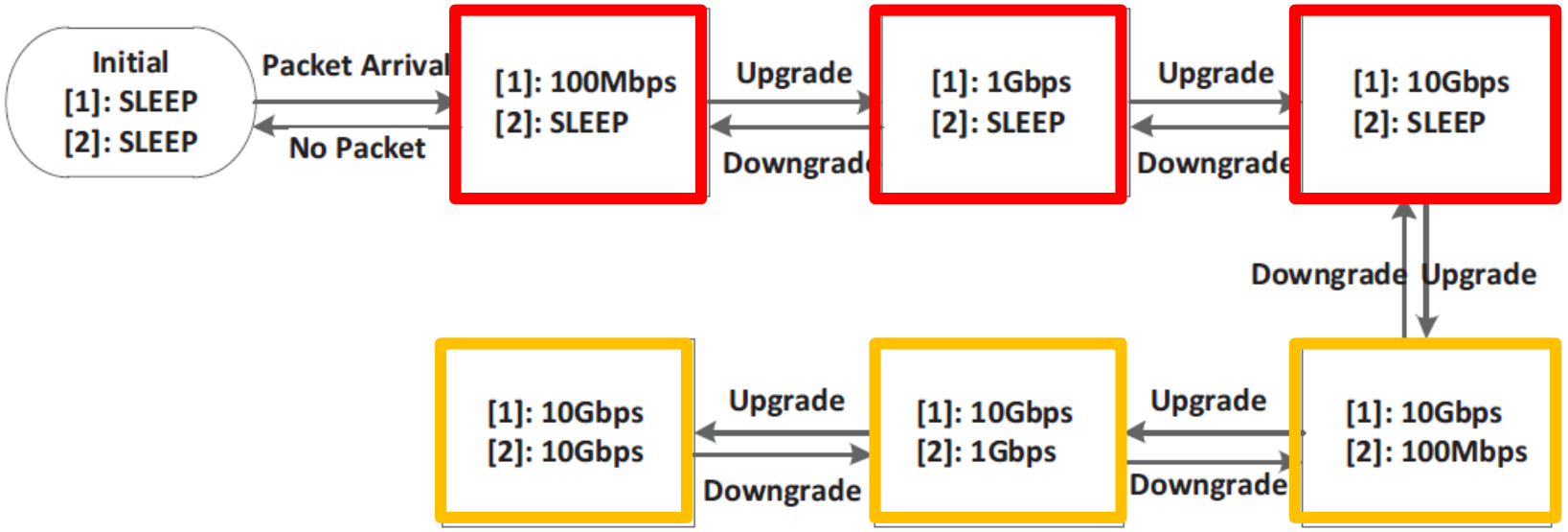


Fig. 5. Switch state transition.

Paper 1

Number of flows	Power saving
10	62.9%
20	40.2%
50	45.3%
100	30.5%
200	22.9%
300	21.7%

TABLE V
ENERGY USAGE COMPARISON.

Number of flows	Energy usage (J)	
	Our solution	FT
10	14009	37800
20	21844	36540
50	21786	36540
100	26250	37800
200	29119	37800
300	29585	37800

