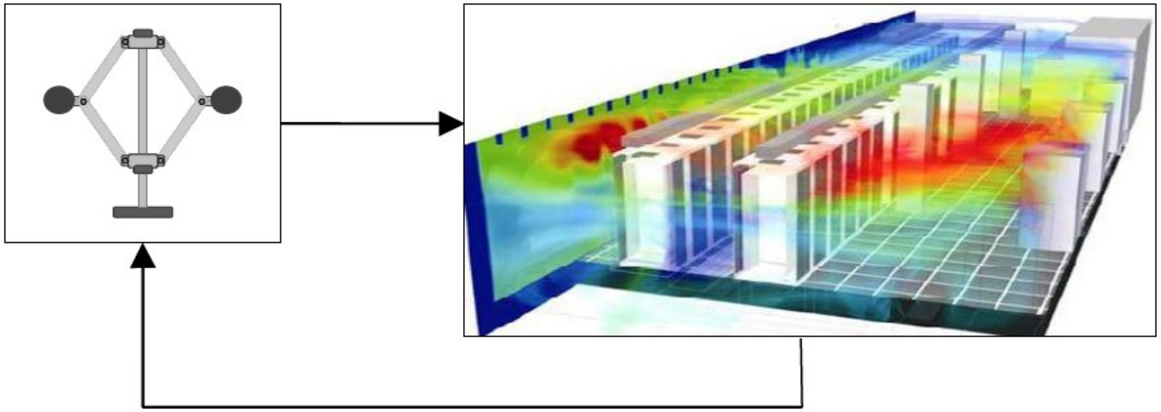
Models and Control Strategies for Data Center Energy Efficiency

Ph.D. Defense

Luca Parolini

Dept. of Electrical and Computer Engineering
Carnegie Mellon University



Feb. 27th 2012

Data center examples

- **Facebook's data center in North Carolina, US**
  - $ 450 million project
  - ~28,000 m^2 (300,000 ft^2)
  - Operated by 35 - 45 full-time employees

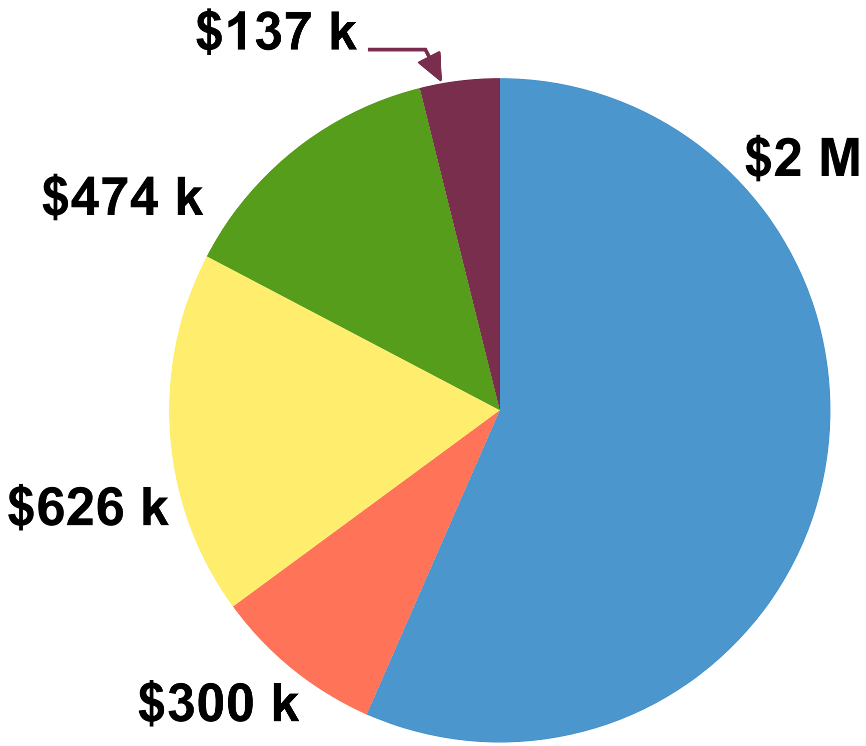
- **Racks**
  - Contain 42 (1U) servers in a rack
  - 1U server: 480mm x 800mm x 44mm

	Idle power	***Peak power***
Server	200 W	350 W
Rack	8.4 kW	~15 kW

http://blog.cluttr.be/2011/01/26/the-case-for-data-center-power-efficiency/
http://www.datacenterknowledge.com/

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



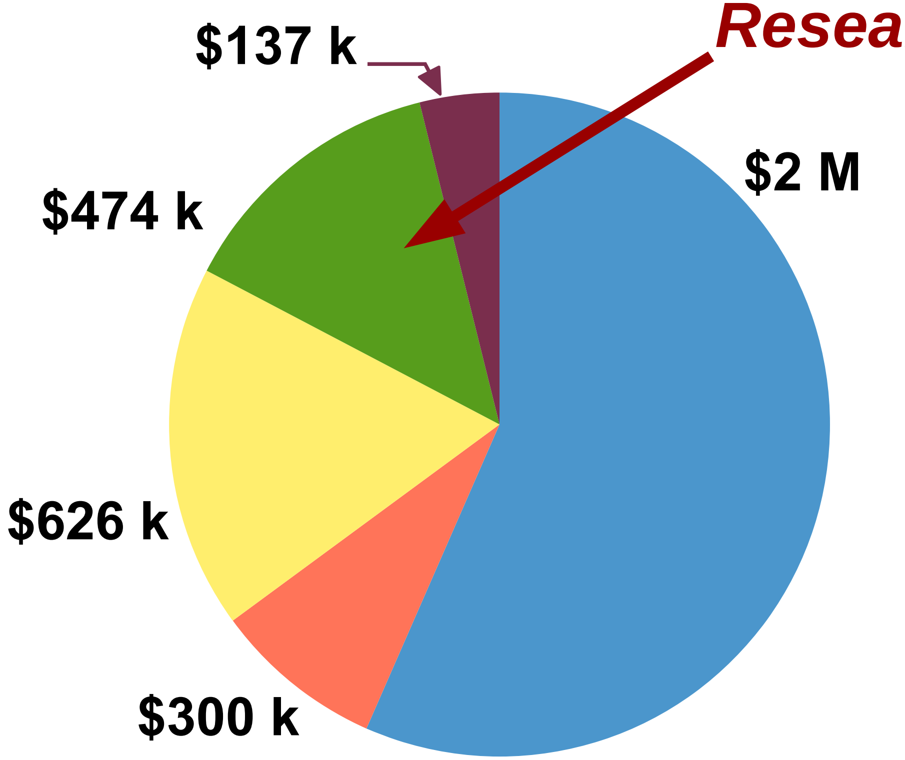
●	Server (hardware cost)	*57%*
●	Networking equipment	*8%*
●	Power distribution and cooling	*18%*
●	Powering cost	*13%*
●	Other infrastructure	*4%*

● **Total: $ 3.5 M**

James Hamilton, http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

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



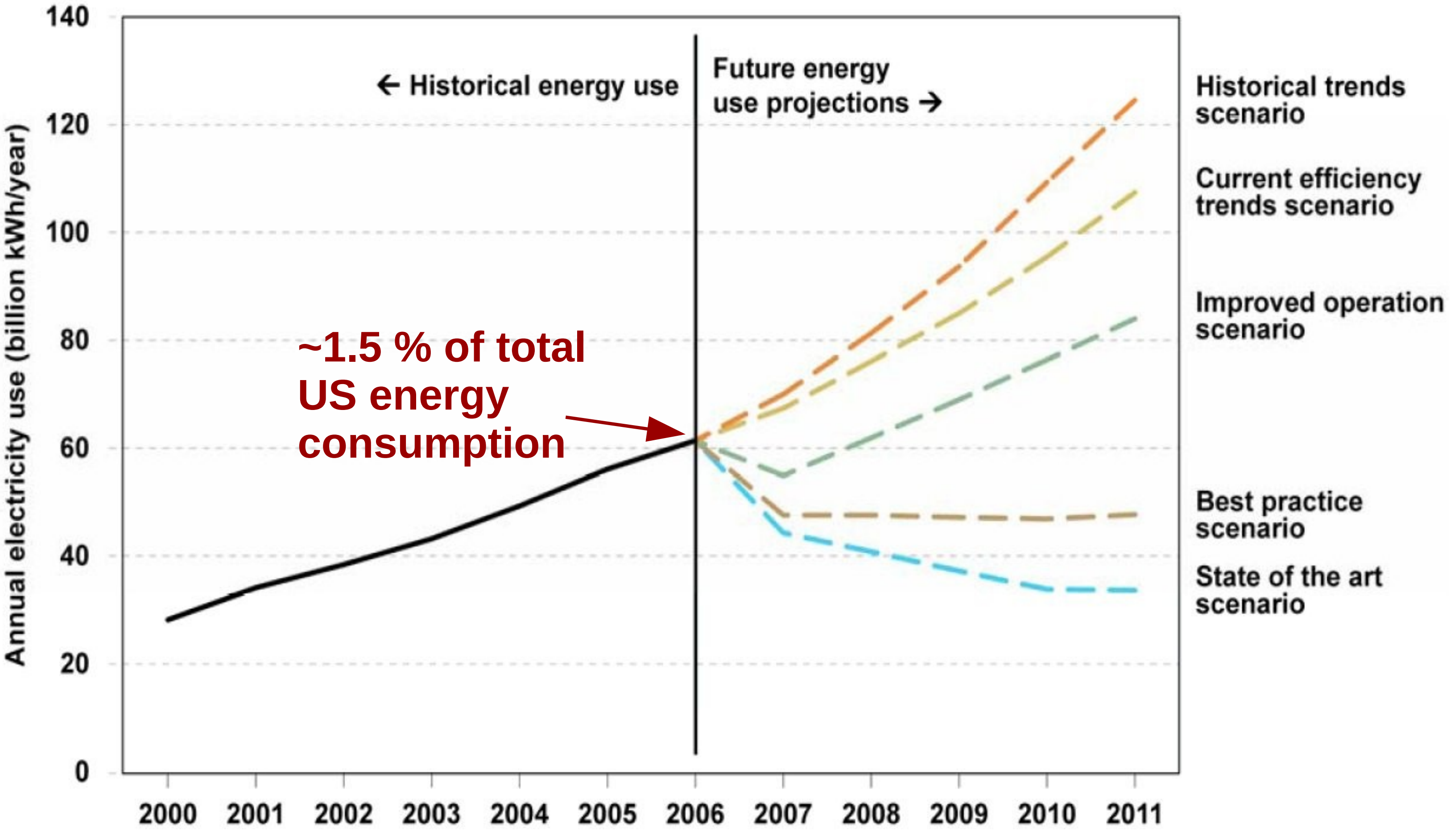
Research focus

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

● **Total: $ 3.5 M**

James Hamilton, http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

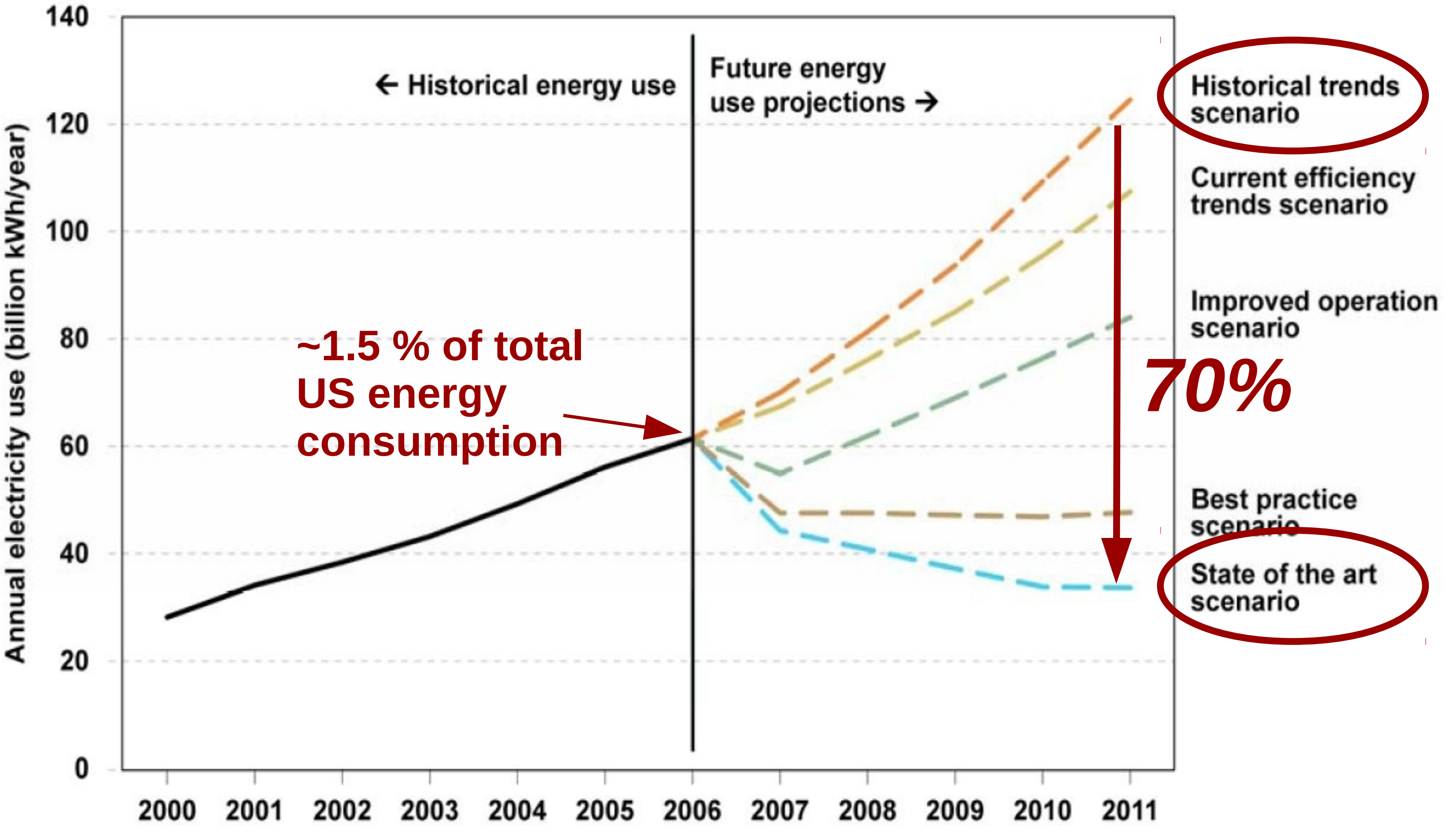
Data center electricity consumption in US



~1.5 % of total US energy consumption

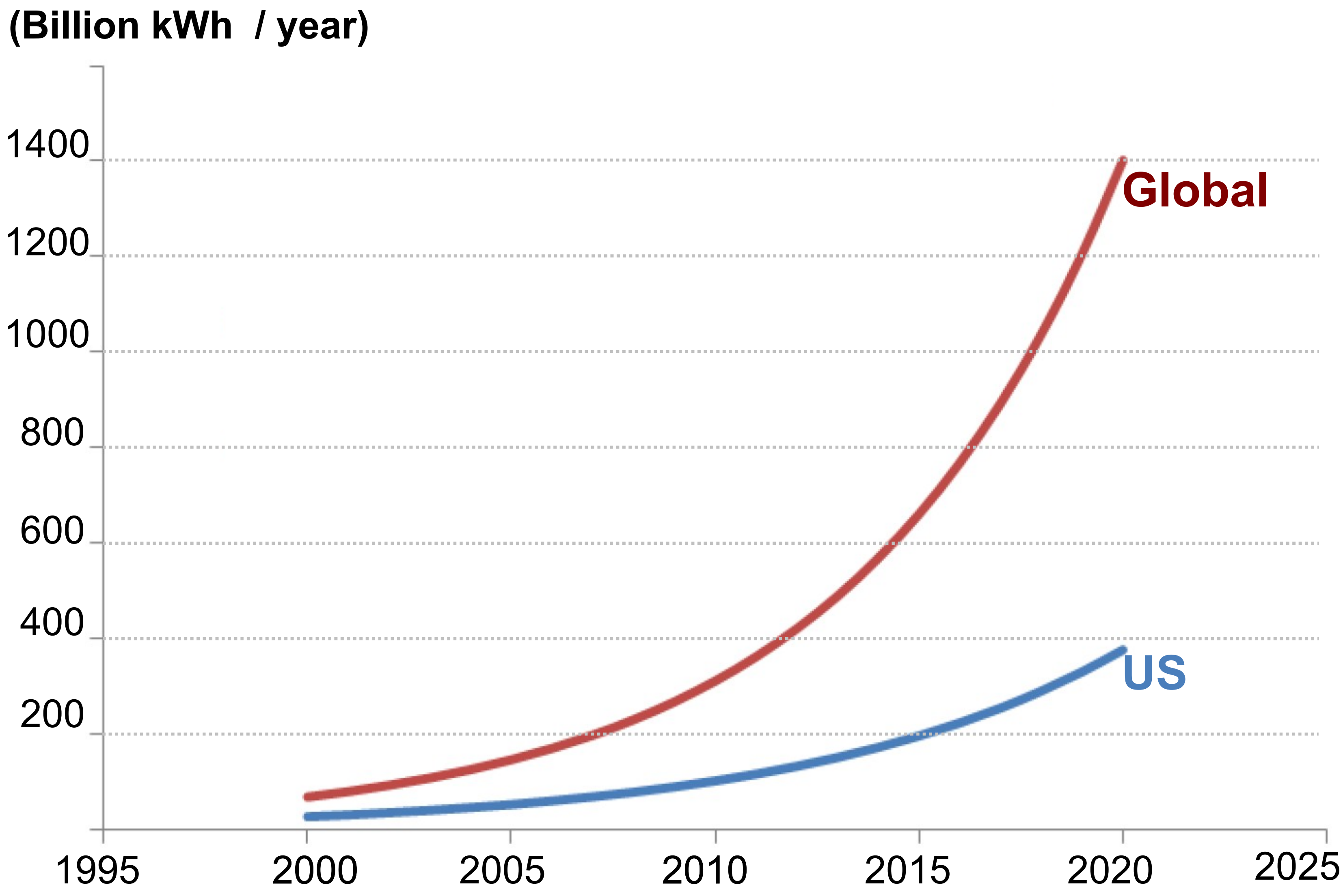
Environmental Protection Agency, Technical report, 2007

