Models and Control Strategies for Data Center Energy Efficiency

Luca Parolini
Bruce Krogh
Bruno Sinopoli
Electrical and Computer Engineering
Carnegie Mellon University

Special Thanks to Zhikui Wang, HP Labs

LCCC Workshop
Lund, Sweden
December 5-7, 2012
Data center examples

- **Facebook's data center in North Carolina, US**
  - 450$ million project
  - ~28,000 m² (300,000 ft²)
  - Operated by 35 - 45 full-time employees

- **Racks**
  - Contain 42 (1U) servers in a rack
  - 1 server: 480mm x 800mm x 44mm

<table>
<thead>
<tr>
<th></th>
<th><strong>Idle power</strong></th>
<th><strong>Peak power</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>200 W</td>
<td>350 W</td>
</tr>
<tr>
<td>Rack</td>
<td>8.4 kW</td>
<td>~15 kW</td>
</tr>
</tbody>
</table>

http://www.datacenterknowledge.com/
Electricity consumption

(Billion of Kwh / year)


Global

US

C. L. Belady, Projecting annual new datacenter construction market size, Microsoft, 2011
Monthly operating cost

- **Large-scale facility, 50K servers**
  - Facility cost amortized over 10 years
  - Server cost amortized over 3 years
  - Servers account for 70% of total power consumption

- **Total: 3.5 M$**

  - Server (hardware cost) 57%
  - Networking equipment 8%
  - Power distribution and cooling 18%
  - Powering cost 13%
  - Other infrastructure 4%

Monthly operating cost

- **Large-scale facility, 50K servers**
  - Facility cost amortized over 10 years
  - Server cost amortized over 3 years
  - Servers consume 70% of total power consumption

**Research focus**
- Server (hardware cost) 57%
- Networking equipment 8%
- Power distribution and cooling 18%
- Powering cost 13%
- Other infrastructure 4%
- **Total: 3.5 M$**

Temperature distribution

$T_{\text{max}} = 36.6^\circ \text{C}$

Server racks

Cold aisles

Hot aisle

Hot aisles

CRACs (Computer Room Air Conditioners)

Outline

- Introduction
- Control-oriented model
- Control strategies
- Simulation results
- Conclusion and future work
Hierarchical control approach

- **Data center level (hr)**
- **Zone level (min)**
- **Intra-zone level (sec)**

Diagram showing the hierarchical control approach in a data center.

- **Predictor**
- **Data center controller**
- **Zone controller**
- **Intra-zone controller**
- **CRAC controller**
- **CRAC unit**
- **Environment**
- **Disturbances**
Zone level control

- Operates at a faster time-scale
- Decides how many servers in the zone should be turned on

Control actions based on
- Desired workload execution rate (predictive control)
- Current resource use in the zone (reactive control)

Consider:
- Time to turn servers on: $T_s$
- Variability of workload arrival rate

Zone level control - simulation results

- Compare proposed control approach against optimal open-loop strategy
  - $\tau$: expected time interval over which the workload arrival rate is constant

![Graph showing power consumption improvement vs. $T_s / \tau$](graph.png)

- Higher is better

Proposed control strategy
Open-loop strategy (normalized)
Zone level control - simulation results

- Compare proposed control approach against optimal open-loop strategy
  - \( \tau \): expected time interval over which the workload arrival rate is constant

- Workload arrival rate varies too fast with respect to the time to turn servers on
- Best solution is to never turn servers on and off
Modeling approach: Data Center Level

- Focuses on processes in the hours time scale

- Groups servers into zones
  - Power consumption of a zone is proportional to the amount of workload executed
  - Data are always available
  - Neglect the time to turn servers On and Off
    - Much shorter than the controller sampling-time

- Considers
  - Computational and thermal dynamics
  - Nonlinear efficiency of the CRAC units
    - Service level agreements (SLAs) with users and the power-grid
Computational network

