E2GC2 Workshop, 2009/10/13 (Tue)

Power-Performance Modeling of Heterogeneous Cluster-Based Web Servers

Hiroshi Sasaki†, Takatsugu Oya†, Masaaki Kondo‡, and Hiroshi Nakamura†

† The University of Tokyo
‡ The University of Electro-Communications

Background

- Small to Large-scale internet services everywhere
  - Parallel computation of a bunch of requests
  - Throughput oriented computing

    We focus on *web servers*
- *Cluster-based* web servers
  - *Heterogeneous* clusters are popular

- Importance of *low power* computing
  - Computation cost, cooling cost, ...

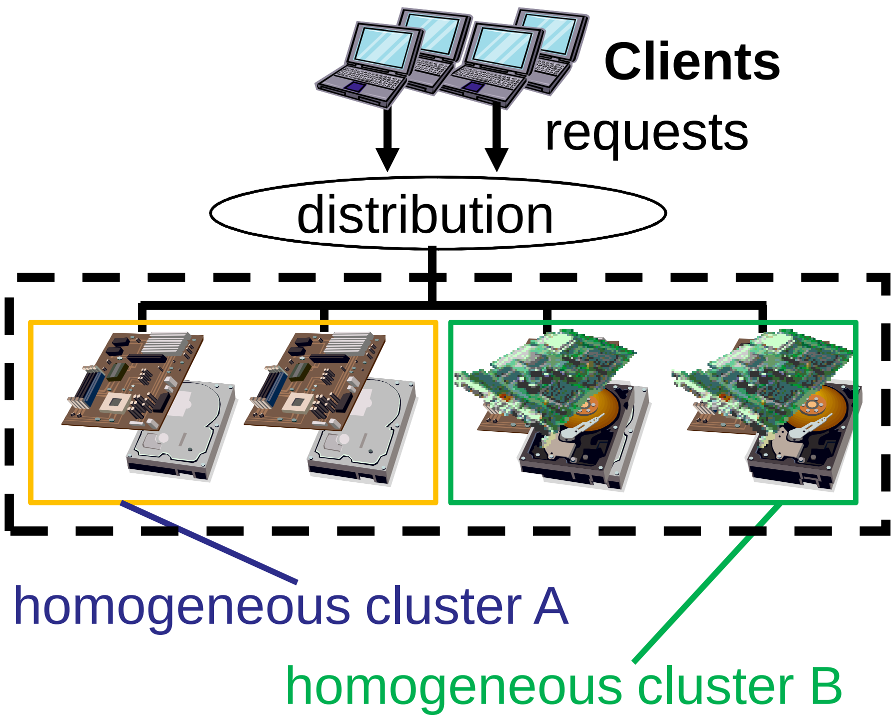
Characteristics of web services

- Requests
  - CPU-bound
    - Dynamic files (CGI, php, Java servlets...)
  - Disk-bound
    - Static files (html web pages, jpeg photos, tar balls...)
- Response time restrictions
  - Guarantee comfortable web services
  - Web servers must be able to handle max loads

Server configuration

- Basic configuration
  - Front-end: handle and distribute the requests, reconfiguration (# of nodes, level of frequencies)
  - Back-end: execute the requests