Data center electricity consumption in US



Environmental Protection Agency, Technical report, 2007

Data center electricity consumption

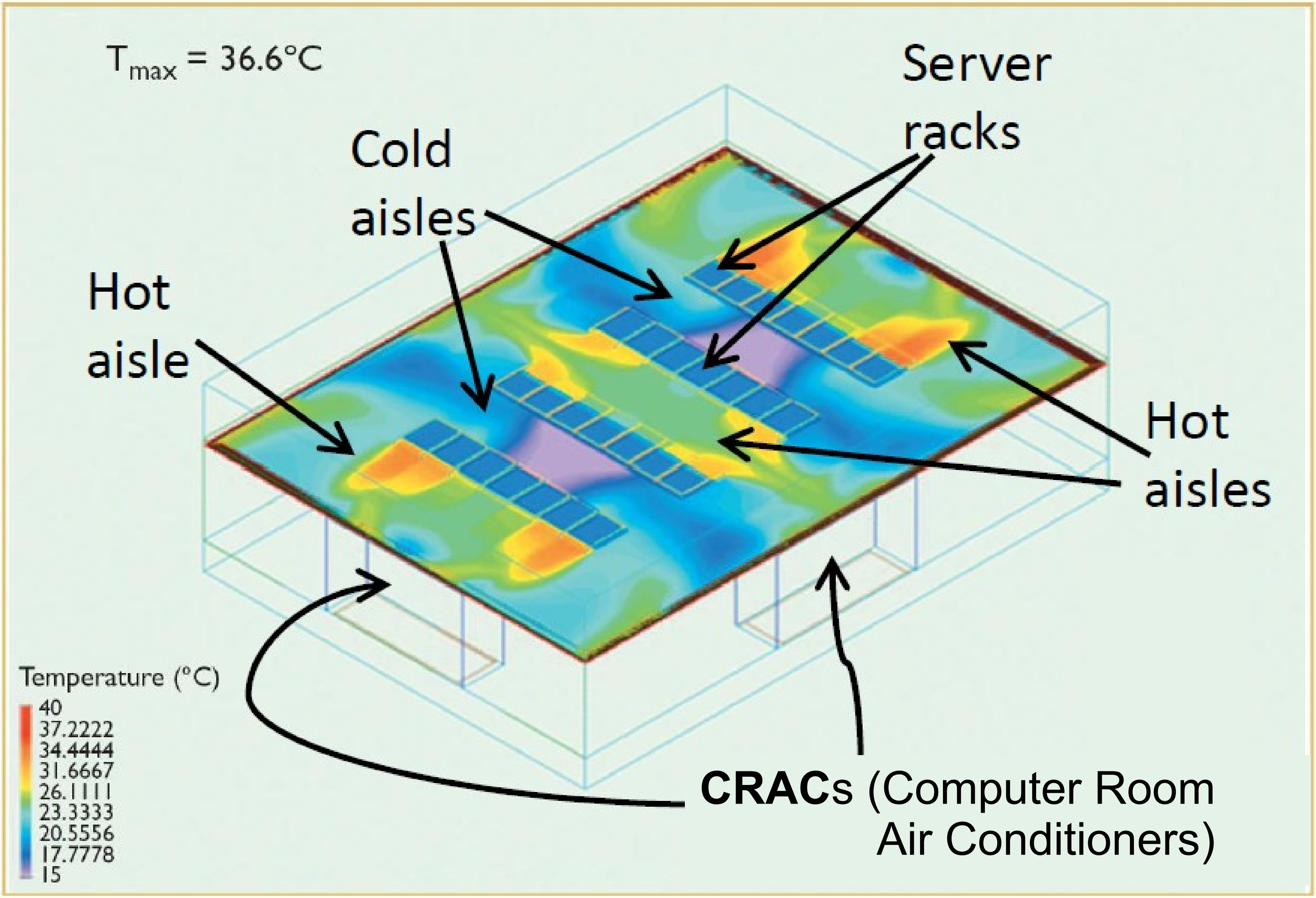


C. L. Belady, Projecting annual new datacenter construction market size, Microsoft, 2011

Thermal constraints

- **Without proper cooling, chip temperatures would exceed safe operation limits**
  - *Almost complete transformation of electrical energy into heat*
- **Chip temperature generally unobservable**
- **Industrial approach**
  - Bound server *inlet air temperature*

Temperature distribution



CRACs (Computer Room Air Conditioners)

R. K. Sharma *et al.* "Balance of Power: Dynamic Thermal Management for Internet Data Centers", IEEE Internet Computing 2005

Contributions of the dissertation

- **A modeling framework for data centers**
  - Facilitates the development of cyber-physical models of data centers
- **A control strategy**
  - Takes advantage of the modularity typically found in data centers
- **Multiple control-oriented models**
  - Make it possible to formulate and to solve, in real-time, optimal control problems
- **Multiple control algorithms**
  - Representative abstractions of different approaches to the data center control
- **A collection of MATLAB routines to simulate the evolution of a data center (Data center simulator)**
  - Makes it possible to analyze the performance of multiple controllers in a variety of scenarios

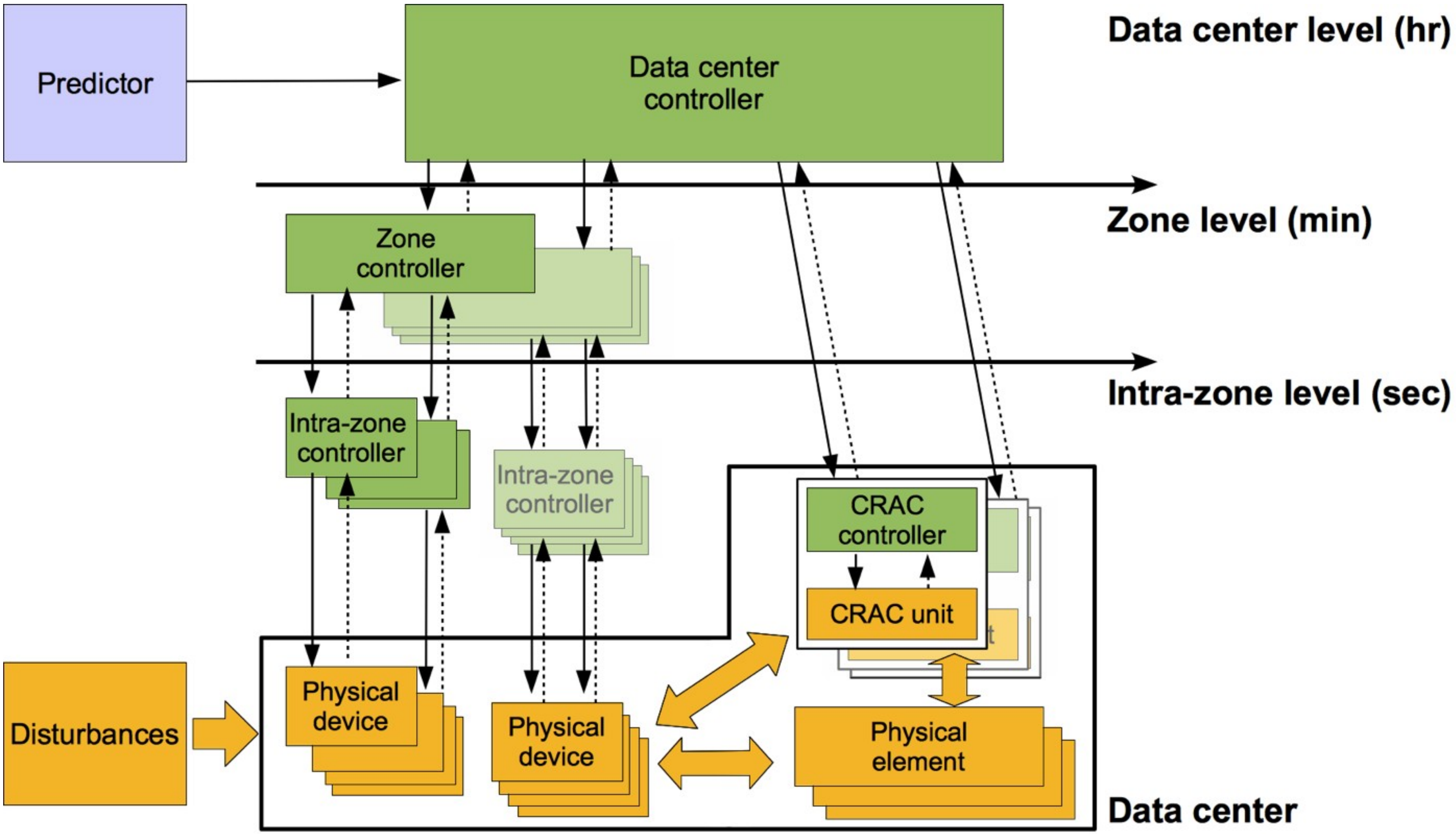
Contributions of the dissertation

- **A modeling framework for data centers**
  - Facilitates the development of cyber-physical models of data centers
- ***A control strategy***
  - Takes advantage of the modularity typically found in data centers
- ***Multiple control-oriented models***
  - Make it possible to formulate and to solve, in real-time, optimal control problems
- ***Multiple control algorithms***
  - Representative abstractions of different approaches to the data center control
- **A collection of MATLAB routines to simulate the evolution of a data center (Data center simulator)**
  - Makes it possible to analyze the performance of multiple controllers in a variety of scenarios

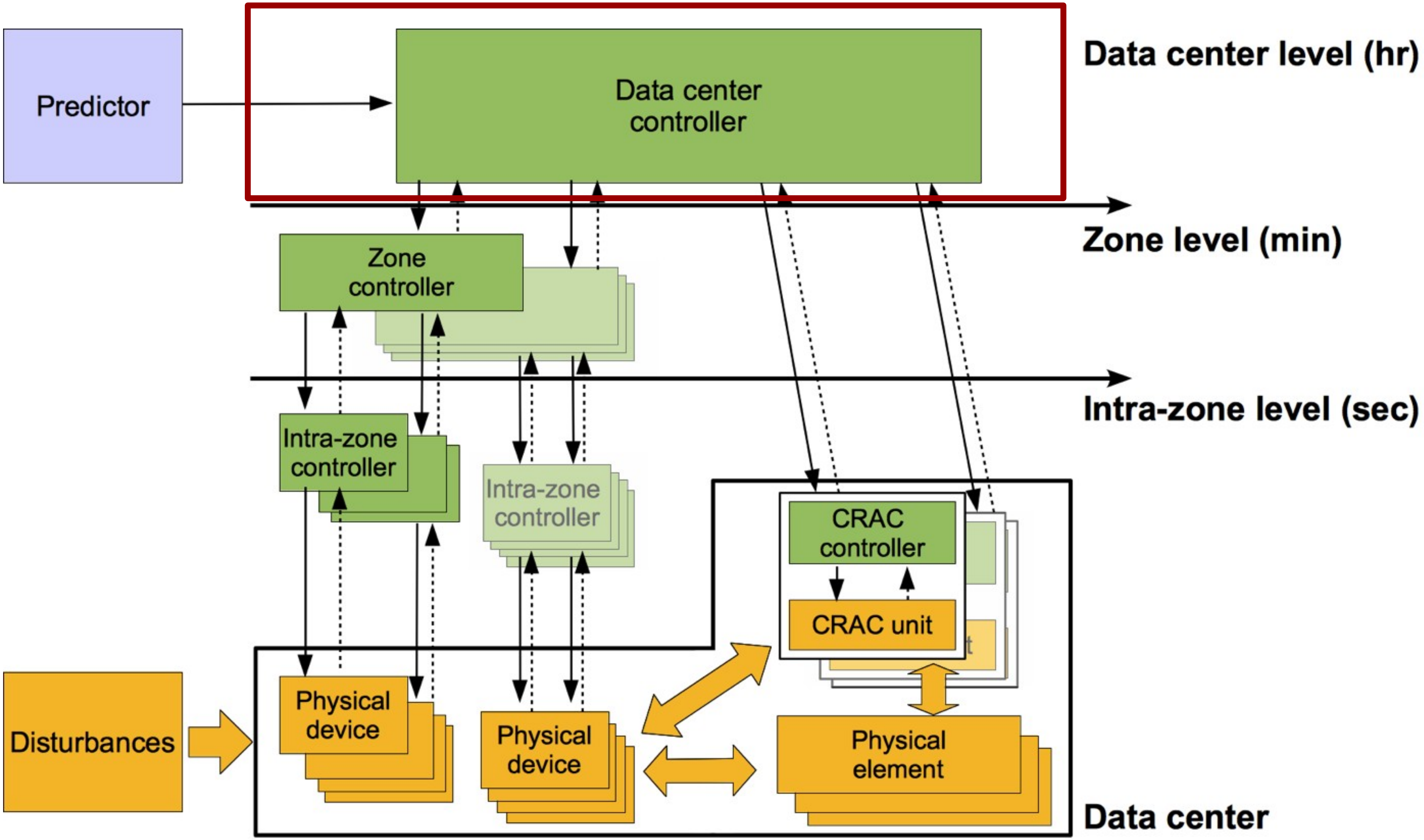
Outline

- Introduction
- **Proposed control strategy**
- Performance analysis for constant job arrival rate
- Interaction with the smart-grid
- Zone-level control
- Conclusion and future work

Hierarchical control strategy



Hierarchical control strategy



Data-center-level controller

- **Focuses on processes in the hours time scale**
  - Considers only the mean values of the predicted processes
- **Groups servers into zones**
  - *Power consumption of a zone is proportional to the amount of jobs executed*

Data-center-level controller

- **Focuses on processes in the hours time scale**
  - Considers only the mean values of the predicted processes
- **Groups servers into zones**
  - *Power consumption of a zone is proportional to the amount of jobs executed*
- **Defines set-points for**
  - Scheduling of jobs among the zones
  - Migration of jobs among the zones
  - Rate at which jobs have to be executed in every zone
  - Reference temperature of CRAC units
- **Considers**
  - Computational and thermal dynamics
  - Nonlinear efficiency of the CRAC units, i.e. coefficient of performance (COP)

Proposed data-center-level controllers

- **Baseline controller**
  - Open-loop controller
  - Set-points set for the worst-case scenario
  - Lower bound of the performance of other control approaches

Proposed data-center-level controllers

- **Baseline controller**
  - Open-loop controller
  - Set-points set for the worst-case scenario
  - Lower bound of the performance of other control approaches
- **Uncoordinated controller**
  - Based on model-predictive-control (MPC)
  - Neglects the coupling between the cyber and the physical parts
  - Represents the control approaches typically found in modern data centers

Proposed data-center-level controllers

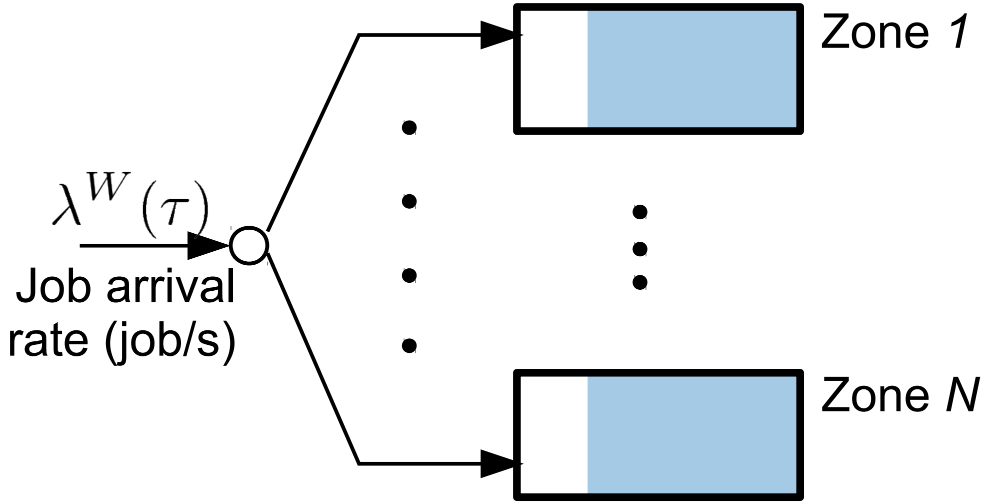
- **Baseline controller**
  - Open-loop controller
  - Set-points set for the worst-case scenario
  - Lower bound of the performance of other control approaches
- **Uncoordinated controller**
  - Based on model-predictive-control (MPC)
  - Neglects the coupling between the cyber and the physical parts
  - Represents the control approaches typically found in modern data centers
- **Coordinated controller**
  - Based on MPC
  - Considers the coupling between the cyber and the physical parts
  - Takes full advantage of a cyber-physical model of a data center