Paper 1

The mean delay among all hosts tends to be less as the number of flows increases

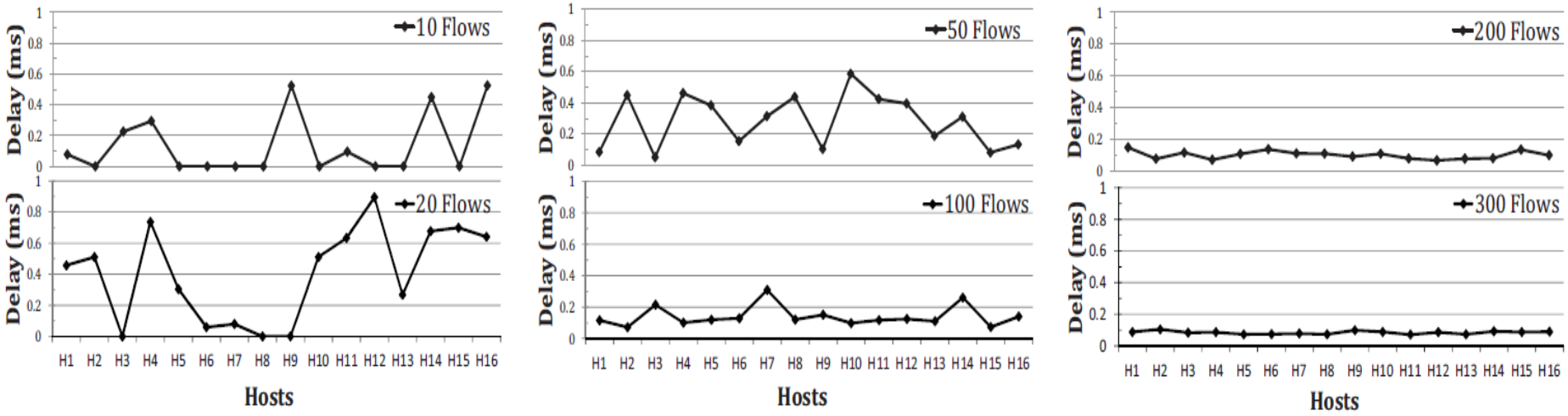


Fig. 9. Hosts delay statistics.

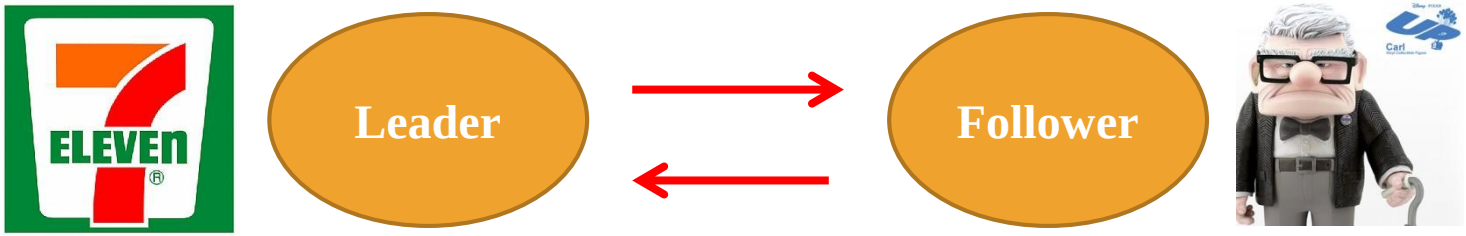
Limits of Energy Saving for the Allocation of Data Center Resources to Networked Applications

Leon, X. ; Navarro, L.

INFOCOM, 2011 Proceedings IEEE

Paper 2

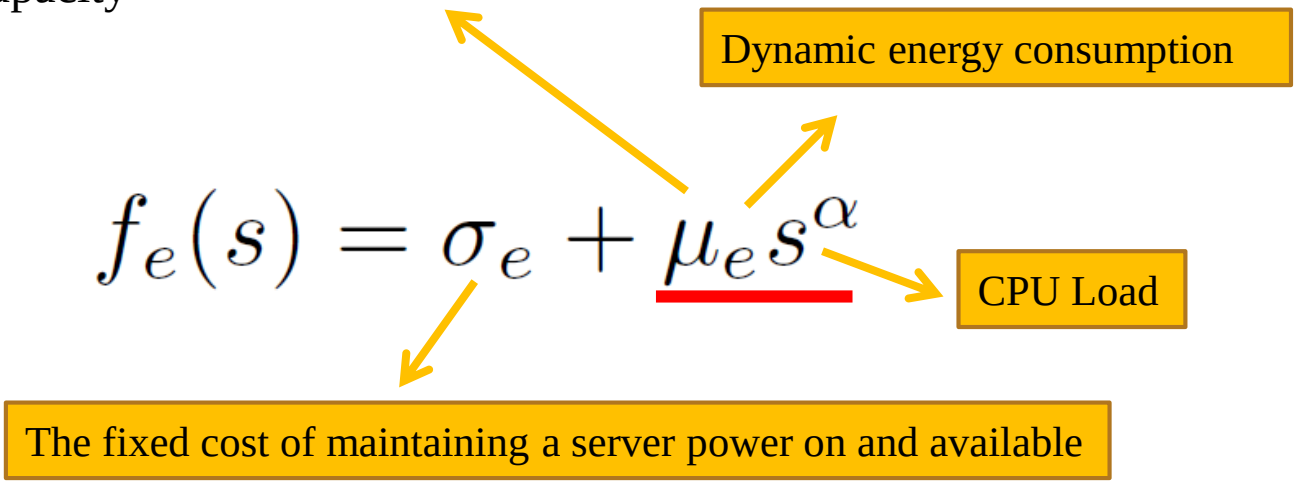
- Power saving in data centers
  - Stackelberg leadership game
    - Leader
      - The infrastructure operator
        - Determining which resources to keep on and off
    - Follower
      - A set of strategic users buying resources as followers
        - Decide their optimal bidding vector of the resources