- Describes the evolution of resource usage in the data center
- Based on a fluid approximation of the workload execution and arrival processes
  - First-order approximation of a queuing system

![Diagram showing workload arrival rate and zones](image)
Computational network

- Describes the evolution of resource usage in the data center
- Based on a fluid approximation of the workload execution and arrival processes
  - First-order approximation of a queuing system

\[ \lambda(\tau) s_1(\tau) \]

\[ \lambda(\tau) \]

Workload arrival rate (job/s)

\[ \lambda(\tau) s_N(\tau) \]

Scheduling action

Zone 1

Zone N
Computational network

- Describes the evolution of resource usage in the data center
- Based on a fluid approximation of the workload execution and arrival processes
  - First-order approximation of a queuing system

\[ \lambda(\tau) s_1(\tau) \rightarrow l_1(\tau) \rightarrow [\mu_1(\tau)] \]

\[ \lambda(\tau) \]

Workload execution rate (job/s)

\[ \lambda(\tau) s_N(\tau) \rightarrow l_N(\tau) \rightarrow [\mu_N(\tau)] \]

Desired workload execution rate
Computational network

- Describes the evolution of resource usage in the data center
- Based on a fluid approximation of the workload execution and arrival processes
  - First-order approximation of a queuing system

\[
\begin{align*}
\lambda(\tau) s_1(\tau) &\quad \rightarrow \quad l_1(\tau) \\
\lambda(\tau) s_N(\tau) &\quad \rightarrow \quad l_N(\tau)
\end{align*}
\]

- Workload arrival rate (job/s)

\[
\begin{align*}
al_1(\tau) &= \lambda(\tau) s_1(\tau) \\
d_1(\tau) &= \eta_1(\tau) \\
\dot{l}_1(\tau) &= a_1(\tau) - d_1(\tau) \\
\eta_1(\tau) &= \begin{cases} 
\mu_1(\tau) & \text{if } l_1(\tau) > 0 \\
\text{or } a_1(\tau) > \mu_1(\tau) \\
a_1(\tau) & \text{otherwise}
\end{cases}
\end{align*}
\]
Computational network

- Describes the evolution of resource usage in the data center
- Based on a fluid approximation of the workload execution and arrival processes
  - First-order approximation of a queuing system

\[
\begin{align*}
\lambda(t) & \rightarrow \sum_{j=1}^{N} \xi_{i,j}(\tau) \rightarrow \sum_{j=1}^{N} \left[ \delta_{i,j}(\tau) \right] \rightarrow l_1(\tau) \rightarrow \eta_1(\tau) \\
\lambda(t) & \rightarrow \lambda(\tau) s_1(\tau) \\
\end{align*}
\]

- \( \lambda(t) \) is the workload arrival rate (job/s)
- \( \lambda(\tau) s_1(\tau) \) is the desired workload migration rate
- \( \sum_{j=1}^{N} \xi_{i,j}(\tau) \) is the workload execution rate
- \( \sum_{j=1}^{N} \left[ \delta_{i,j}(\tau) \right] \) is the migration rate

\[
\begin{align*}
a_1(\tau) &= \lambda(\tau) s_1(\tau) + \sum_{j=1}^{N} \xi_{1,j}(\tau) \\
d_1(\tau) &= \eta_1(\tau) + \sum_{j=1}^{N} \xi_{j,1}(\tau) \\
l_1(\tau) &= a_1(\tau) - d_1(\tau) \\
\eta_1(\tau) &= \begin{cases} 
\mu_1(\tau) & \text{if } l_1(\tau) > 0 \\
\text{or } a_1(\tau) > \mu_1(\tau) \\
a_1(\tau) & \text{otherwise}
\end{cases}
\end{align*}
\]
Thermal network

\[ T_{in} = \Psi T_{out} \]

Q. Tang et al. “Sensor-based fast thermal evaluation model for energy efficient high-performance data centers”, 2006
Thermal network