Proposed data-center-level controllers

- **Baseline controller**
  - Open-loop controller
  - Set-points set for the worst-case scenario
  - Lower bound of the performance of other control approaches
- **Uncoordinated controller**
  - Based on model-predictive-control (MPC)
  - Neglects the coupling between the cyber and the physical parts
  - Represents the control approaches typically found in modern data centers
- **Coordinated controller**
  - Based on MPC
  - Considers the coupling between the cyber and the physical parts
  - Takes full advantage of a cyber-physical model of a data center
- ***The different controllers help us to understand the impact of considering the cyber-physical nature of data centers***

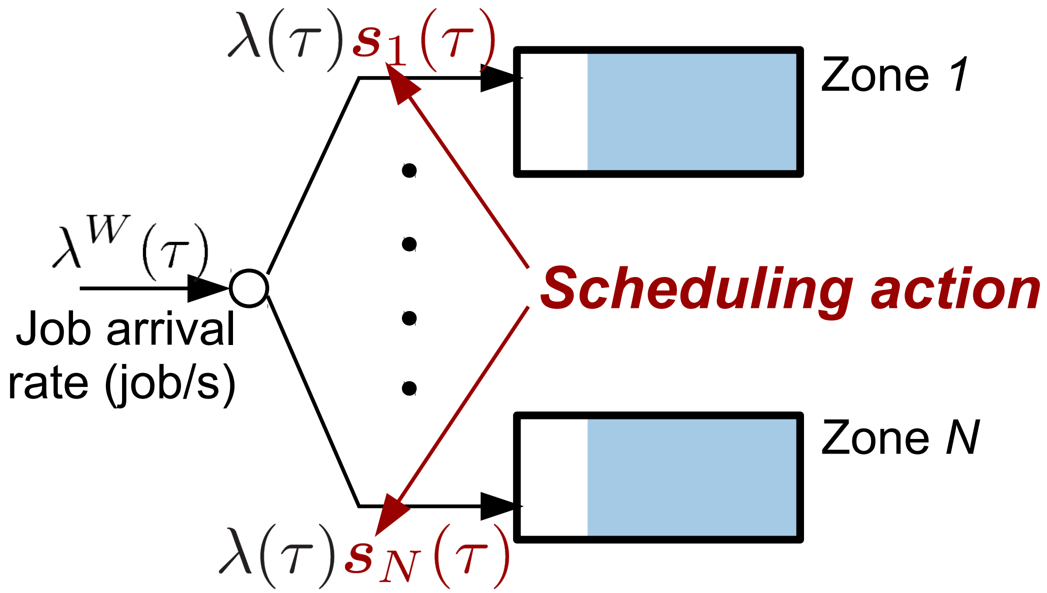
Computational network

- **Describes the process of data exchange within the data center and between the data center the external users**
- **Based on a fluid approximation of the job execution and arrival processes**
  - First-order approximation of a queuing system



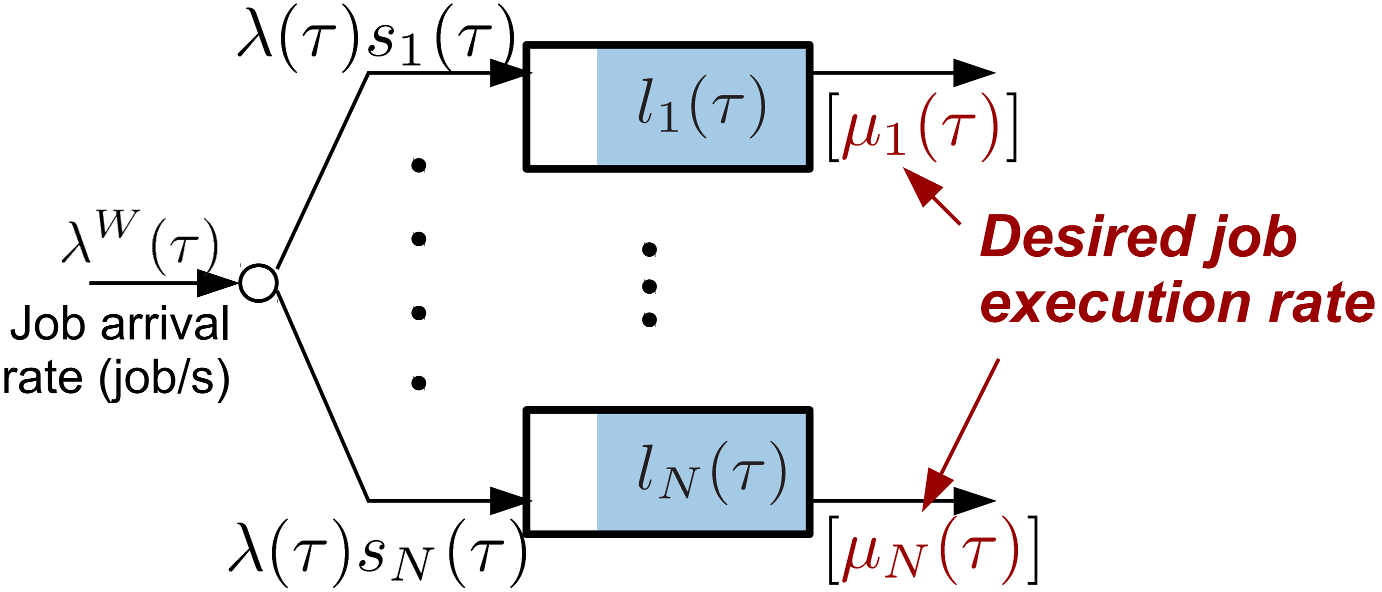
Computational network

- **Describes the process of data exchange within the data center and between the data center the external users**
- **Based on a fluid approximation of the job execution and arrival processes**
  - First-order approximation of a queuing system



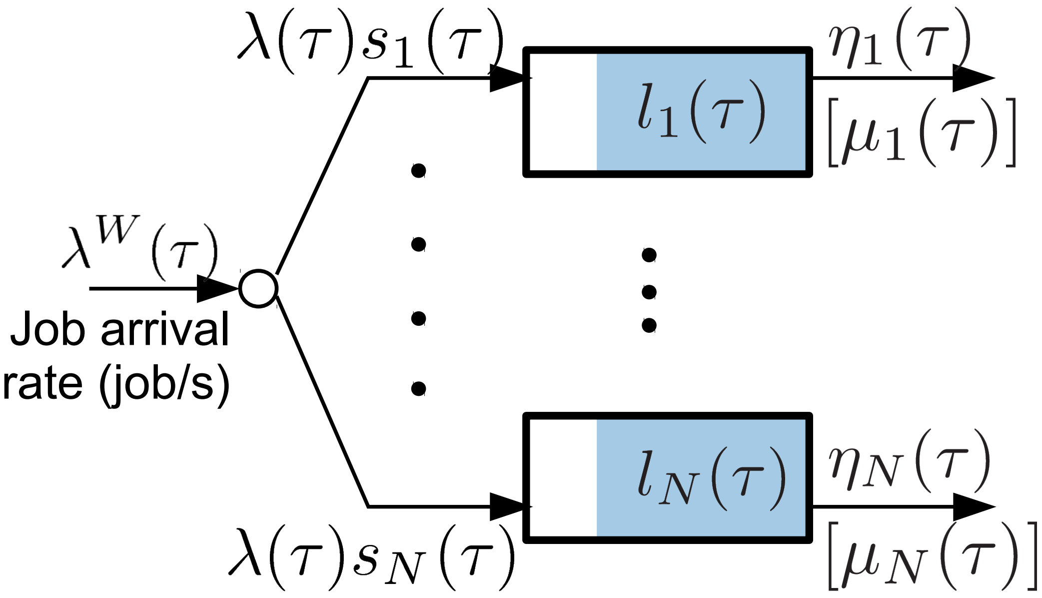
Computational network

- **Describes the process of data exchange within the data center and between the data center the external users**
- **Based on a fluid approximation of the job execution and arrival processes**
  - First-order approximation of a queuing system



Computational network

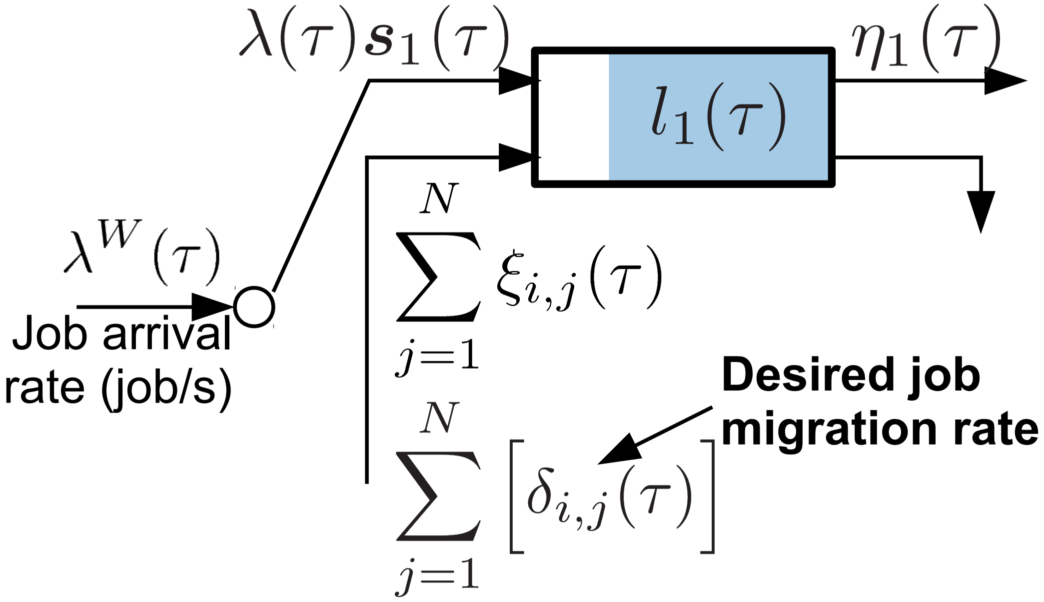
- **Describes the process of data exchange within the data center and between the data center the external users**
- **Based on a fluid approximation of the job execution and arrival processes**
    - First-order approximation of a queuing system



$$a_1(\tau) = \lambda(\tau)s_1(\tau)$$
$$d_1(\tau) = \eta_1(\tau)$$
$$\dot{l}_1(\tau) = a_1(\tau) - d_1(\tau)$$
$$\eta_1(\tau) = \begin{cases} \mu_1(\tau) & \text{if } l_1(\tau) > 0 \\ & \text{or } a_1(\tau) > \mu_1(\tau) \\ a_1(\tau) & \text{otherwise} \end{cases}$$

Computational network

- **Describes the process of data exchange within the data center and between the data center the external users**
- **Based on a fluid approximation of the job execution and arrival processes**
  - First-order approximation of a queuing system



Job arrival rate (job/s)

Desired job migration rate

$$a_1(\tau) = \lambda(\tau)\boldsymbol{s}_1(\tau) + \sum_{j=1}^{N} \xi_{1,j}(\tau)$$

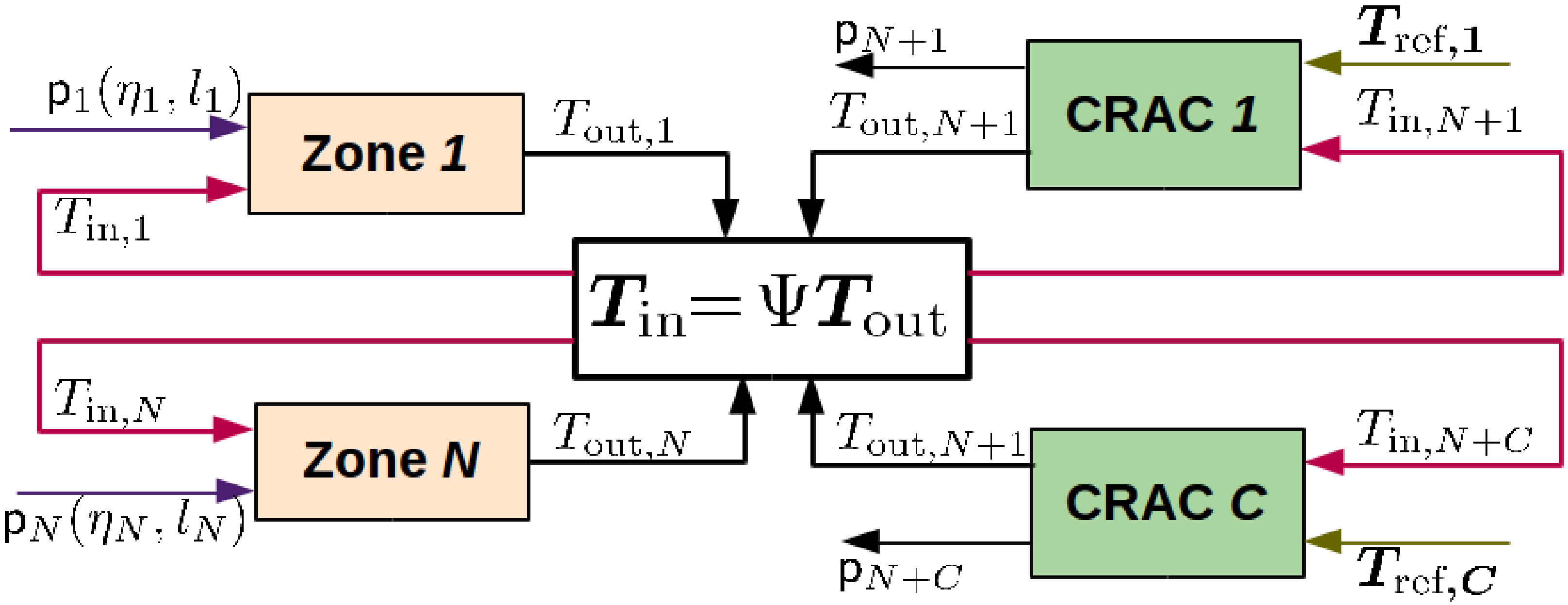
$$d_1(\tau) = \eta_1(\tau) + \sum_{j=1}^{N} \xi_{j,1}(\tau)$$

$$\dot{l}_1(\tau) = a_1(\tau) - d_1(\tau)$$

$$\eta_1(\tau) = \begin{cases} \mu_1(\tau) & \text{if } l_1(\tau) > 0 \\ & \text{or } a_1(\tau) > \mu_1(\tau) \\ a_1(\tau) & \text{otherwise} \end{cases}$$

Thermal network

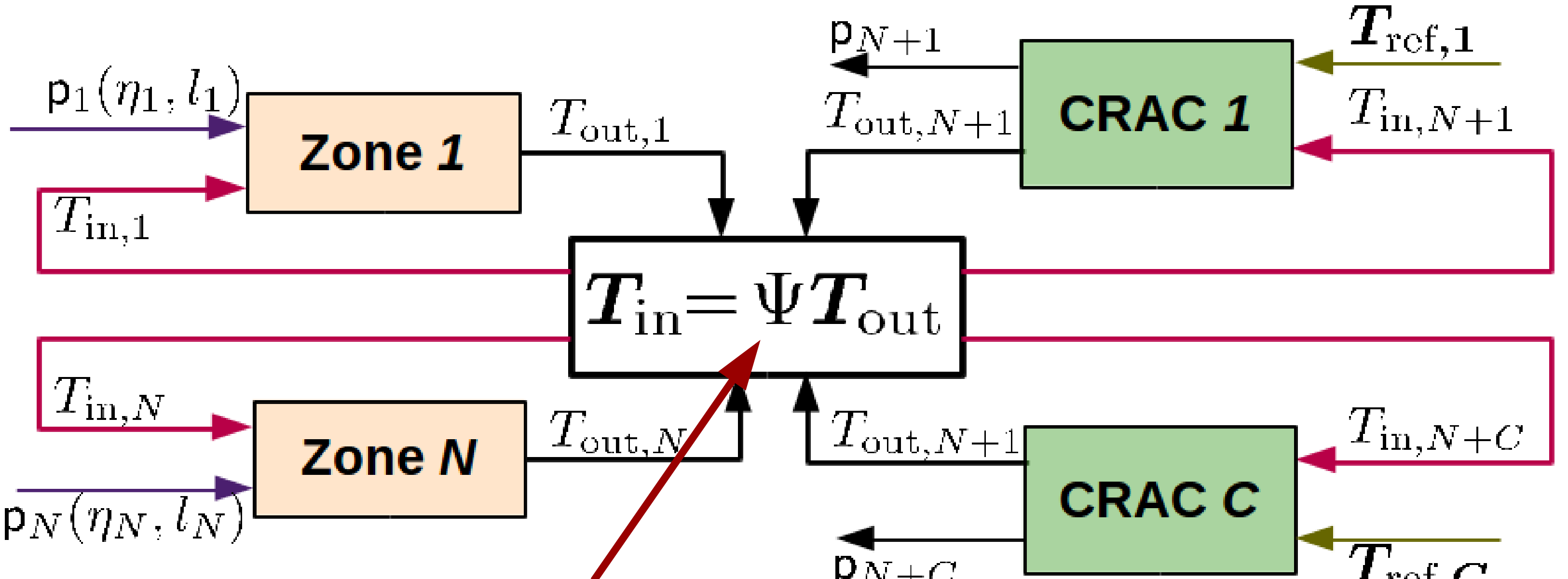
- **Describes the process of energy exchange within the data center and between the data center the external environment**