Paper 2

- Energy consumption model

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



Paper 2

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

The number of bids on resource *j*

The price of the resource *j*

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

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

Resource state(on/off)

The resource share obtained from resource k

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

User *i's* private preference for resource

$$\max U_i(x_1, \ldots, x_m) = \sum_{j=1}^{m} q_{ij} \frac{x_{ij}}{x_{ij} + y_j}$$

Paper 2

Fig. 1. User i's best response algorithm

Require: ϕ {user i's minimum share}
Require: X {user i's budget}
Require: $\{y_1, \ldots, y_m\}$ {list of resource prices}
Require: $\{q_1, \ldots, q_m\}$ {list of resource states (on/off)}
$M = \{y_j : q_j = 1\}$ {list of prices of *on* machines}
Sort the set M by y_j in increasing order
Compute largest k such that

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

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 $(x_1, \ldots, x_m)$

$$x_j \geq \frac{y_j \phi_i}{1 - \phi_i}$$

Paper 2

- Providers model → Leader

Profit maximize

Resource state(on/off)

A non-negative bid of user *i* for resource *j*

$$\max \quad P(q_1, \ldots, q_m) = \sum_{j=1}^{m} (q_j \sum_{i=1}^{n} x_{ij}) - \sum_{j=1}^{m} q_j c_j$$

The actual cost of maintaining the infrastructure

Paper 2

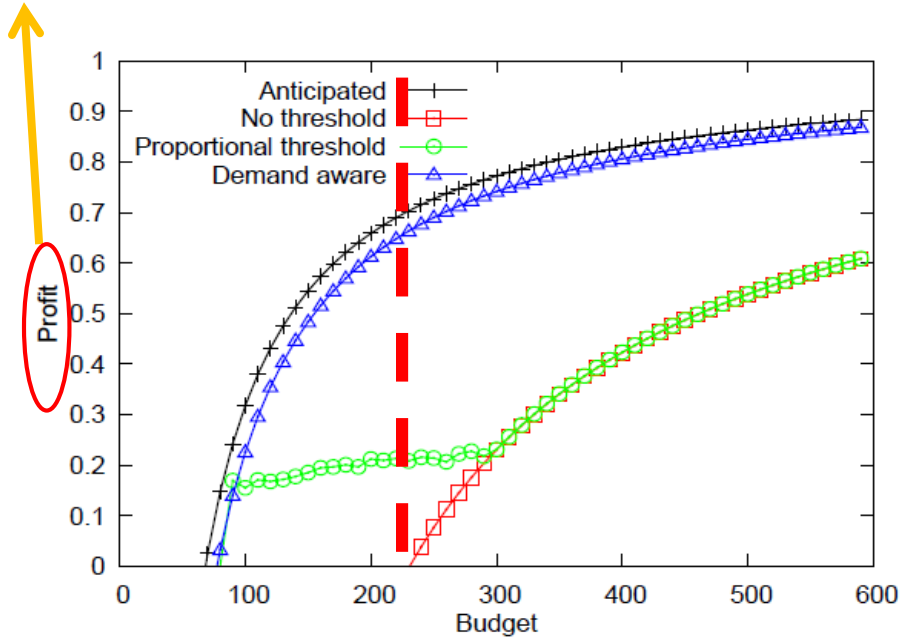
Fig. 2. Provider's best response algorithm