Inlet temperature constraint \( T_{in}(\tau) = \Psi T_{out}(\tau) \leq T_{in} \)

Heat removed rate (W)

CRAC power consumption 
\[ p_i(t) = \frac{\dot{Q}_i(t)}{COP_i(T_{out,i}(t))} \]

Q. Tang et al. “Sensor-based fast thermal evaluation model for energy efficient high-performance data centers”, 2006
# Variables at step $k$

<table>
<thead>
<tr>
<th>Input</th>
<th>Controllable</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Job scheduling</td>
<td>$s(k)$</td>
</tr>
<tr>
<td></td>
<td>Resource allocation</td>
<td>$\mu(k)$</td>
</tr>
<tr>
<td></td>
<td>Job migration</td>
<td>$\delta(k)$</td>
</tr>
<tr>
<td></td>
<td>CRAC unit reference temperature</td>
<td>$T_{\text{ref}}(k)$</td>
</tr>
<tr>
<td>Uncontrollable</td>
<td>Job arrival</td>
<td>$\lambda(k)$</td>
</tr>
</tbody>
</table>

| Output         | Zone power consumption                      | $p_N(k)$        |
|                | Power consumption of CRAC nodes             | $p_c(k)$        |
|                | Input temperatures of zones                | $T_{\text{in}}(k)$ |

| State          | Number of jobs in every zones              | $l(k)$          |
|                | Output temperatures of CRACs and zones     | $T_{\text{out}}(k)$ |
Outline

- Introduction
- Control-oriented model
- Control strategies
- Simulation results
- Conclusion and future work
Control strategies

- **Baseline**
  - Open-loop strategy
  - Sets control variables for the worst-case scenario

- **Uncoordinated**
  - Manages the computational and the thermal resources independently
  - Neglects the thermal-computational coupling in the data center

- **Coordinated**
  - Manages the computational and the thermal resources in a single optimization problem
Baseline & uncoordinated approaches

- **Baseline**

\[
\mu(\tau) = \bar{\mu} \quad \delta(\tau) = 0 \quad s(\tau) = 1 \frac{1}{N} \quad T_{\text{ref}}(\tau) = T_{\text{ref}}
\]

- **Uncoordinated controller**

\[
\min_{\mathcal{M}, \mathcal{S}, \mathcal{D}} \sum_{h=k}^{k+T-1} 1^T \hat{p}_N(h|k)
\]

\[s.t. \quad \hat{l}(k|k) = l(k) \quad \text{for all } h = k, \ldots, k + T - 1\]

\[\text{computational dynamics} \]

\[\text{QoS constraints}\]

\[0 \leq \hat{\mu}(h|k) \leq \bar{\mu}, \quad \hat{\delta}(h|k) = 0, \quad 0 \leq \hat{s}(h|k) \leq 1, \quad 1^T \hat{s}(h|k) \leq 1\]

\[\mathcal{D} = \left\{ \hat{\delta}(k|k), \ldots, \hat{\delta}(k+T-1|k) \right\} \]

\[\mathcal{T}_{\text{ref}} = \left\{ T_{\text{ref}}(k|k), \ldots, T_{\text{ref}}(k+T-1|k) \right\} \]

\[\mathcal{M} = \left\{ \hat{\mu}(k|k), \ldots, \hat{\mu}(k+T-1|k) \right\} \]

\[\mathcal{S} = \left\{ \hat{s}(k|k), \ldots, \hat{s}(k+T-1|k) \right\} \]
Baseline & uncoordinated approaches

**Baseline**

\[
\mu(\tau) = \bar{\mu} \quad \delta(\tau) = 0 \quad s(\tau) = 1 \frac{1}{N} \quad T_{\text{ref}}(\tau) = T_{\text{ref}}
\]