Q. Tang *et al.* "Sensor-based fast thermal evaluation model for energy efficient high-performance data centers", 2006

Thermal network

- **Describes the process of energy exchange within the data center and between the data center the external environment**



Constant airflow vector and exchange matrix

- **Inlet temperature constraint** $\boldsymbol{T}_{\mathrm{in}}(\tau) = \Psi \boldsymbol{T}_{\mathrm{out}}(\tau) \leq \overline{\boldsymbol{T}_{\mathrm{in}}}$
- **CRAC power consumption** $\mathsf{p}_i(t) = \dfrac{\dot{Q}_i(t)}{COP_i(T_{\mathrm{out},i}(t))}$

Heat removed rate (W)

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

Variables at step *k*

- **No environment nodes considered in this table**

			Variables
Input	Controllable	Job scheduling	$\mathbf{s}(k)$
		Desired job execution rate	$\boldsymbol{\mu}(k)$
		Desired job migration rate	$\boldsymbol{\delta}(k)$
		CRAC unit reference temperature	$\mathbf{T}_{\mathrm{ref}}(k)$
	Uncontrollable	Job arrival rate	$\lambda^{W}(k)$
Output	Job departure rates from every zone		$\boldsymbol{\eta}(k)$
	Power consumption of CRAC nodes		$\mathbf{p}_C(k)$
	Input temperatures of zones and CRAC nodes		$\mathbf{T}_{\mathbf{in}}(k)$
	Zone power consumption		$\mathbf{p}_N(k)$
State	Number of jobs in every zones		$\mathbf{l}(k)$
	Output temperatures of zones and CRAC nodes		$\mathbf{T}_{\mathrm{out}}(k)$

Baseline & uncoordinated controllers

- **Baseline controller**

$$\boldsymbol{\mu}(k) = \overline{\boldsymbol{\mu}} \qquad \boldsymbol{\delta}(k) = \mathbf{0} \qquad \boldsymbol{s}(k) = \mathbf{1}\frac{1}{N} \qquad \boldsymbol{T}_{\mathrm{ref}}(k) = \underline{\boldsymbol{T}_{\mathrm{ref}}}$$

Baseline & uncoordinated controllers

- **Baseline controller**

$$\boldsymbol{\mu}(k) = \overline{\boldsymbol{\mu}} \qquad \boldsymbol{\delta}(k) = \mathbf{0} \qquad \boldsymbol{s}(k) = \mathbf{1}\frac{1}{N} \qquad \boldsymbol{T}_{\mathrm{ref}}(k) = \underline{\boldsymbol{T}_{\mathrm{ref}}}$$

- **Uncoordinated controller**

$$\min_{\mathcal{M},\mathcal{S},\mathcal{D}} \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{N}}(h|k)$$

Minimize expected zone power consumption

$$\begin{aligned}
\textbf{s.t.} \quad & \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k) \\
& \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\
& \text{computational dynamics} \\
& \text{QoS constraints} \\
& \left.\begin{array}{ll} \mathbf{0} \leq \hat{\boldsymbol{\mu}}(h|k) \leq \overline{\boldsymbol{\mu}}, & \hat{\boldsymbol{\delta}}(h|k) = \mathbf{0} \\ \mathbf{0} \leq \hat{\boldsymbol{s}}(h|k) \leq \mathbf{1}, & \mathrm{B}_s \hat{\boldsymbol{s}}(h|k) = \mathbf{1} \end{array}\right\}
\end{aligned}$$

- Computational dynamics
- Quality of service (QoS) constraints
- Control constraints

$$\mathcal{D} = \left\{\hat{\boldsymbol{\delta}}(k|k), \ldots, \hat{\boldsymbol{\delta}}(k+\mathcal{T}-1|k)\right\} \qquad \mathcal{T}_{ref} = \left\{\hat{\boldsymbol{T}}_{\mathrm{ref}}(k|k), \ldots, \hat{\boldsymbol{T}}_{\mathrm{ref}}(k+\mathcal{T}-1|k)\right\}$$

$$\mathcal{M} = \left\{\hat{\boldsymbol{\mu}}(k|k), \ldots, \hat{\boldsymbol{\mu}}(k+\mathcal{T}-1|k)\right\} \qquad \mathcal{S} = \left\{\hat{\boldsymbol{s}}(k|k), \ldots, \hat{\boldsymbol{s}}(k+\mathcal{T}-1|k)\right\}$$

Baseline & uncoordinated controllers

- **Baseline controller**

$$\boldsymbol{\mu}(k) = \overline{\boldsymbol{\mu}} \qquad \boldsymbol{\delta}(k) = \mathbf{0} \qquad \boldsymbol{s}(k) = \mathbf{1}\frac{1}{N} \qquad \boldsymbol{T}_{\text{ref}}(k) = \underline{\boldsymbol{T}_{\text{ref}}}$$

- **Uncoordinated controller**

$$\begin{aligned} \min_{\mathcal{M},\mathcal{S},\mathcal{D}} \ & \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{N}}(h|k) \\ \textbf{s.t.} \quad & \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k) \\ & \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\ & \text{computational dynamics} \\ & \text{QoS constraints} \\ & \mathbf{0} \leq \hat{\boldsymbol{\mu}}(h|k) \leq \overline{\boldsymbol{\mu}}, \quad \hat{\boldsymbol{\delta}}(h|k) = \mathbf{0} \\ & \mathbf{0} \leq \hat{\boldsymbol{s}}(h|k) \leq \mathbf{1}, \quad B_s \hat{\boldsymbol{s}}(h|k) = \mathbf{1} \end{aligned}$$

$$\xrightarrow{\{\hat{\mathbf{p}}_{\mathcal{N}}(h|k)\}}$$

$$\begin{aligned} \min_{\mathcal{T}_{ref}} \ & \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{C}}(h|k) \\ \textbf{s.t.} \quad & \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k) \\ & \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\ & \text{thermal dynamics} \\ & \underline{\boldsymbol{T}_{\text{ref}}} \leq \hat{\boldsymbol{T}}_{\text{ref}}(h|k) \leq \overline{\boldsymbol{T}_{\text{ref}}} \\ & \hat{\boldsymbol{T}}_{\text{in}}(h+1|k) \leq \overline{\boldsymbol{T}_{\text{in}}} \end{aligned}$$

$$\mathcal{D} = \left\{\hat{\boldsymbol{\delta}}(k|k), \ldots, \hat{\boldsymbol{\delta}}(k+\mathcal{T}-1|k)\right\} \qquad \mathcal{T}_{ref} = \left\{\hat{\boldsymbol{T}}_{\text{ref}}(k|k), \ldots, \hat{\boldsymbol{T}}_{\text{ref}}(k+\mathcal{T}-1|k)\right\}$$

$$\mathcal{M} = \left\{\hat{\boldsymbol{\mu}}(k|k), \ldots, \hat{\boldsymbol{\mu}}(k+\mathcal{T}-1|k)\right\} \qquad \mathcal{S} = \left\{\hat{\boldsymbol{s}}(k|k), \ldots, \hat{\boldsymbol{s}}(k+\mathcal{T}-1|k)\right\}$$

Coordinated controller

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

$$\min_{\mathcal{M},\mathcal{S},\mathcal{D},\mathcal{T}_{ref}} \left(\sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{N}}(h|k) + \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{C}}(h|k) \right)$$

Minimize expected zone and CRAC power consumption

$$\begin{aligned}
\textbf{s.t.}\quad & \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k), \quad \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k) \\
& \text{for all } h = k, \ldots, k + \mathcal{T} - 1 \\
& \left.\begin{array}{l}\text{computational dynamics,} \\ \text{thermal dynamics,}\end{array}\right\} \\
& \text{QoS constraints,} \\
& \mathbf{0} \leq \hat{\boldsymbol{\mu}}(h|k) \leq \overline{\boldsymbol{\mu}}, \quad \hat{\boldsymbol{\delta}}(h|k) = \mathbf{0} \\
& \mathbf{0} \leq \hat{\boldsymbol{s}}(h|k) \leq \mathbf{1}, B_s \hat{\boldsymbol{s}}(h|k) = \mathbf{1}, \\
& \underline{\boldsymbol{T}_{\text{ref}}} \leq \hat{\boldsymbol{T}}_{\text{ref}}(h|k) \leq \overline{\boldsymbol{T}_{\text{ref}}}, \quad \hat{\boldsymbol{T}}_{\text{in}}(h+1|k) \leq \overline{\boldsymbol{T}_{\text{in}}}, \\
& \hat{\mathbf{p}}(h|k) = \mathsf{A}_{\alpha} \hat{\boldsymbol{\eta}}(h|k) + \mathsf{B}_{\beta} \hat{\boldsymbol{l}}(h|k)
\end{aligned}$$

Considers both thermal and computational dynamics

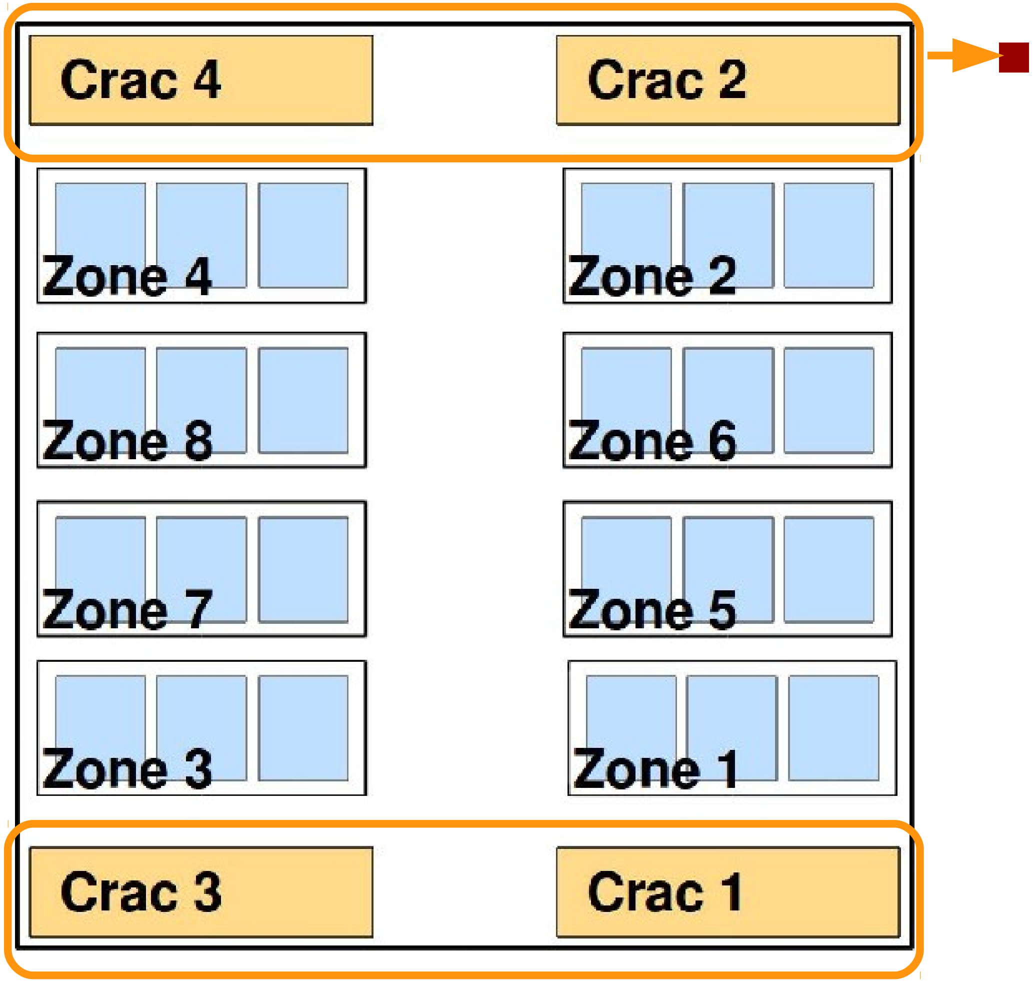
Thermal-computational coupling

Outline

- Introduction
- Proposed control strategy
- **Performance analysis for constant job arrival rate**
- Interaction with the smart-grid
- Zone-level control
- Conclusion and future work

Data center layout

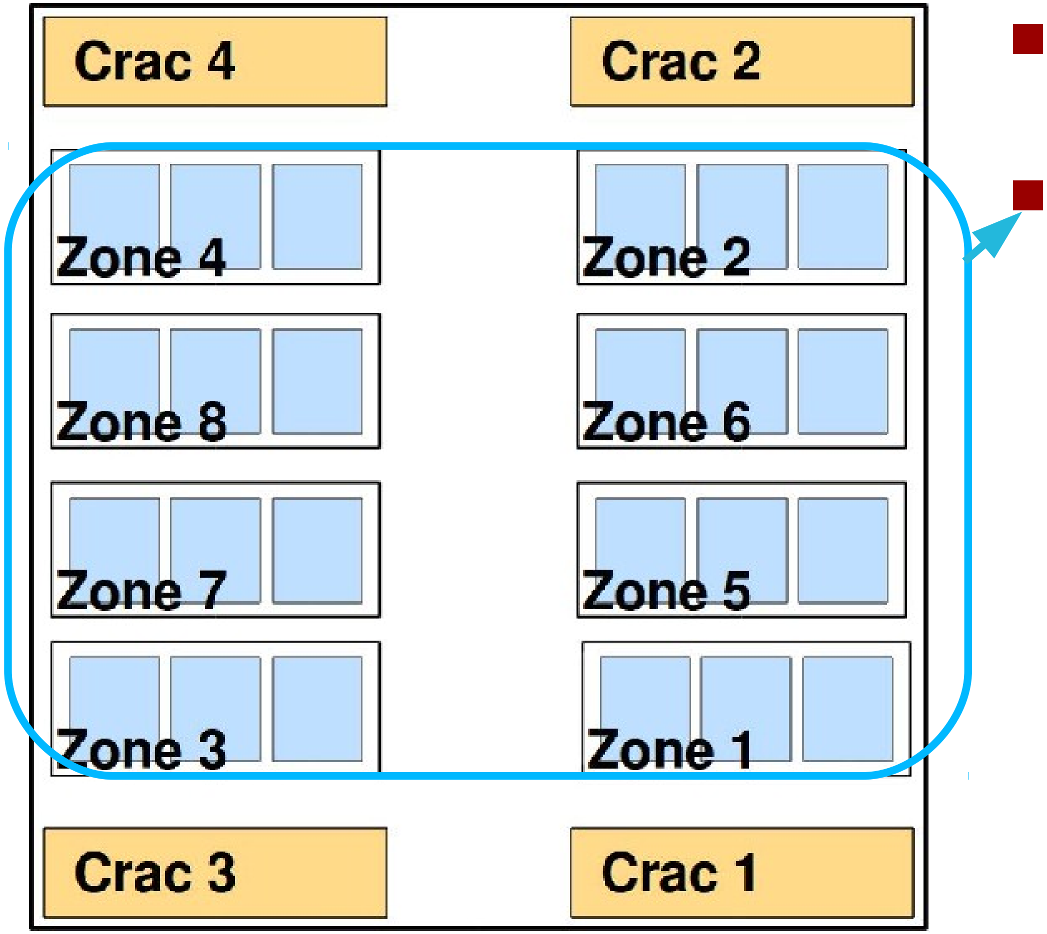
- **Highlights the effects of choosing energy efficient servers or efficient cooled servers**