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(ϕ_i, τ_i, X_i)
end for
until convergence
until $\forall_i |S_i| \geq \tau_i$ OR $k = m$
return $(q_1, \ldots, q_m)$

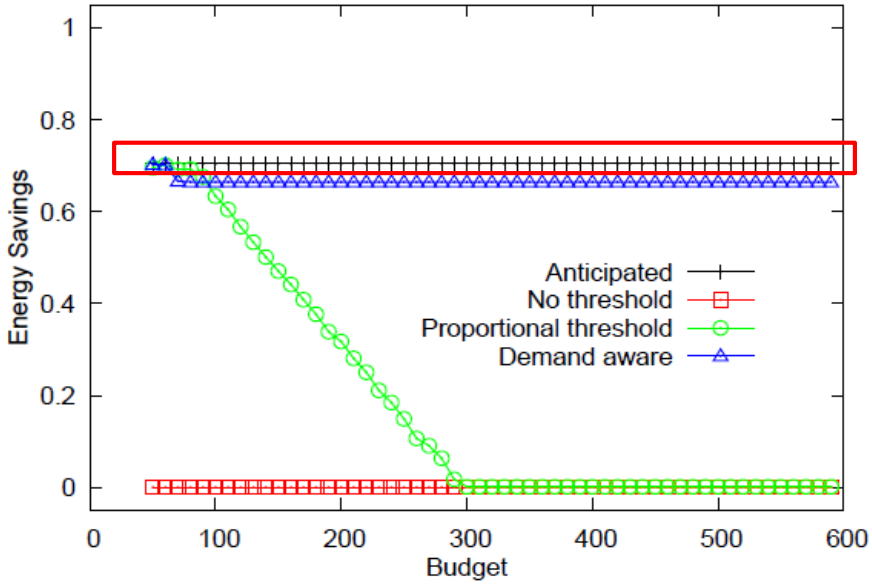
Paper 2

The number m of nodes = 1000
The number n of users = 250

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



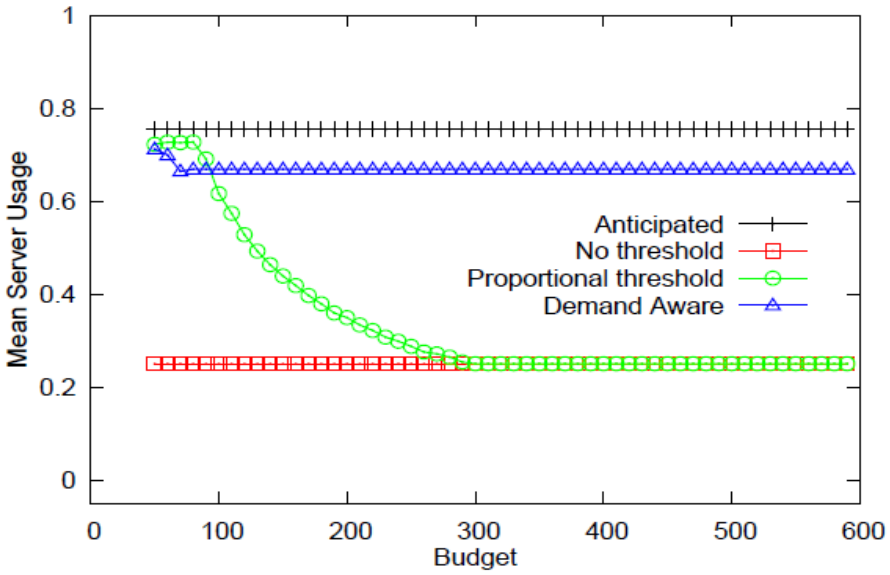
(a) Profit



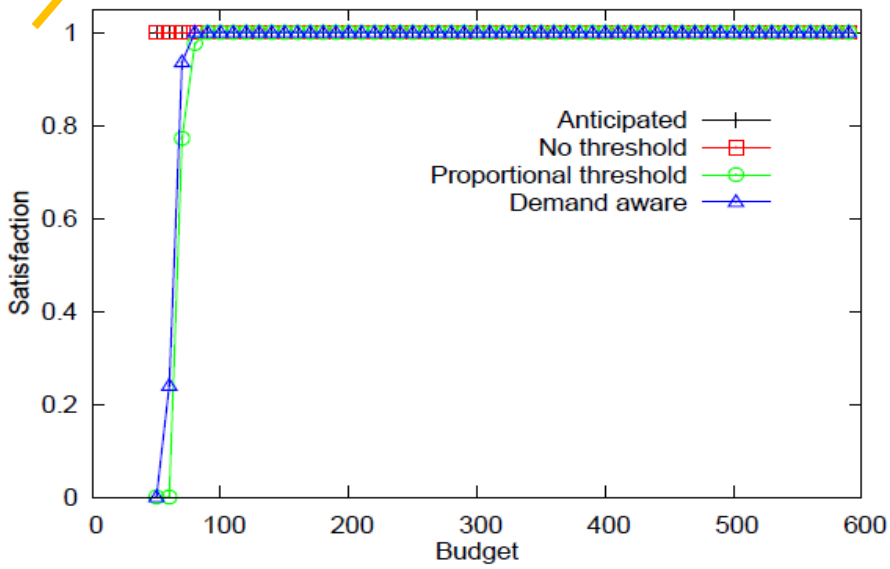
(b) Energy Savings

Paper 2