**Uncoordinated controller**

\[
\min_{\mathcal{M}, \mathcal{S}, \mathcal{D}} \sum_{h=k}^{k+\mathcal{T}-1} 1^T \hat{p}_N(h|k)
\]

\[
\text{s.t.} \quad \hat{l}(k|k) = l(k)
\]

for all \( h = k, \ldots, k + \mathcal{T} - 1 \)

computational dynamics

QoS constraints

\[
0 \leq \hat{\mu}(h|k) \leq \bar{\mu}, \quad \hat{\delta}(h|k) = 0
\]

\[
0 \leq \hat{s}(h|k) \leq 1, \quad 1^T \hat{s}(h|k) \leq 1
\]

\[
\mathcal{D} = \left\{ \hat{\delta}(k|k), \ldots, \hat{\delta}(k+\mathcal{T}-1|k) \right\} \quad \mathcal{T}_{\text{ref}} = \left\{ \hat{T}_{\text{ref}}(k|k), \ldots, \hat{T}_{\text{ref}}(k+\mathcal{T}-1|k) \right\}
\]

\[
\mathcal{M} = \left\{ \hat{\mu}(k|k), \ldots, \hat{\mu}(k+\mathcal{T}-1|k) \right\} \quad \mathcal{S} = \left\{ \hat{s}(k|k), \ldots, \hat{s}(k+\mathcal{T}-1|k) \right\}
\]

\[
\min_{\mathcal{T}_{\text{ref}}} \sum_{h=k}^{k+\mathcal{T}-1} 1^T \hat{p}_C(h|k)
\]

\[
\text{s.t.} \quad \hat{T}_{\text{out}}(k|k) = T_{\text{out}}(k)
\]

for all \( h = k, \ldots, k + \mathcal{T} - 1 \)

thermal dynamics

\[
T_{\text{ref}} \leq \hat{T}_{\text{ref}}(h|k) \leq \bar{T}_{\text{ref}}
\]

\[
\hat{T}_{\text{in}}(h+1|k) \leq \overline{T}_{\text{in}}
\]
Coordinated approach

- Considers the computational and the thermal dynamics in the same optimization problem

\[
\min_{M,S,D,T_{\text{ref}}} \left( \sum_{h=k}^{k+T-1} 1^T \hat{p}_N(h|k) + 1^T \hat{p}_C(h|k) \right)
\]

s.t. \( \hat{l}(k|k) = l(k), \quad \hat{T}_{\text{out}}(k|k) = T_{\text{out}}(k) \)
for all \( h = k, \ldots, k + T - 1 \)
computational dynamics, thermal dynamics,
QoS constraints,
\( 0 \leq \hat{\mu}(h|k) \leq \bar{\mu}, \quad \hat{\delta}(h|k) = 0 \)
\( 0 \leq \hat{s}(h|k) \leq 1, \quad 1^T \hat{s}(h|k) \leq 1, \)
\( T_{\text{ref}} \leq \hat{T}_{\text{ref}}(h|k) \leq \bar{T}_{\text{ref}}, \quad \hat{T}_{\text{in}}(h+1|k) \leq \bar{T}_{\text{in}}, \)
\( \hat{p}(h|k) = B_\eta \hat{\eta}(h|k) \)

Minimize expected zone and CRAC power consumption

Considers both thermal and computational dynamics

Thermal-computational coupling
Outline

- Introduction
- Control-oriented model
- Control strategies
- Simulation results
- Conclusion and future work
Simulation parameters

- **4 CRAC units**
  - Identical to each other
Simulation parameters

- **4 CRAC units**
  - Identical to each other

- **8 Zones**
  - 3 Racks each (126 servers per zone)
Simulation parameters

- **4 CRAC units**
  - Identical to each other

- **8 Zones**
  - 3 Racks each (126 servers per zone)
  - Energy efficient servers

![Diagram of CRAC units and zones with 4 CRAC units and 8 zones, each containing 3 racks and 126 servers.](image_url)
Simulation parameters