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

Data center layout

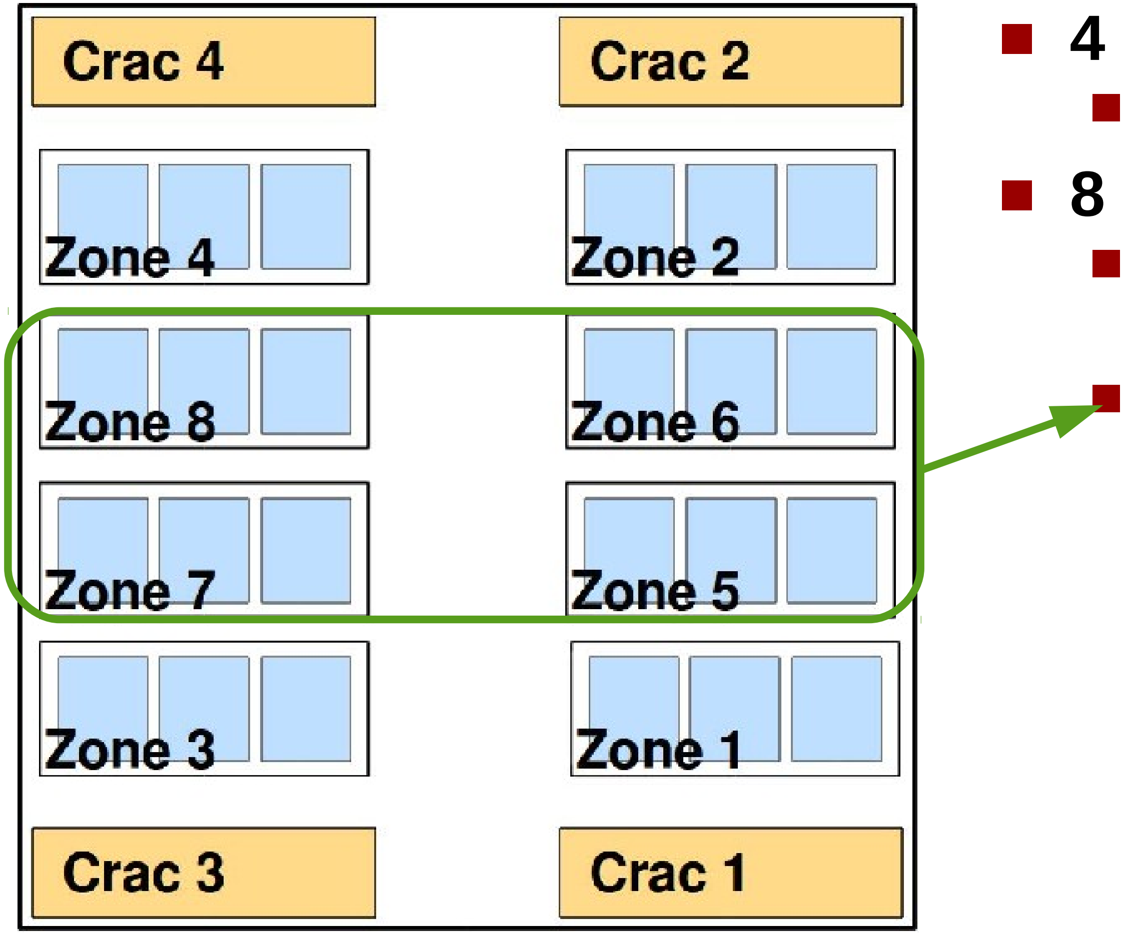
- **Highlights the effects of choosing energy efficient servers or efficient cooled servers**



- **4 CRAC units**
  - Identical to each other
- **8 Zones**
  - 3 Racks each (126 servers per zone)

Data center layout

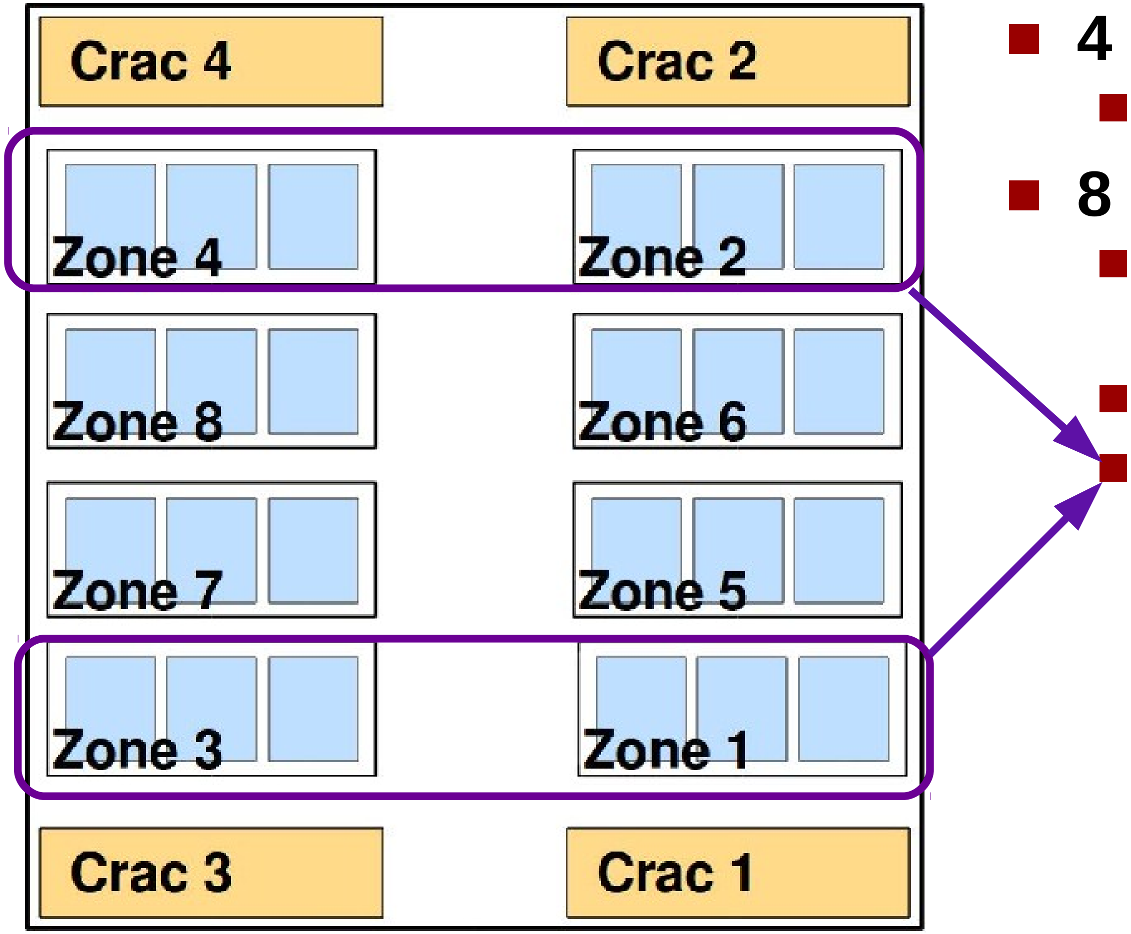
- **Highlights the effects of choosing energy efficient servers or efficient cooled servers**



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

Data center layout

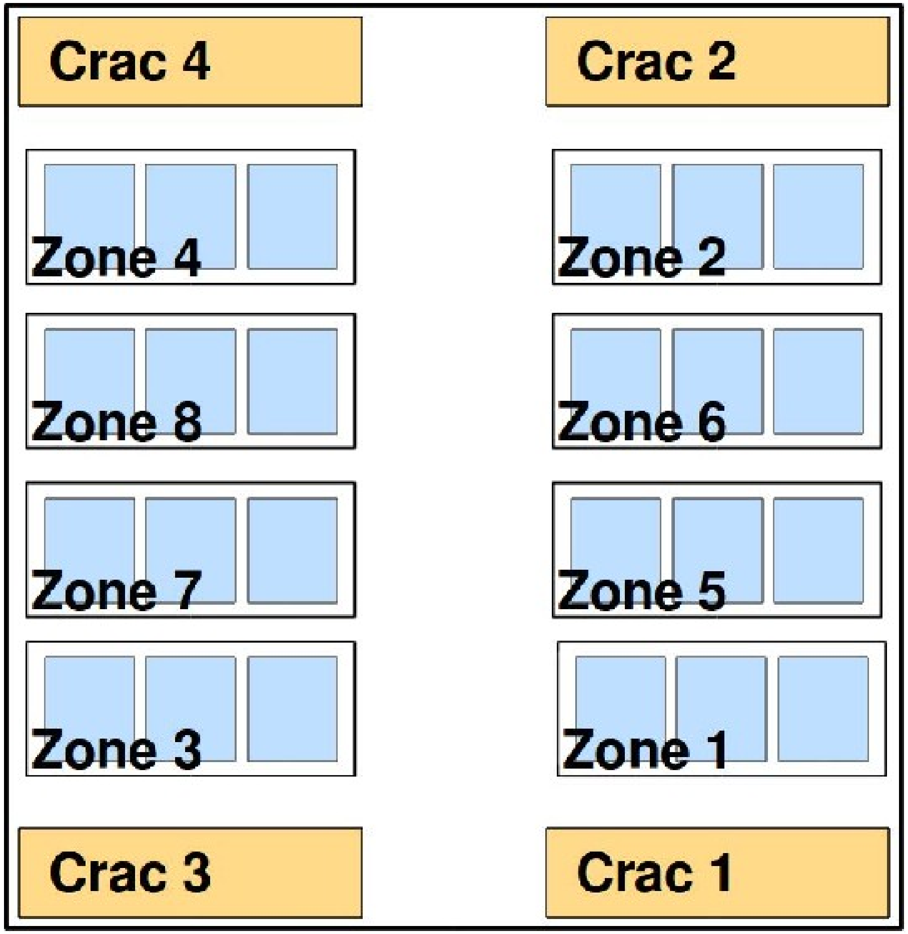
- **Highlights the effects of choosing energy efficient servers or efficient cooled servers**



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

Data center layout

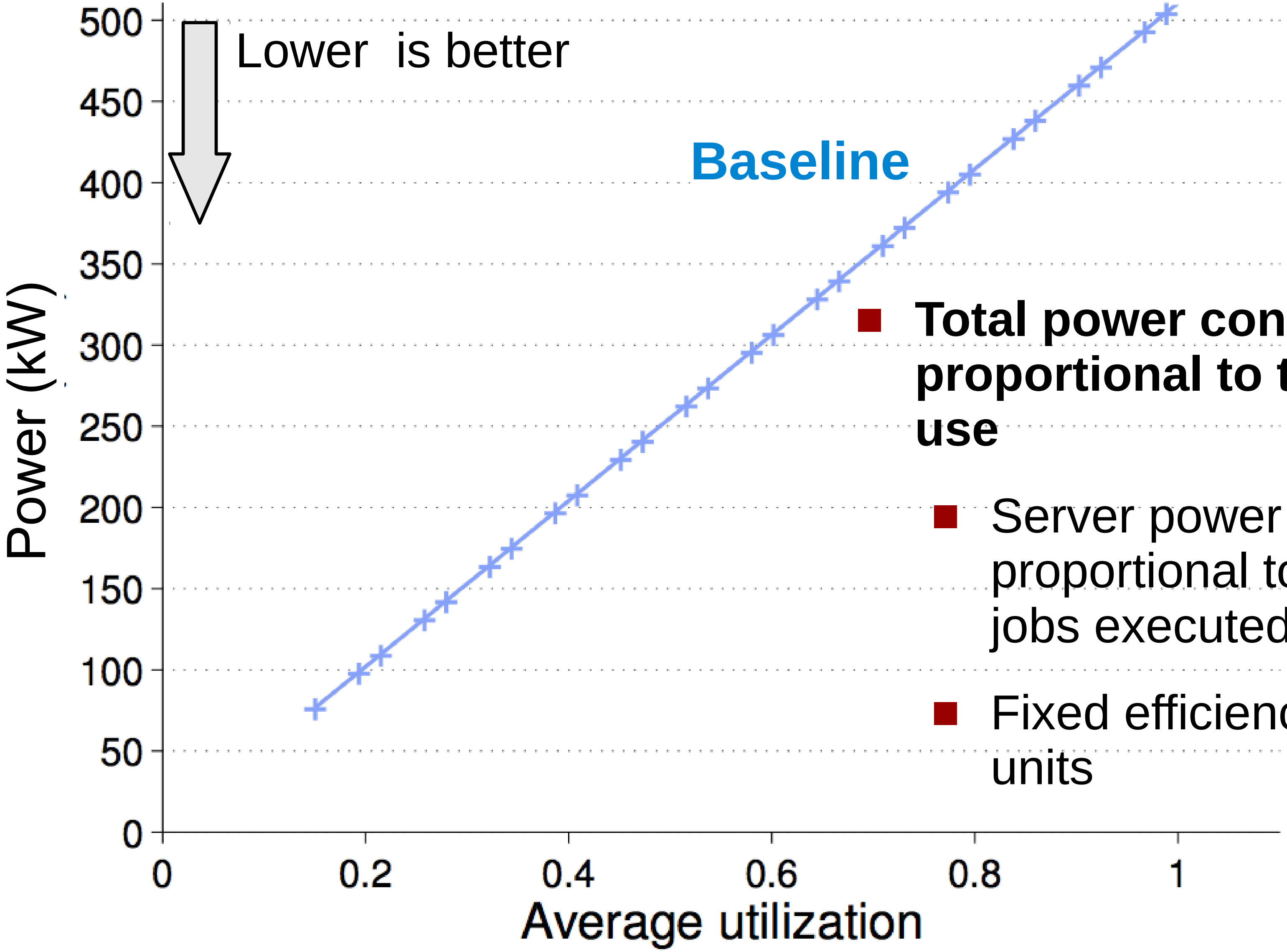
- **Highlights the effects of choosing energy efficient servers or efficient cooled servers**



- **4 CRAC units**
  - Identical to each other
- **8 Zones**
  - 3 Racks each (126 servers per zone)
  - Energy efficient servers
  - Efficiently cooled servers
- **Optimal control problems**
  - Written in TomSym
  - Solver: KNITRO 7.0

http://tomsym.com/ http://www.ziena.com/knitro.htm

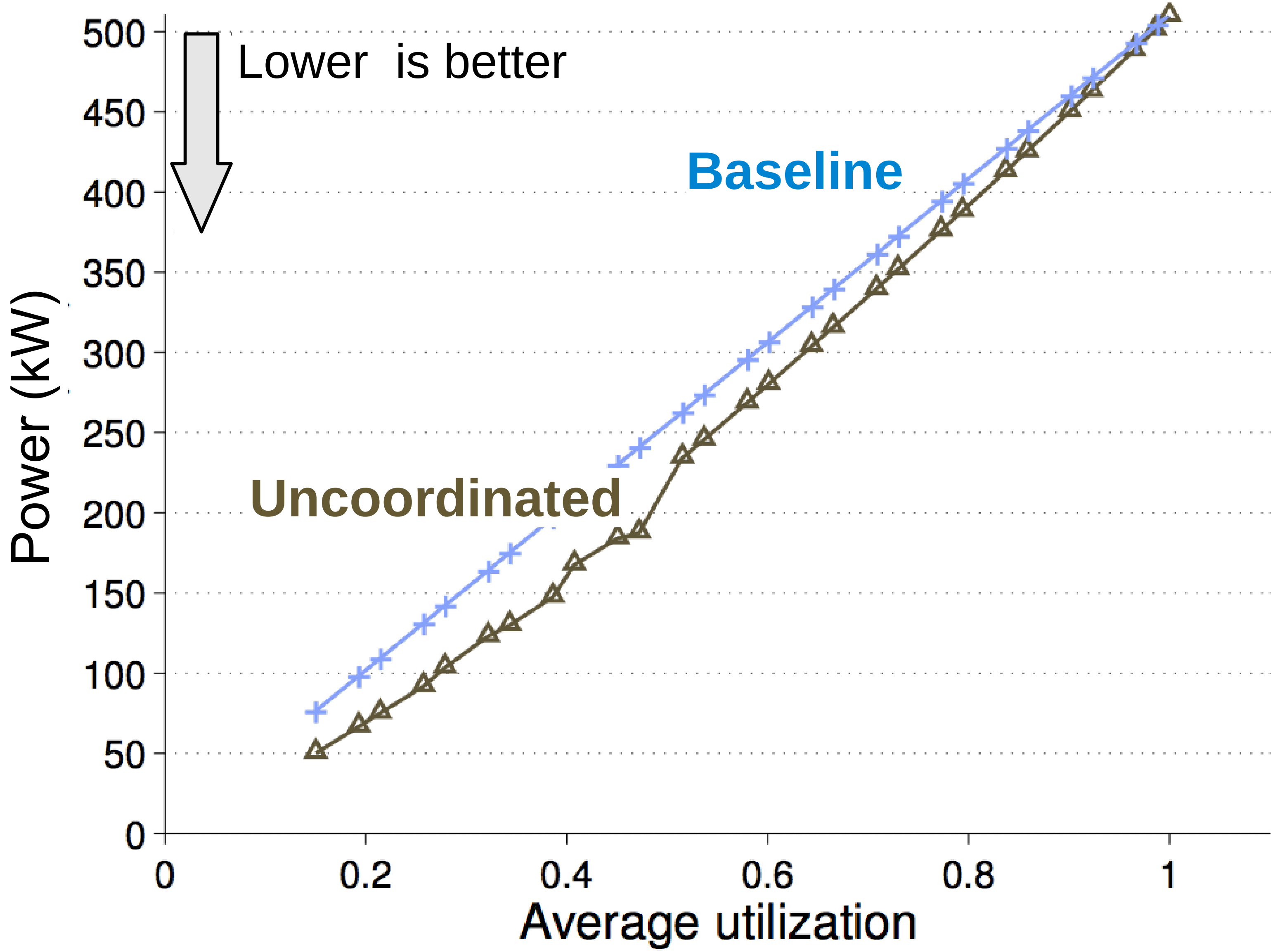
Total power consumption



- **Total power consumption proportional to the average use**
  - Server power consumption proportional to the amount of jobs executed
  - Fixed efficiency of CRAC units