The ratio between the number k of satisfied users and the total number n of users



(c) Server Usage



(d) Satisfaction

HERO: Hierarchical energy optimization for datacenter networks

Yan Zhang ; Ansari, N.; Communications (ICC), 2012 IEEE International Conference on

Paper 3

- 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

Paper 3

- 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

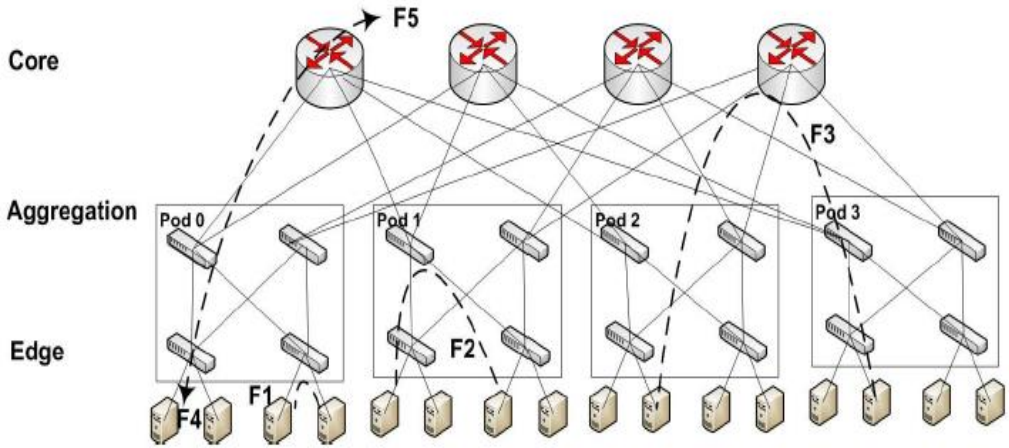


Fig. 1: Data center network topologies and traffic patterns.