- Back-end servers
  - Composed from several homogeneous clusters



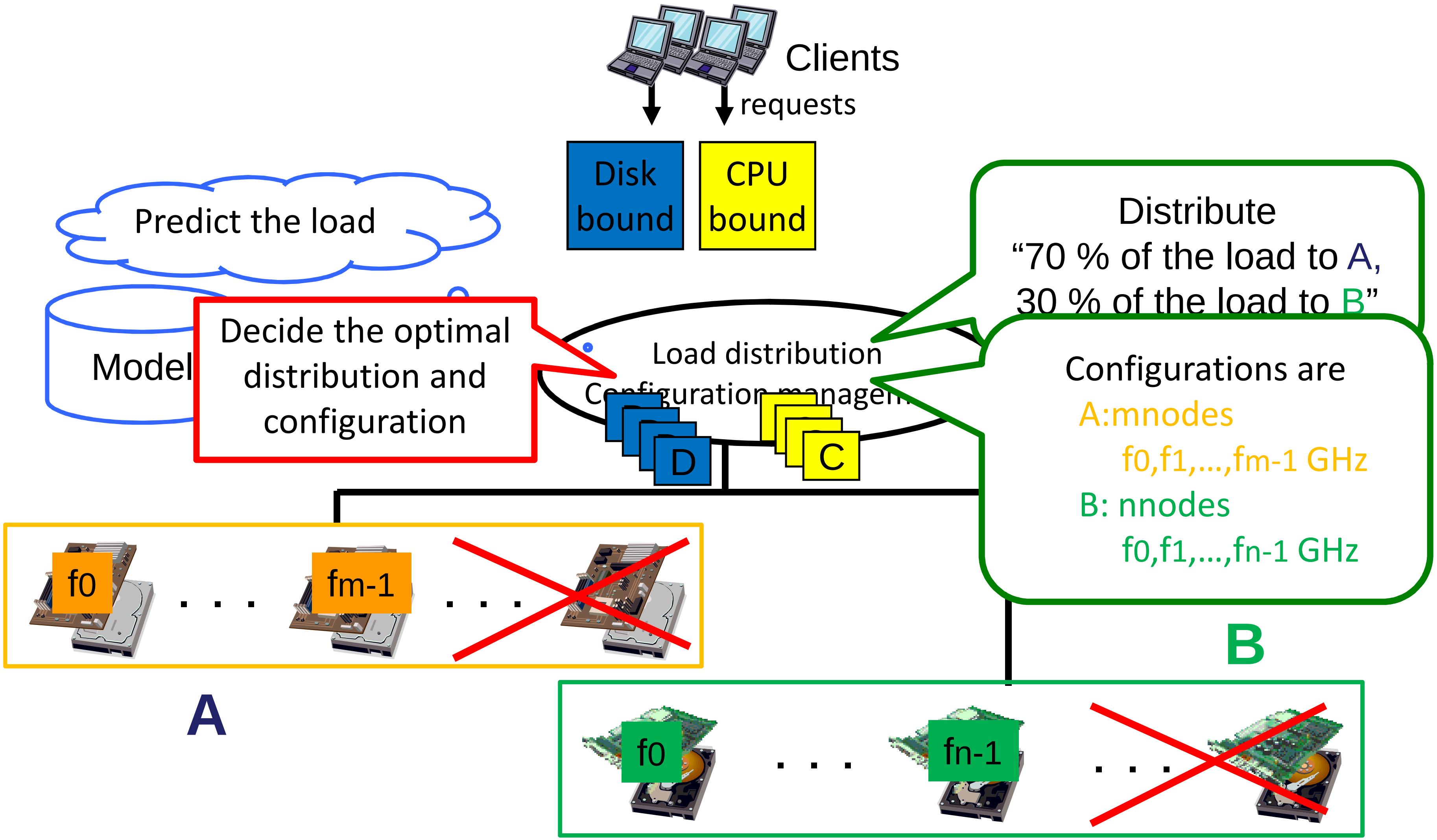
Objective of this work

- Power reduction of heterogeneous cluster-based web servers

Requests: CPU-bound and Disk-bound

- Satisfy the response time
- Minimize the power consumption
- Dynamically select the optimal configuration (# of nodes, levels of frequencies)

Overall picture



Overview of the proposed technique

1. Power-performance modeling
   - Performance model
     - ✧ How much load can a certain configuration handle within the response time restriction?
   - Power model
     - ✧ How much power will a certain configuration consume?

   Constructing a model for a single node is enough
     - ✧ All the requests are parallel
     - ✧ Power and performance are just a sum
2. Derive the optimal configuration
   - Homogeneous -> heterogeneous
   - Mathematically derive from the constructed model

Modeling: load definition

- What is a load?
  - CPU-bound requests: the time to execute a page (ms)
  - Disk-bound requests: the size of a page (KB)

- To handle it more effectively, we define the load as a single dimensional value
  - Actual amount of requests/max amount of requests
  - CPU-bound load: $Load_C$; Disk-bound load: $Load_D$
  - $0 <= Load <= 1$

Performance modeling

- Below are two equations a CPU should satisfy to execute both $Load_C$ and $Load_D$ simultaneously
  - CPU

    Performance_for_$Load_C$(= f1($Load_C$))

    + Performance_for_$Load_D$(= f2($Load_D$))

    <= CPU performance
  - Memory bus

    Bandwidth_for_$Load_C$(= g1($Load_C$))

    + Bandwidth_for_$Load_D$(= g2($Load_D$))

    <= Memory bus bandwidth

    Details in the paper...