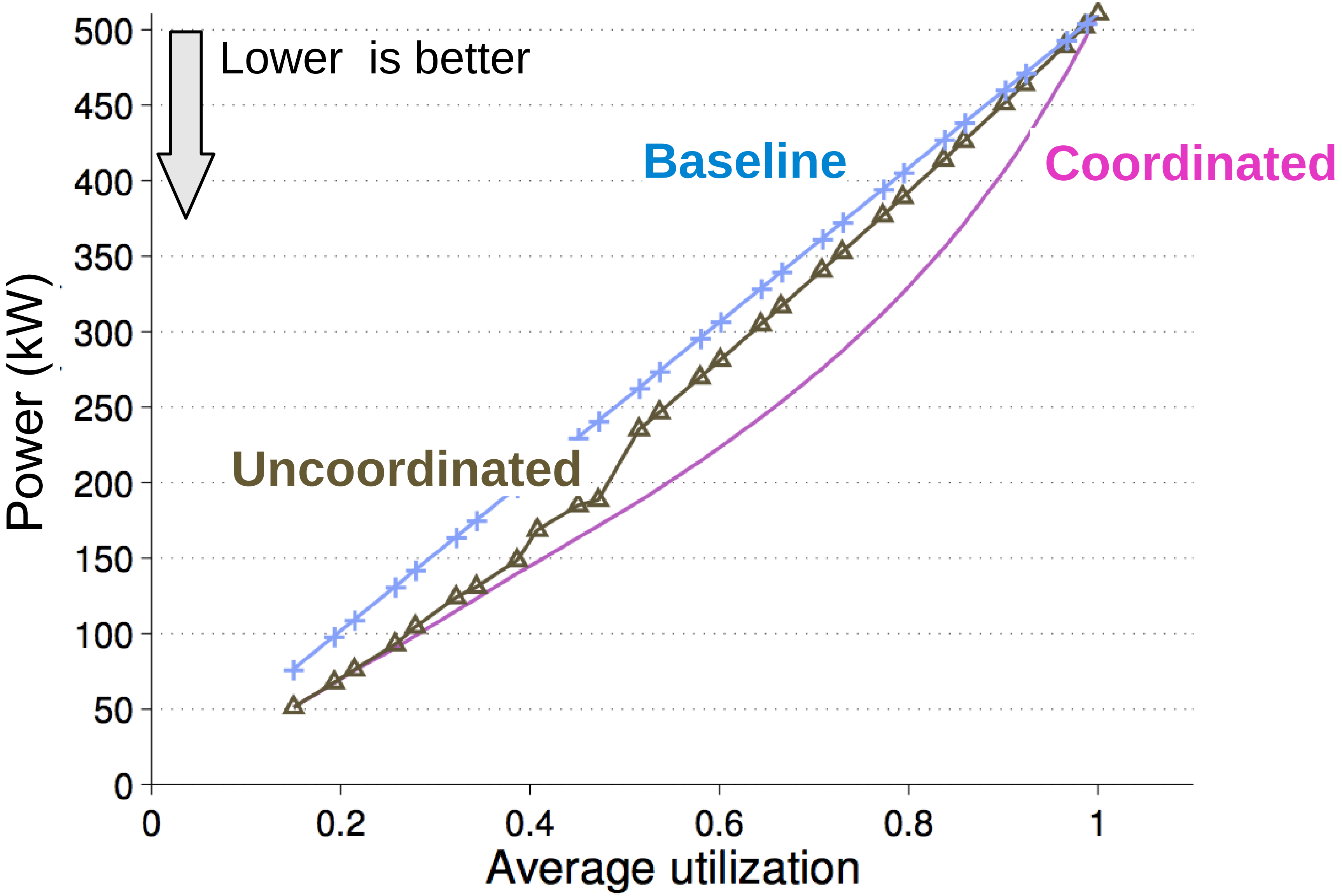
Parolini *et al.* Proceedings of the IEEE 2011

Total power consumption



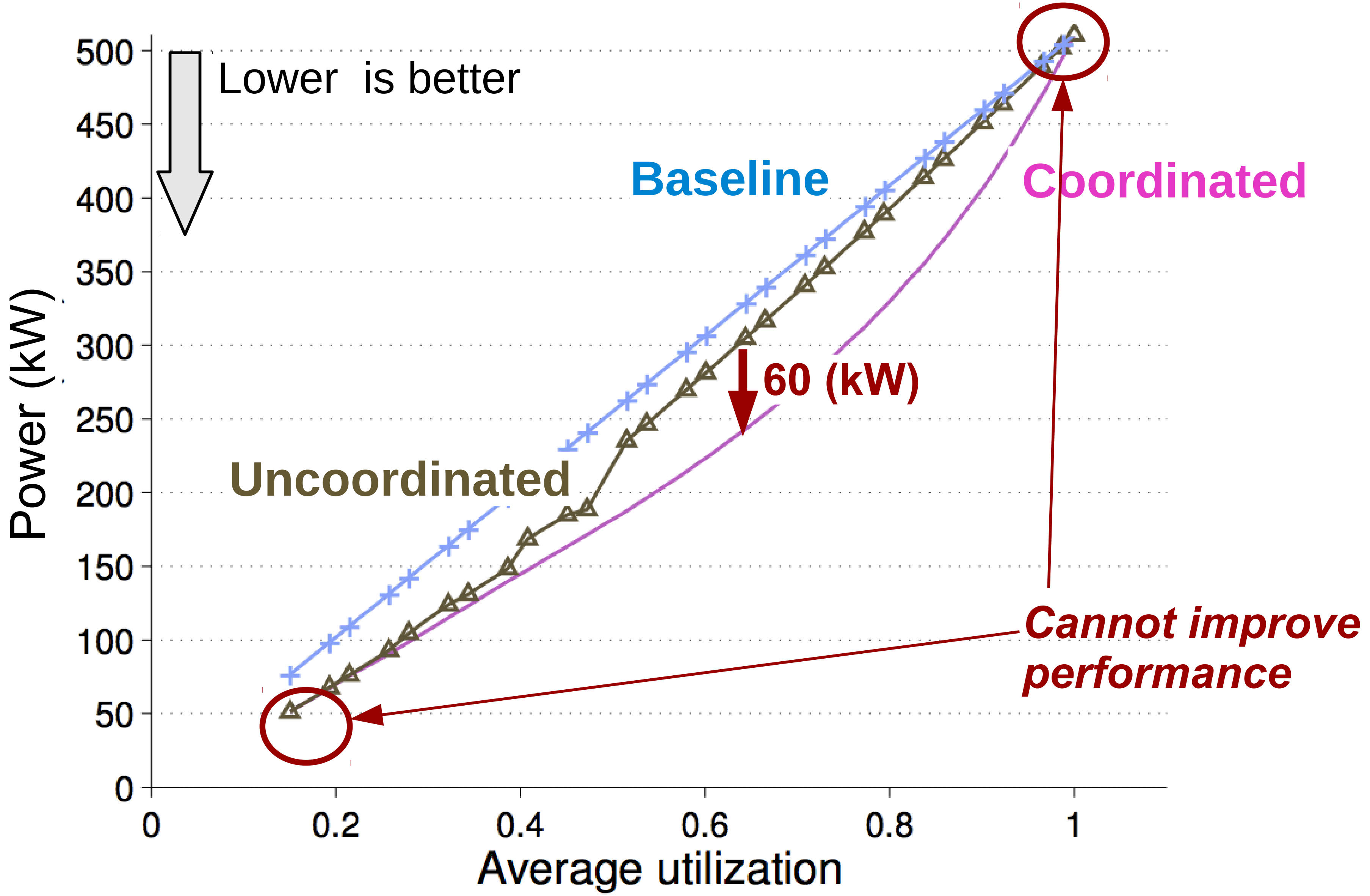
Parolini *et al.* Proceedings of the IEEE 2011

Total power consumption



Parolini *et al.* Proceedings of the IEEE 2011

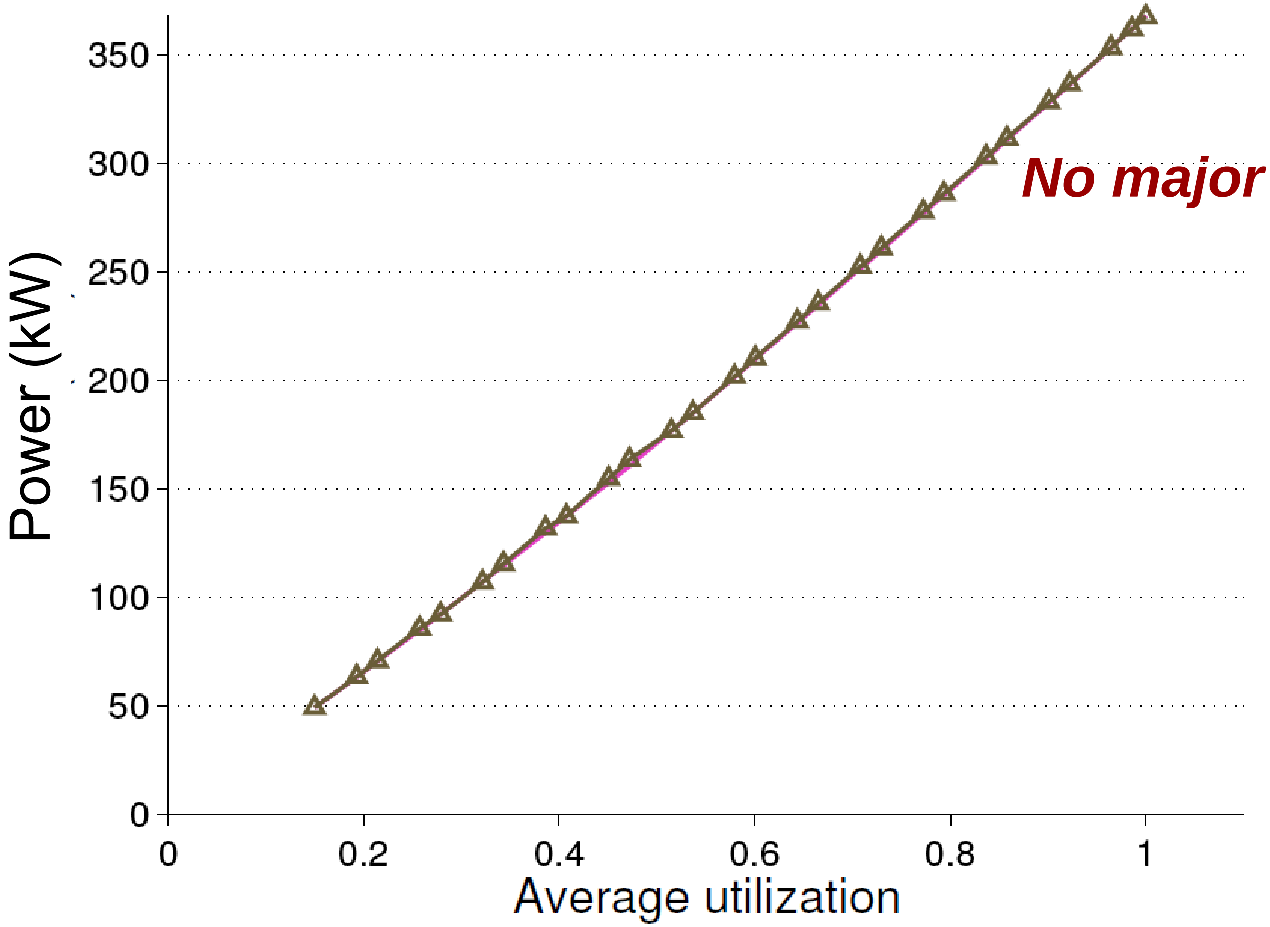
Total power consumption



Parolini *et al.* Proceedings of the IEEE 2011

Total power consumption – case 2

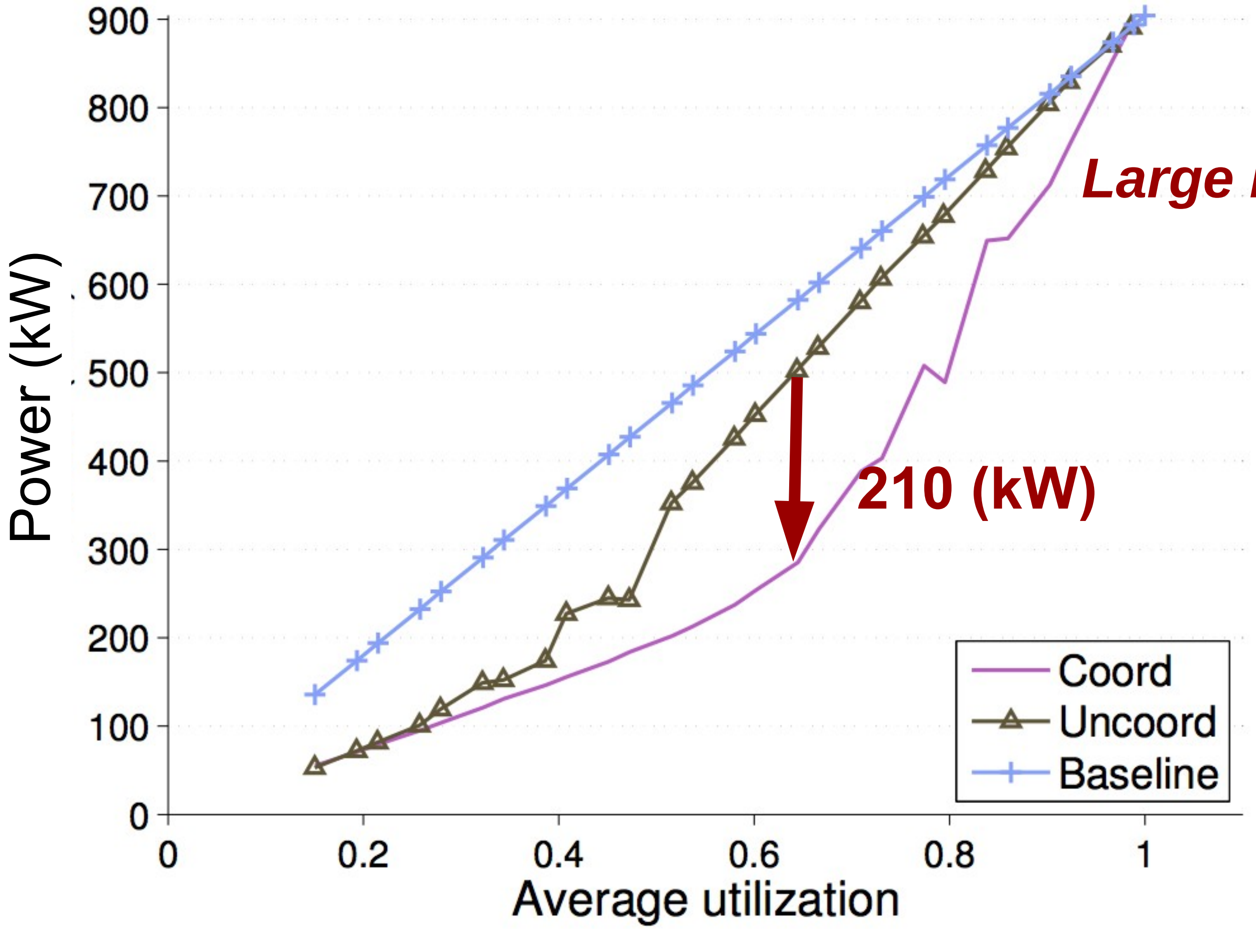
- **How do the controllers perform when all of the zones are efficiently cooled?**



No major improvements

Total power consumption – case 3

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



Large improvements

210 (kW)

Cyber-Physical index

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

Cyber-Physical index

- **Given a data center**
  - How much energy can we save by using 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 relative sensitivity of the zones**

$$\boldsymbol{CPI} = k\,\mathrm{std}\left(\left[\begin{array}{ccc} S_1 & \dots & S_N \end{array}\right]\right)$$