Paper 3

- 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

Paper 3

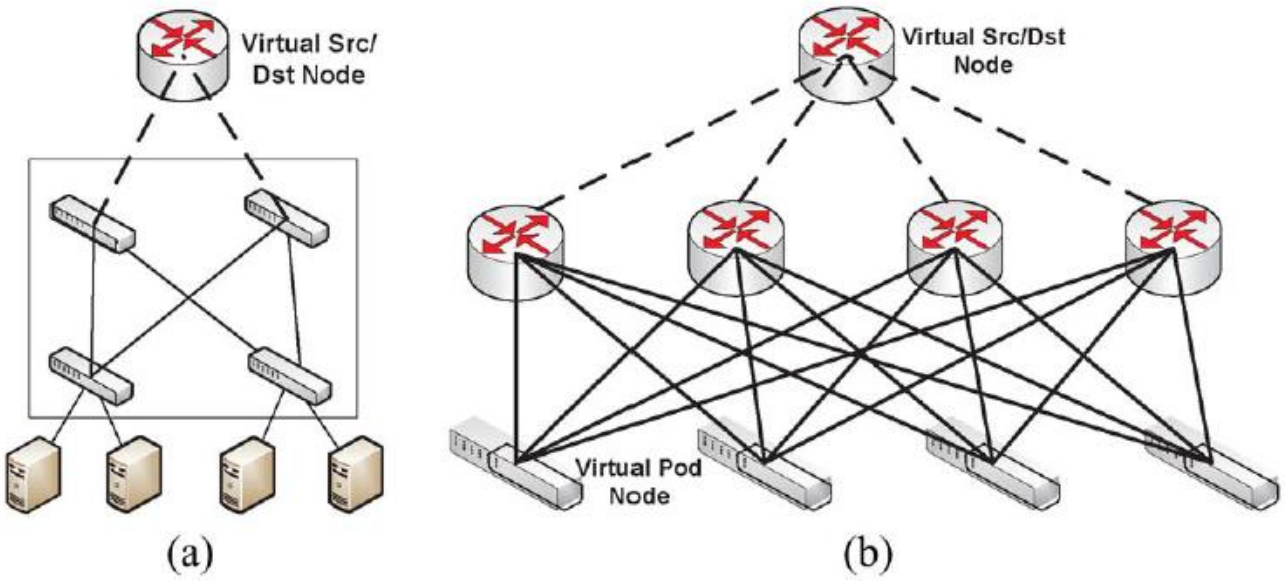


Fig. 2. Subnetwork topologies of a 4-ary fat-tree network. (a) Pod-level subgraph. (b) Core-level subgraph.

The potential benefit of hierarchical energy optimization is to simplify energy optimization problem by reducing optimization variables greatly

Paper 3

- 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

Paper 3

- 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

$$P^{p_m}_{total} = \sum_{i=1}^{N^p_m}\sum_{j=1}^{N^p_m} x_{ij} P^{L}_{ij} + \sum_{i=1}^{N^p_m} y_i P^{N}_{i}$$

Paper 3

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.

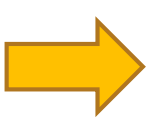
Stage 3: Solve the pod-level CMCF optimization problem.

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.

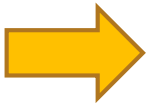
end for

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



Core-level

Determine the state of switches and links



Pod-level

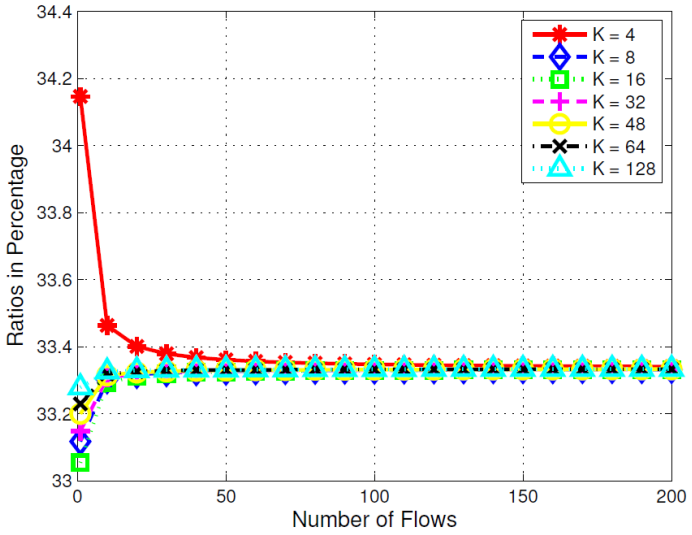
Paper 3

- 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

Paper 3

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



(a) The ratio of the total number of variables.