Power modeling

- Power = Base power
  $+ \text{Power\_for\_Load}_{C}(= F(\text{Load}_{C}))$
  $+ \text{Power\_for\_Load}_{D}(= G(\text{Load}_{D}))$

Optimization *(homogenous)*

- For a given amount of load, the optimal configuration is to

Distribute the load equally to every nodes

$\Downarrow$

All the frequencies will be the same

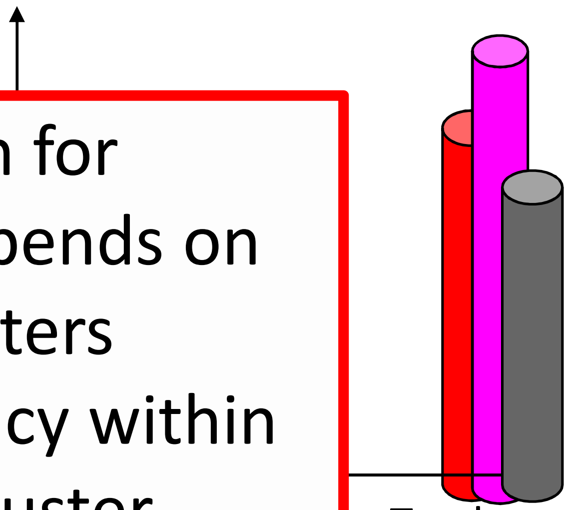
Details in the paper...

Challenges for optimizing heterogeneous cluster

Total load 1.0

Power

Distribution



Optimal configuration for heterogeneous cluster depends on

- o # of homogeneous clusters
- o # of nodes and frequency within each homogeneous cluster
- o Distribution ratio of the load

-> Derive from the model

# of nodes	32	3	. . .
frequency	0.8G 2.0G	1.8G 1.4G	. . .

Optimization (heterogenous)

- Unknowns: Distribution ratio of the load, # of nodes and frequency within each homogeneous clusters
- Known: Load

I. Within homogeneous clusters
   i. Frequency (= f(Distribution ratio, # of nodes)):
      substitute a load (for single node) for performance model and derive the min frequency
   ii. # of nodes (= g(Distribution ratio)):
      substitute the frequency for power model and derive the # of nodes which minimizes the power