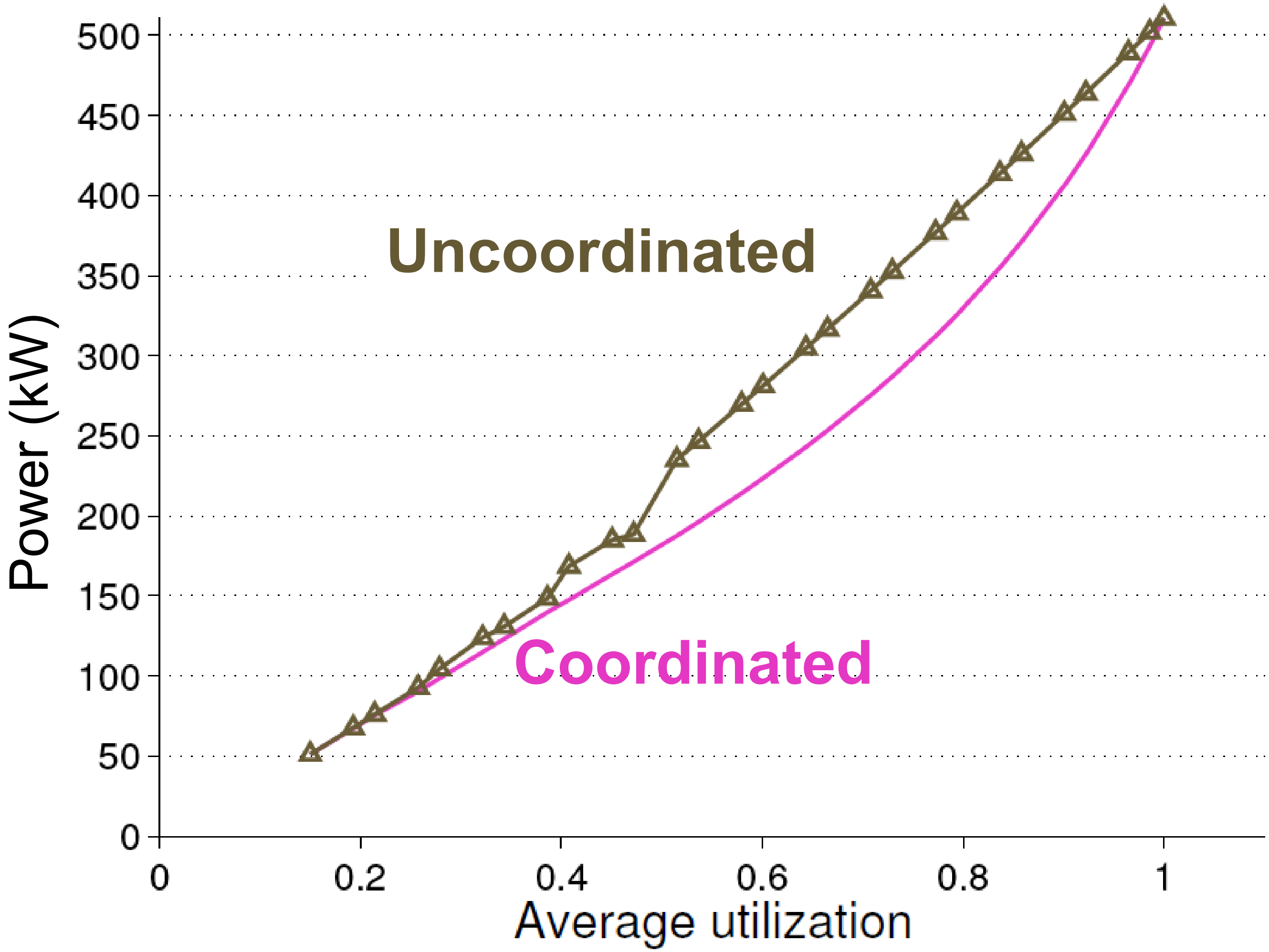
$$S_i = \left\|\left.\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})}\right\|_2 \Big/ \left\|\left.\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})}\right\|_2$$

Relative sensitivity of the ith zone

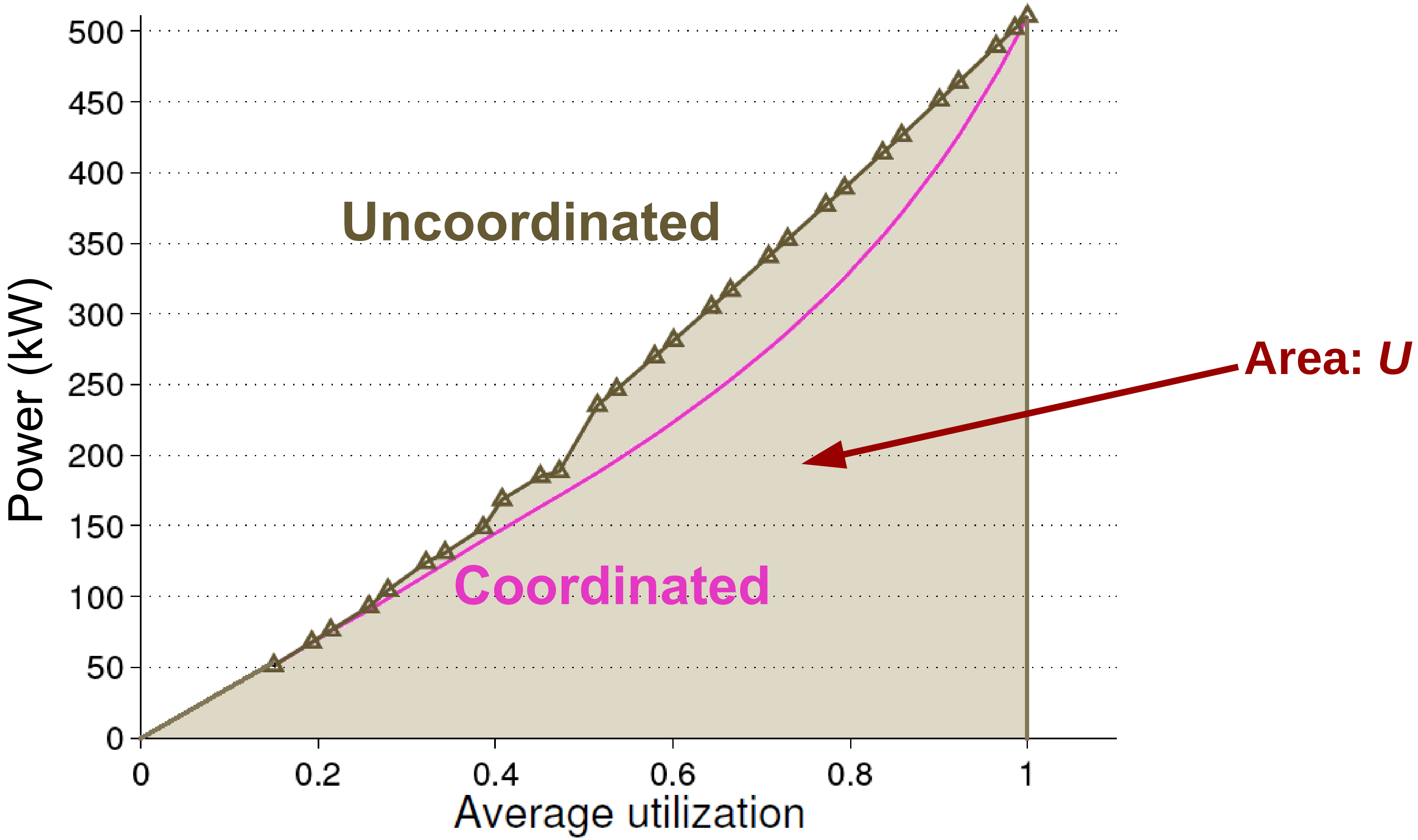
$$\boldsymbol{z} = \left[\begin{array}{cc} \boldsymbol{T}_{\mathrm{ref}}{}^T & \boldsymbol{\eta}^T \end{array}\right]^T$$