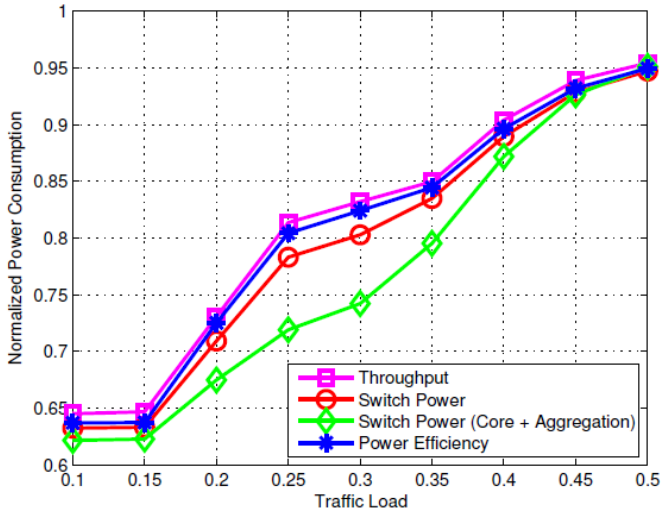


Fig. 5: The power consumption results of different heuristic algorithms.

Simplify energy optimization problem and decrease complexity

Paper 3

The power consumptions of HERO and the non-hierarchical model are almost the same under different traffic loads.

Large Traffic Flows

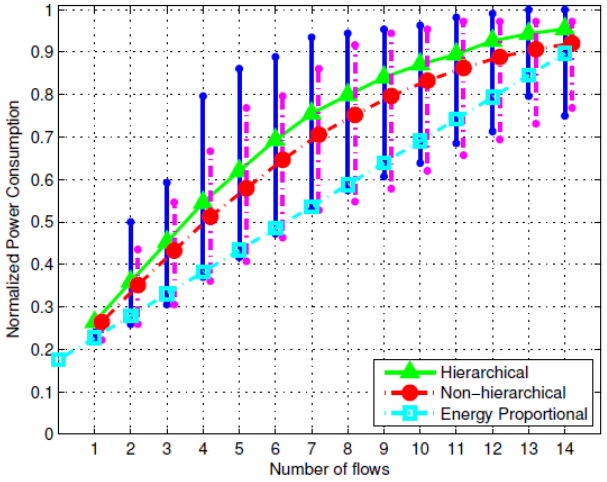


Fig. 3: The power consumption of 4-ary Fat-tree data center networks with different number of traffic flows.

Small Traffic Flows

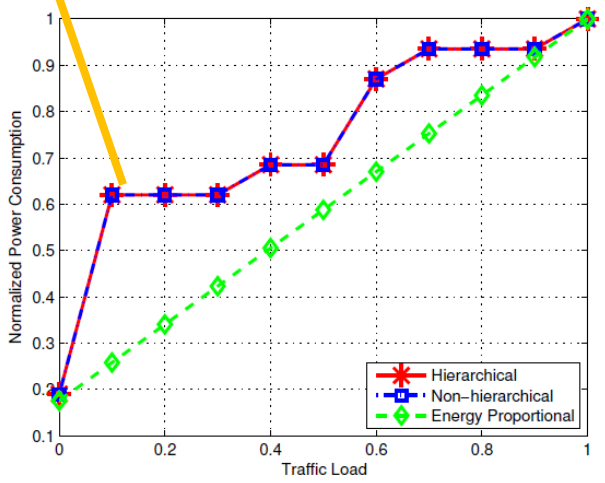


Fig. 4: The power consumption of a 4-ary Fat-tree data center network with all-to-all traffic under different traffic load.

Conclusion

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

Conclusion

	Power Optimization	Resource Allocation	Load Balance	QoS	Fault Tolerance
[1]	●	●	●		●
[2]	●	●	●		●
[3]	●			●	●