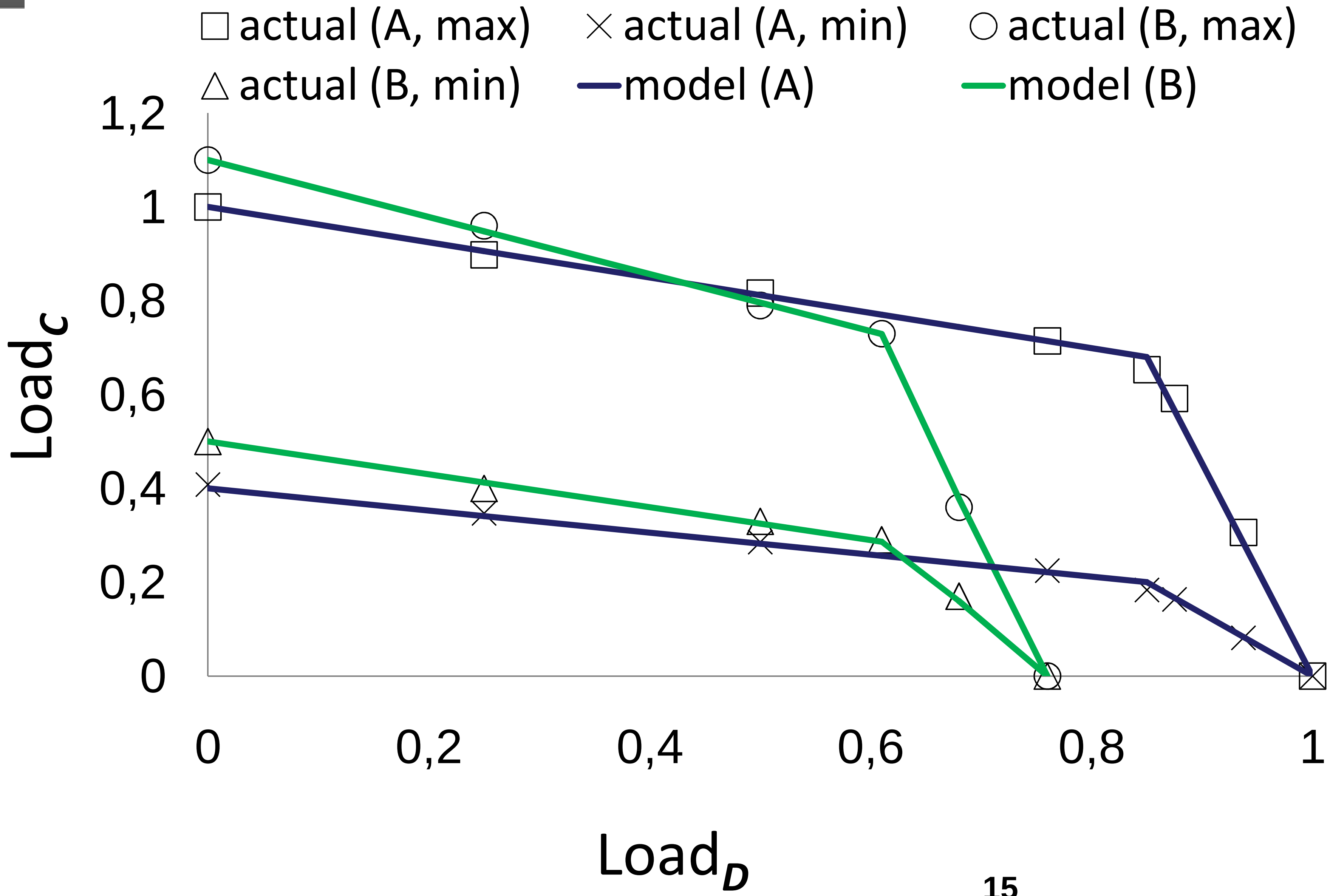
II. Derive the optimal distribution ratio that minimizes the sum of the power of each homogeneous clusters