- 4 CRAC units
  - Identical to each other
- 8 Zones
  - 3 Racks each (126 servers per zone)
  - Energy efficient servers
  - Efficiently cooled servers
Simulation parameters

- 4 CRAC units
  - Identical to each other

- 8 Zones
  - 3 Racks each (126 servers per zone)
  - Energy efficient servers
  - Efficiently cooled servers

- Analyze the average power consumption for different workload arrival rates

- Modeling language: TomSym

- Solver: KNITRO 7.0
Total power consumption

- Lower is better

Baseline

- Total power consumption proportional to the average utilization
- Server power consumption proportional to the average utilization
- Fixed efficiency of CRAC units
Total power consumption

Power (kW)

Baseline

Uncoordinated

Lower is better

Average utilization
Total power consumption

Lower is better

Baseline

Coordinated

Uncoordinated
Total power consumption

Lower is better

Baseline

Coordinated

Uncoordinated

Cannot improve performance

Power (kW)

Average utilization

60 (kW)
Average reference temperatures

- Coordinated controller
- Uncoordinated controller
Total power consumption

- How do the controllers perform when all of the zones are efficiently cooled?

![Graph showing power consumption vs. average utilization]

No major improvements

Power (kW) vs. Average utilization
Total power consumption

- How do the controllers perform when large variability exists among the zone efficiency cooling?

![Graph showing power consumption vs. average utilization with large improvements highlighted.](image-url)
Cyber-Physical index

- Given a data center
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?
Cyber-Physical index

- **Given a data center**
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?

- **Cyber-Physical index (CPI), values in [0,1]**
  - When CPI is close to 1, then a coordinated approach is advisable
  - When CPI is close to 0, then an uncoordinated approach tends to be as efficient as a coordinated approach

- **CPI is function of the sensitivity of the zones with respect to variations of the workload departure rate and of reference temperatures**
Relative efficiency

Uncoordinated

Coordinated
Relative efficiency

Uncoordinated

Coordinated

Area: $U$
Relative efficiency

Relative efficiency = \frac{U - C}{U}

Area: \( U - C \)

Area: \( U \)

Uncoordinated

Coordinated
Relative savings

- Every point represents a different data center

Data center cases discussed
Relative savings

- Every point represents a different data center

An uncoordinated approach can be considered
A coordinated approach should be considered

Data center cases discussed
Interaction with the smart grid

- **Time-varying electricity price**
  - Used by the smart-grid to cap the average power consumption of the data center.
Run-time cost

- Difference between income due to the workload processing and the cost of powering the data center

- Depends on two service level agreements (SLAs)
  - SLA_U: sets the income based on the quality of service (QoS)
    - Approximated by the ratio between required and assigned hardware resources
  - SLA_G: sets the data center's powering cost
    - The energy cost is time-varying and power consumption dependent
Total data center power consumption

- High cost of dropping jobs
- Low cost of dropping jobs
Conclusion

- Discussed a control-oriented data center model
  - Considers both the computational and the physical characteristics of a data center
  - The model can be extended to consider electricity cost (interaction with the smart-grid) and other data center equipment

- Introduced two control strategies
  - Representative of different approaches to data center control

- Compared the performance of the control strategies under the same scenario

- Proposed a cyber-physical index
  - First attempt to characterize the thermal and computational characteristics of the data center within a single index

- Analyzed the impact of time-varying electricity prices
Future work

- Controllers can take advantage of service level agreements (SLAs) with both the users and the power grid
  - Uncoordinated control approach can be as optimal as the coordinated approach
  - Depending on the SLA with the grid, the data center may induce large variations on the real-time electricity price

- Given a data center, where should we locate its server so as to reduce its CPI?

- Feasibility and stability of coordinated and uncoordinated controller
  - The coordinated controller is always feasible, but does not lead to a stable equilibrium point
  - How should the cost function be formulated so that the closed-loop system has a stable, economically optimal, equilibrium point?