Evaluation environment

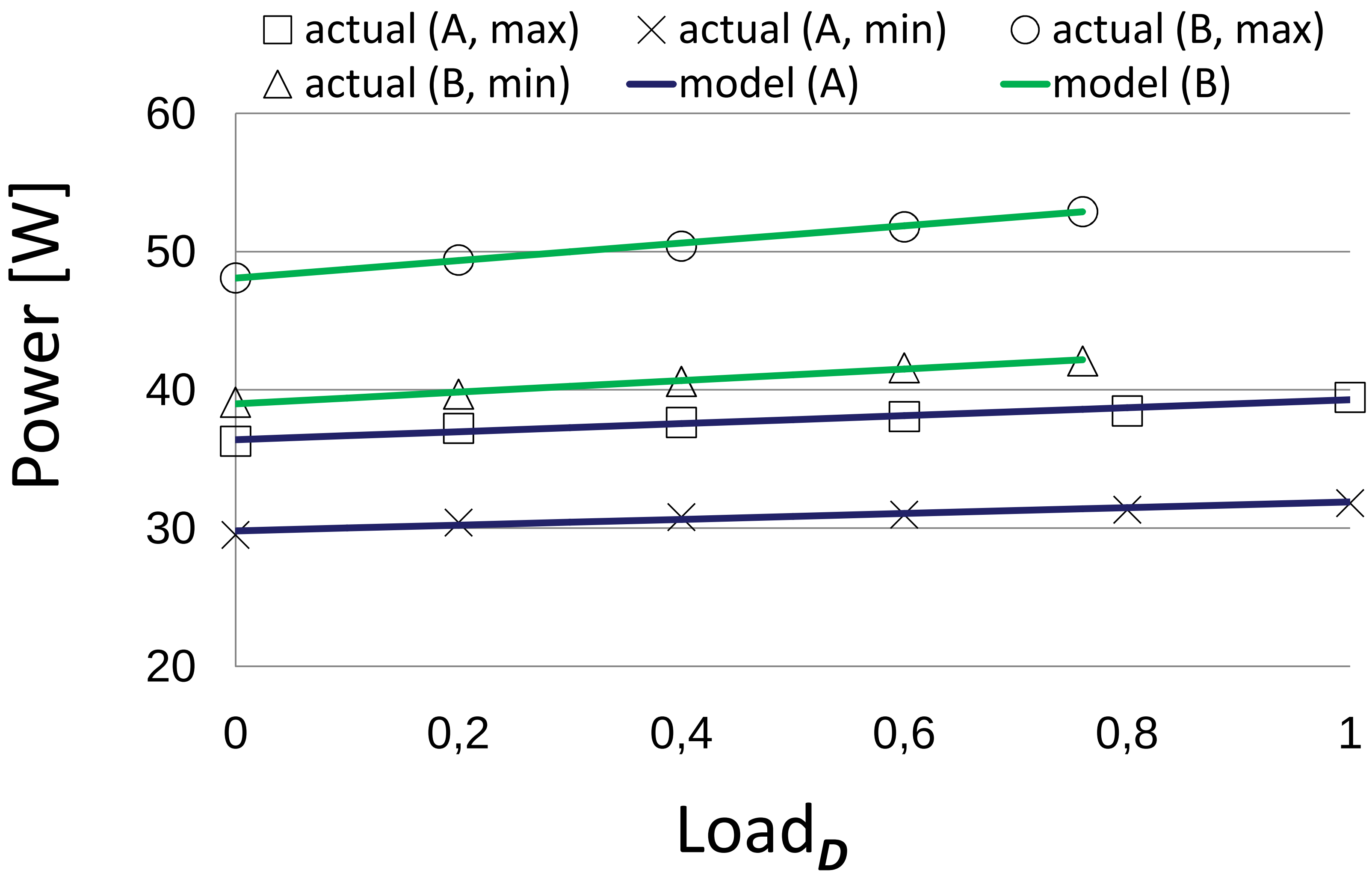
type	A	B
CPU	Intel Pentium M 760 (0.8-2.0 GHz)	AMD Opteron 150 (1.0-2.2 GHz)
memory	DDR2-SDRAM 1GB PC2-4300	DDR-SDRAM 1GB PC-3200
Disk	80GB 7200rpm SATA3.0GB/s seek time 8.8ms	80GB 7200rpm SATA3.0GB/s seek time 8.8ms
OS	Linux kernel-2.6.11	Linux kernel-2.6.16
ServerSW	Apache 2.2.3	Apache 2.2.3

- Clients: httperf 0.8 (by HP)
- Loads: CPU-bound (cgi), Disk-bound (html)
- Response time restriction: 200ms for both types of loads

Validation: performance model



Validation: power model

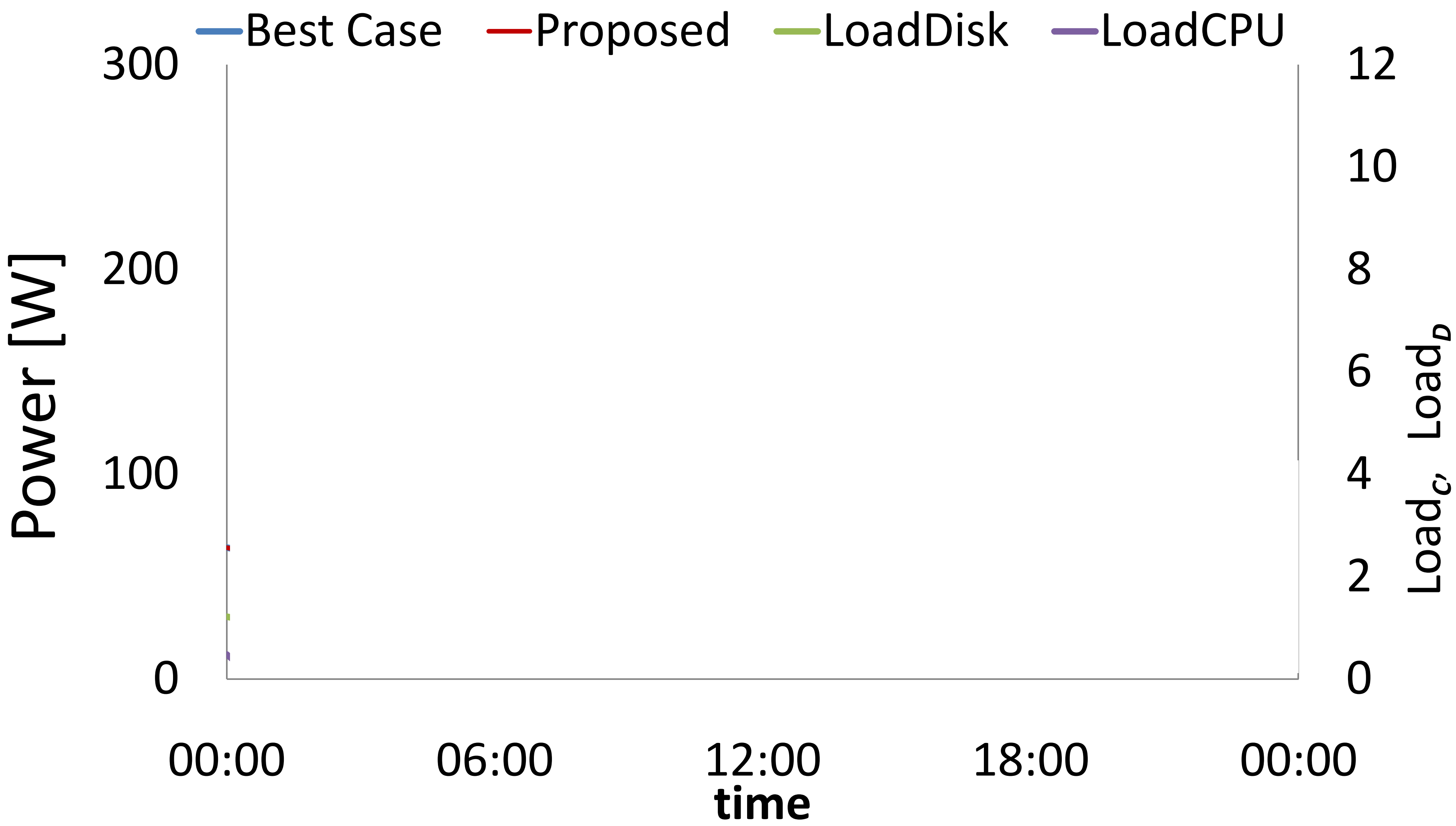


- Coefficients are in the paper

Evaluation

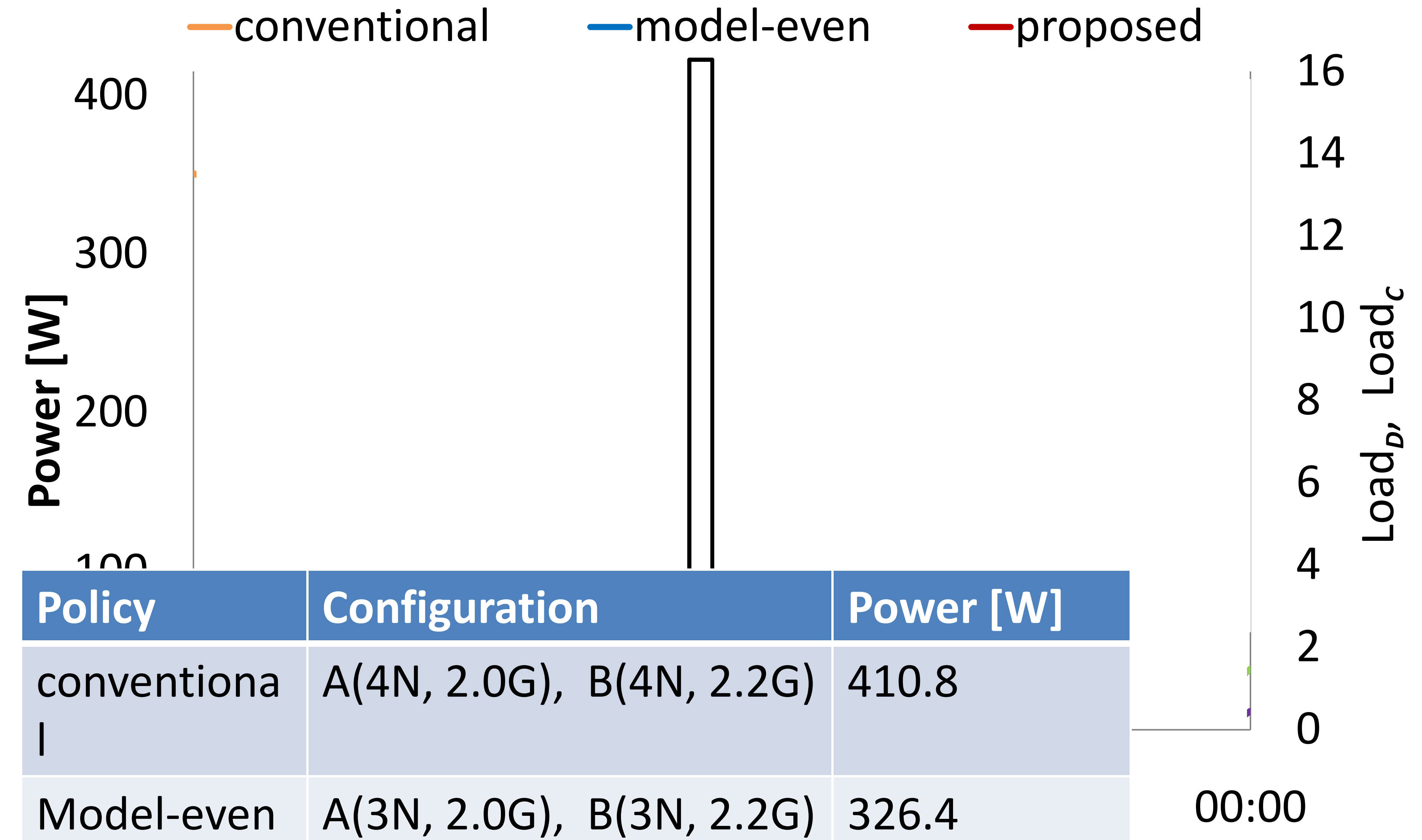
1. Optimizing within homogeneous cluster
   - Best case vs. proposed (derived from the model)
2. Optimizing heterogeneous cluster
   - Compare the three policies below
   1. ***Conventional***
      Load: distribute equally
      Configuration: all nodes are on and max frequency
   2. ***Model-even***
      Load: distribute equally
      Configuration: derive from the model
   3. ***Proposed***
      Load: derive from the model
      Configuration: derive from the model

Result 1 (A: 8 nodes)



- Proposed mostly (94%) selected the optimal configuration

Result 2 (A: 4 nodes B: 4 nodes)



Policy	Configuration	Power [W]
conventional	A(4N, 2.0G), B(4N, 2.2G)	410.8
Model-even	A(3N, 2.0G), B(3N, 2.2G)	326.4
proposed	A(4N, 2.0G), B(1N, 1.4G)	276.2

Conclusions and future work

- Conclusions
  - Objective: power reduction of a heterogeneous cluster-based web servers
  - Constructed a power-performance model
  - Derived the configuration from the model
    - Showed that proposed technique can reduce significant power
- Future work
  - Control the power and performance of other devices (HDD, DRAM Memory, ...)
  - Implement our technique in the OS (power on/off, suspend, dynamic prediction, recovery from mispredictions, ...)