Parolini *et al.* Proceedings of the IEEE 2011

Relative efficiency

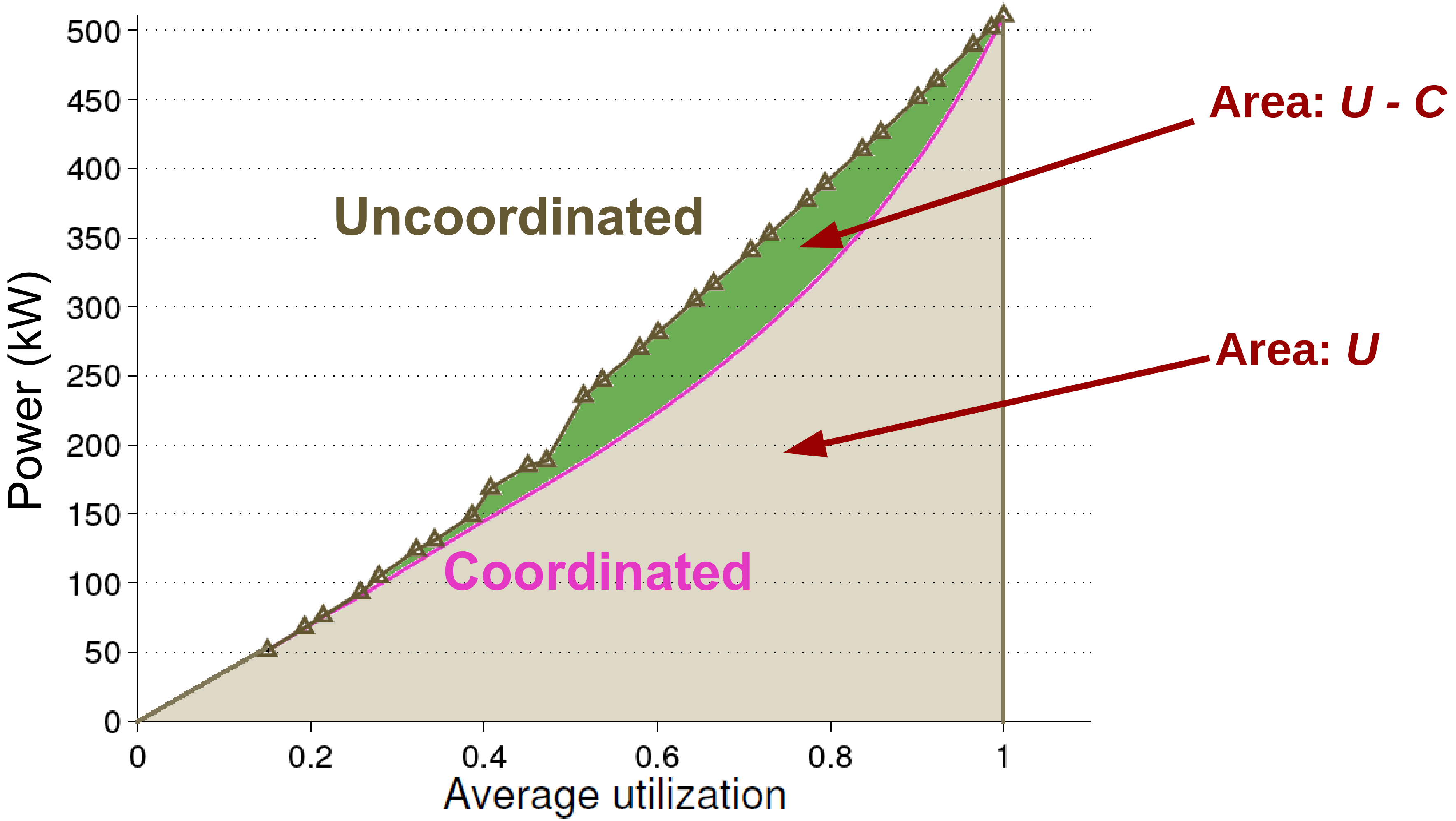


Relative efficiency



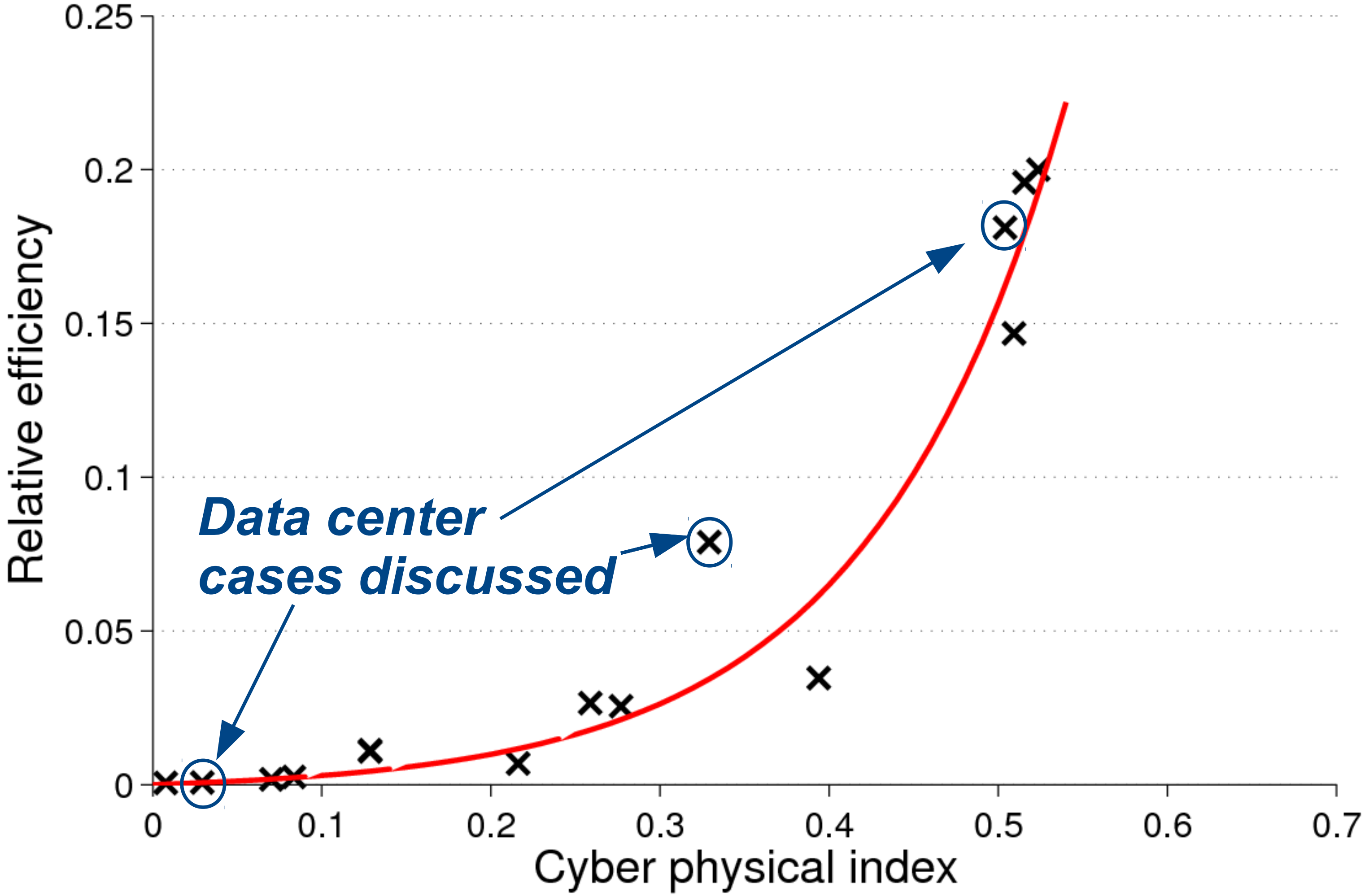
Relative efficiency

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



CPI and relative efficiency

- **Every point represents a different data center**

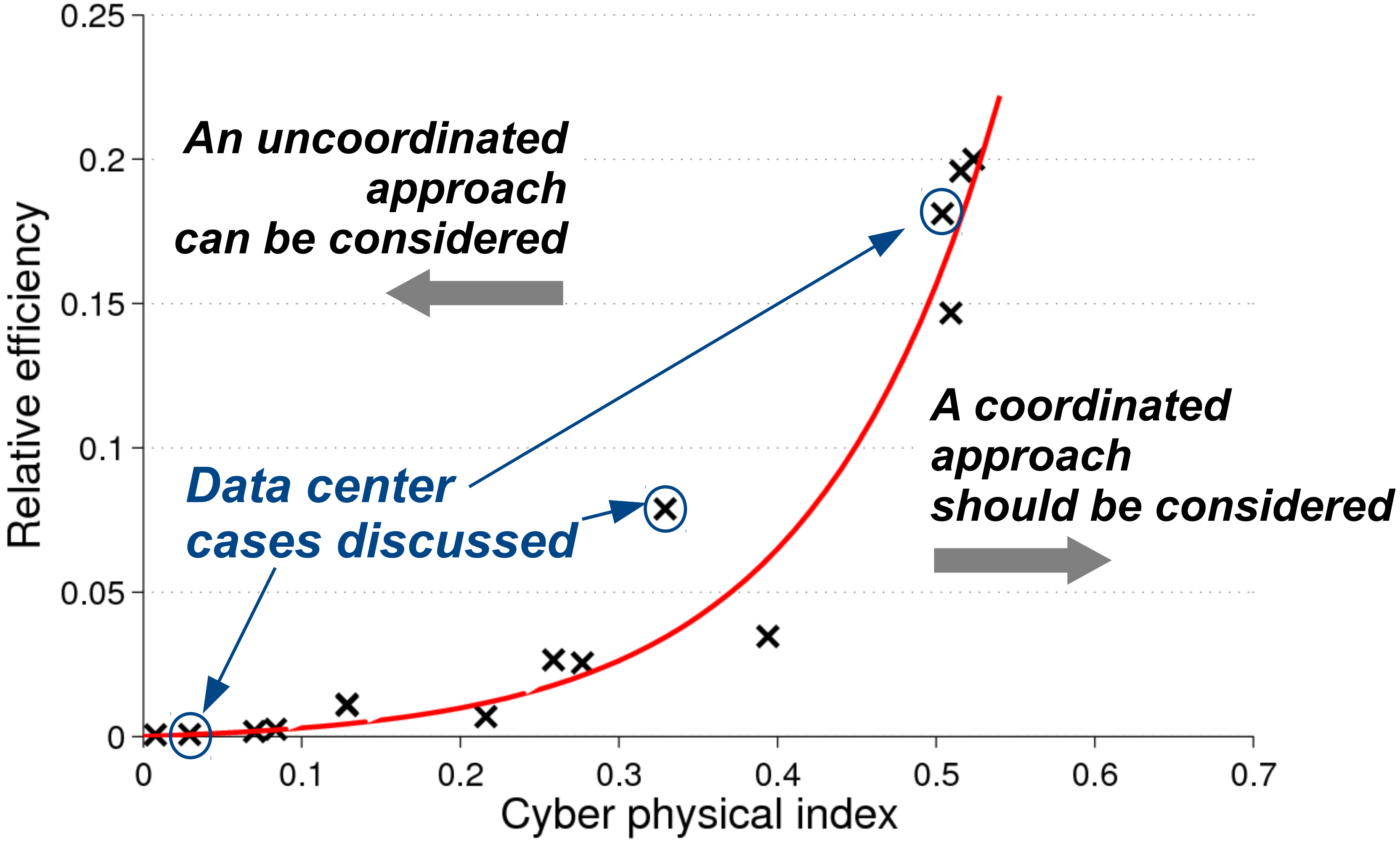


Data center cases discussed

Parolini *et al.* Proceedings of the IEEE 2011

CPI and relative efficiency

- Every point represents a different data center



An uncoordinated approach can be considered

A coordinated approach should be considered

Data center cases discussed

Parolini *et al.* Proceedings of the IEEE 2011

Outline

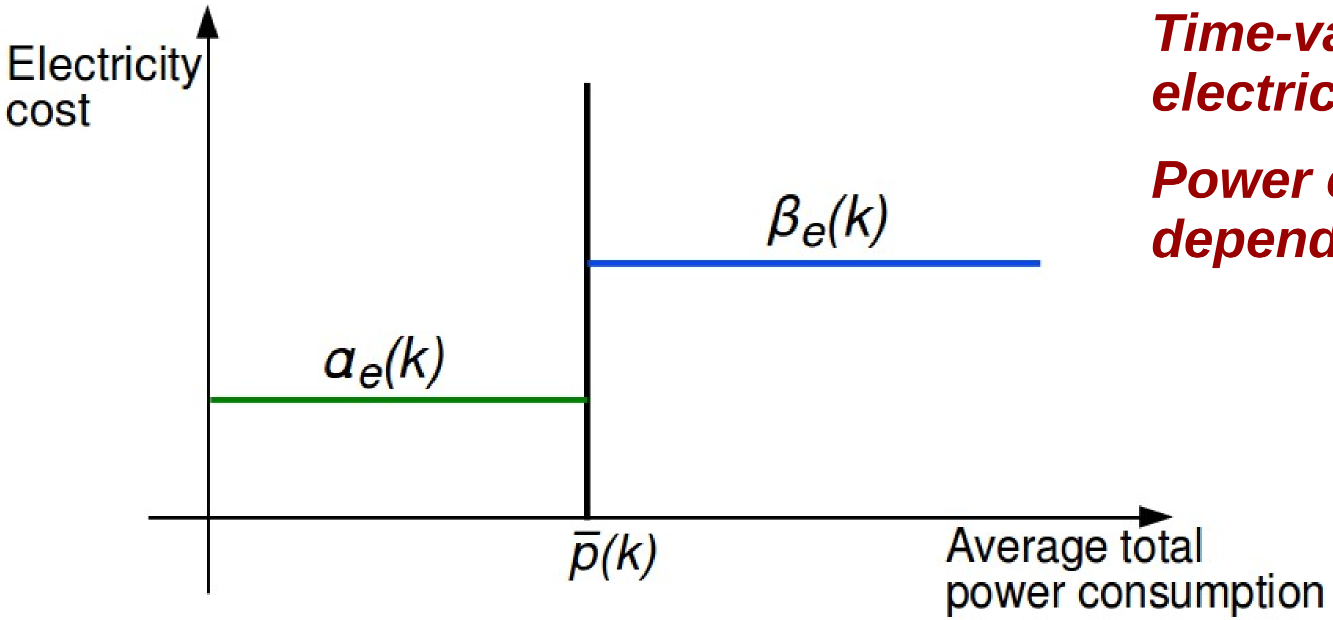
- Introduction
- Proposed control strategy
- Performance analysis for constant job arrival rate
- **Interaction with the smart-grid**
- Zone-level control
- Conclusion and future work

Interaction with the smart-grid

- **Goal of the data-center-level controller**
  - Minimize run-time operating cost
- **Run-time operating cost**
  - Cost of powering the data center
    - Defined by the service level agreement (SLA) with the power-grid, SLA_G
  - Revenue induced by executing the jobs with a certain QoS
    - Defined by the SLA with the users, SLA_U
  - Cost of migrating jobs among the zones

Interaction with the smart-grid

- **Goal of the data-center-level controller**
  - Minimize run-time operating cost
- **Run-time operating cost**
  - Cost of powering the data center
    - Defined by the service level agreement (SLA) with the power-grid, SLA_G

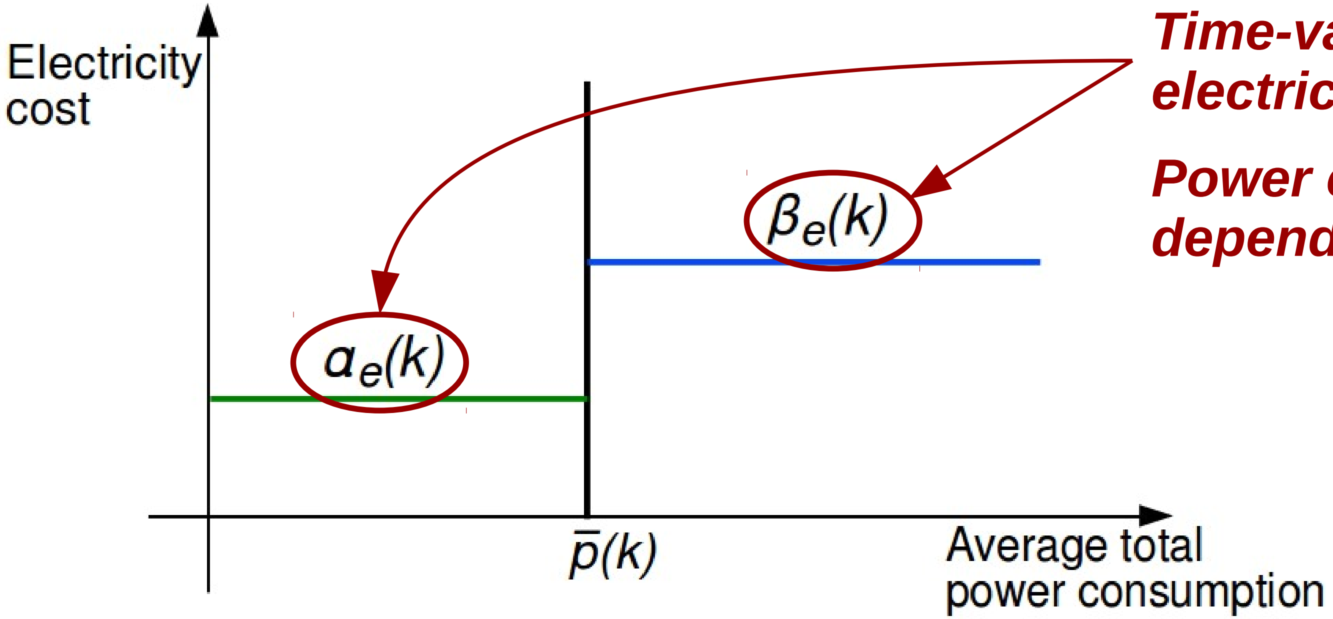


Time-varying electricity price

Power consumption dependent

Interaction with the smart-grid

- **Goal of the data-center-level controller**
  - Minimize run-time operating cost
- **Run-time operating cost**
  - Cost of powering the data center
    - Defined by the service level agreement (SLA) with the power-grid, SLA_G

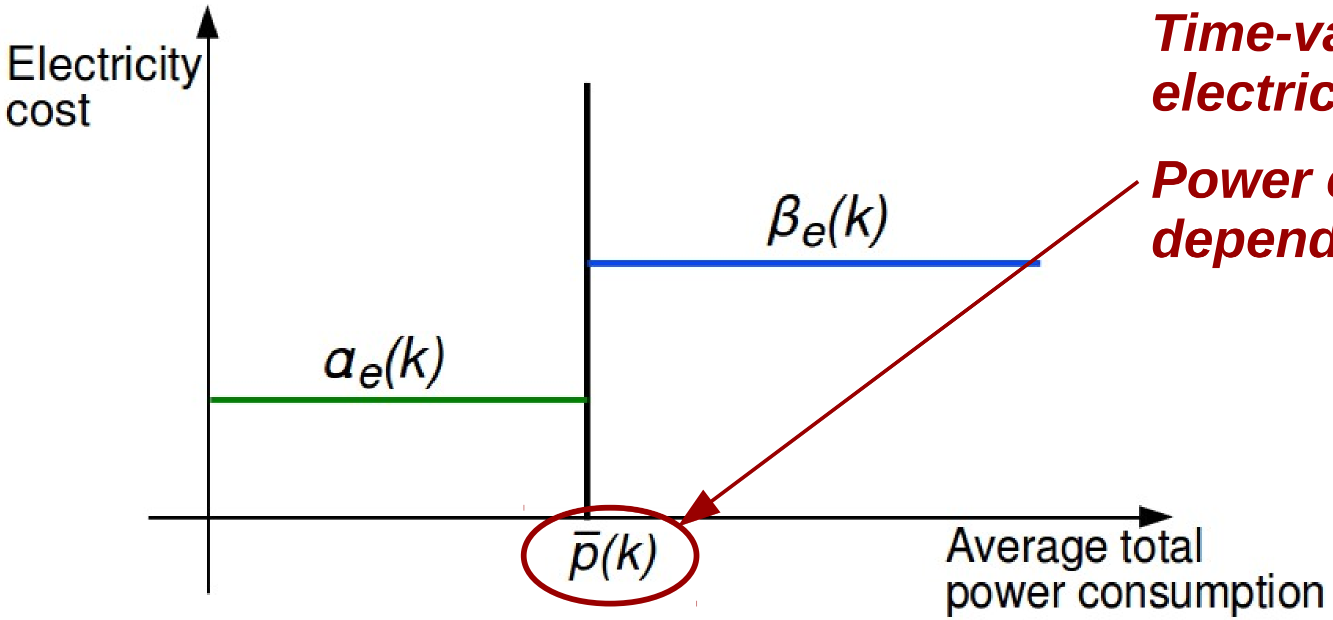


Time-varying electricity price

Power consumption dependent

Interaction with the smart-grid

- **Goal of the data-center-level controller**
  - Minimize run-time operating cost
- **Run-time operating cost**
  - Cost of powering the data center
    - Defined by the service level agreement (SLA) with the power-grid, SLA_G



Time-varying electricity price

Power consumption dependent

Interaction with the smart-grid

- **Goal of the data-center-level controller**
  - Minimize run-time operating cost
- **Run-time operating cost**
  - Cost of powering the data center
    - Defined by the service level agreement (SLA) with the power-grid, SLA_G
  - Revenue induced by executing the jobs with a certain QoS
    - Defined by the SLA with the users, SLA_U
  - Cost of migrating jobs among the zones
- **Controllers have to find the best trade-off between executing jobs with high QoS and reducing the powering cost**

Simulation set-up

- **Neglects the baseline controller**
  - Baseline controller does not consider a cost function in the synthesis of its control strategy

- **Two cases are considered**
  - 1) A coordinated and an uncoordinated controller
    - Controllers are forced to execute jobs with the highest QoS, e.g., neglect SLA_U
    - The uncoordinated controller only considers the reduced electricity price ($\alpha_e(k)$)

Simulation set-up

- **Neglects the baseline controller**
  - Baseline controller does not consider a cost function in the synthesis of its control strategy

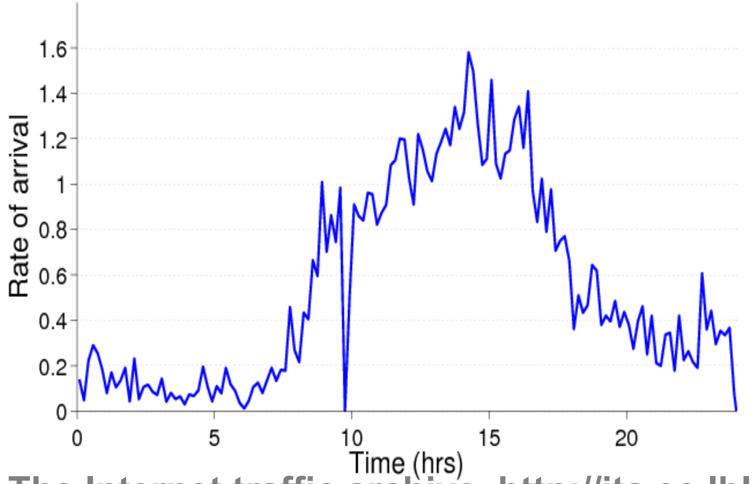
- **Two cases are considered**
  - 1) A coordinated and an uncoordinated controller
    - Controllers are forced to execute jobs with the highest QoS, e.g., neglect SLA_U
    - The uncoordinated controller only considers the reduced electricity price ($\alpha_e(k)$)
  - 2) Two coordinated controllers
    - The first coordinated controller is forced to execute jobs at the highest QoS
    - The parameters in the cost function of the second coordinated controller make it more favorable to drop jobs

Simulation set-up

- **Neglects the baseline controller**
  - Baseline controller does not consider a cost function in the synthesis of its control strategy
- **Two cases are considered**
  - 1) A coordinated and an uncoordinated controller
    - Controllers are forced to execute jobs with the highest QoS, e.g., neglect SLA_U
    - The uncoordinated controller only considers the reduced electricity price ($\alpha_e(k)$)
  - 2) Two coordinated controllers
    - The first coordinated controller is forced to execute jobs at the highest QoS
    - The parameters in the cost function of the second coordinated controller make it more favorable to drop jobs

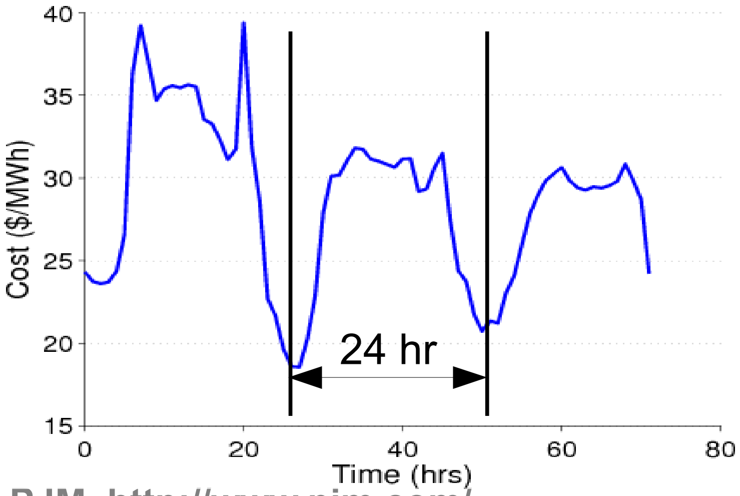
Job arrival rate and electricity costs

- **Real data**
  - Job arrival rate



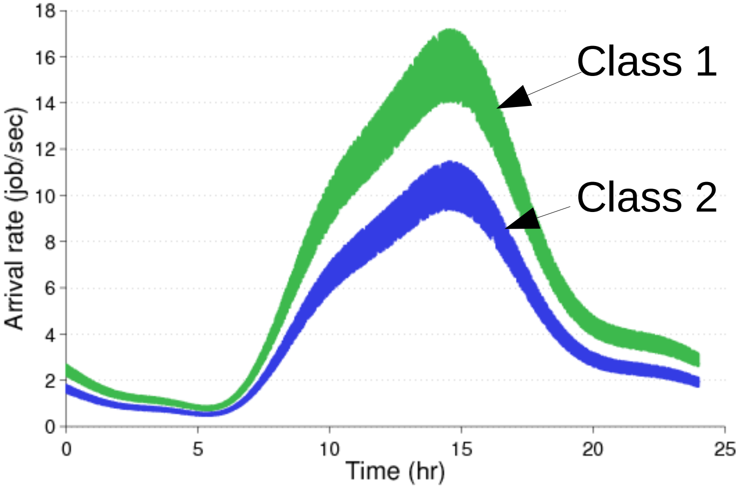
The Internet traffic archive, http://ita.ee.lbl.gov/

  - Day-ahead electricity cost

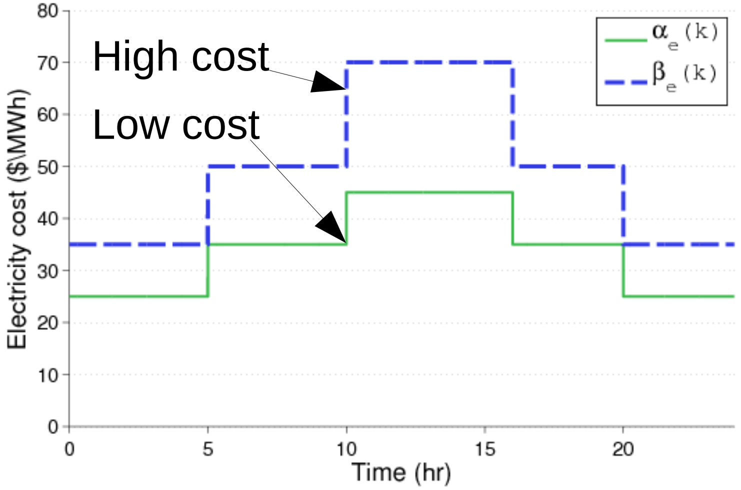


PJM, http://www.pjm.com/

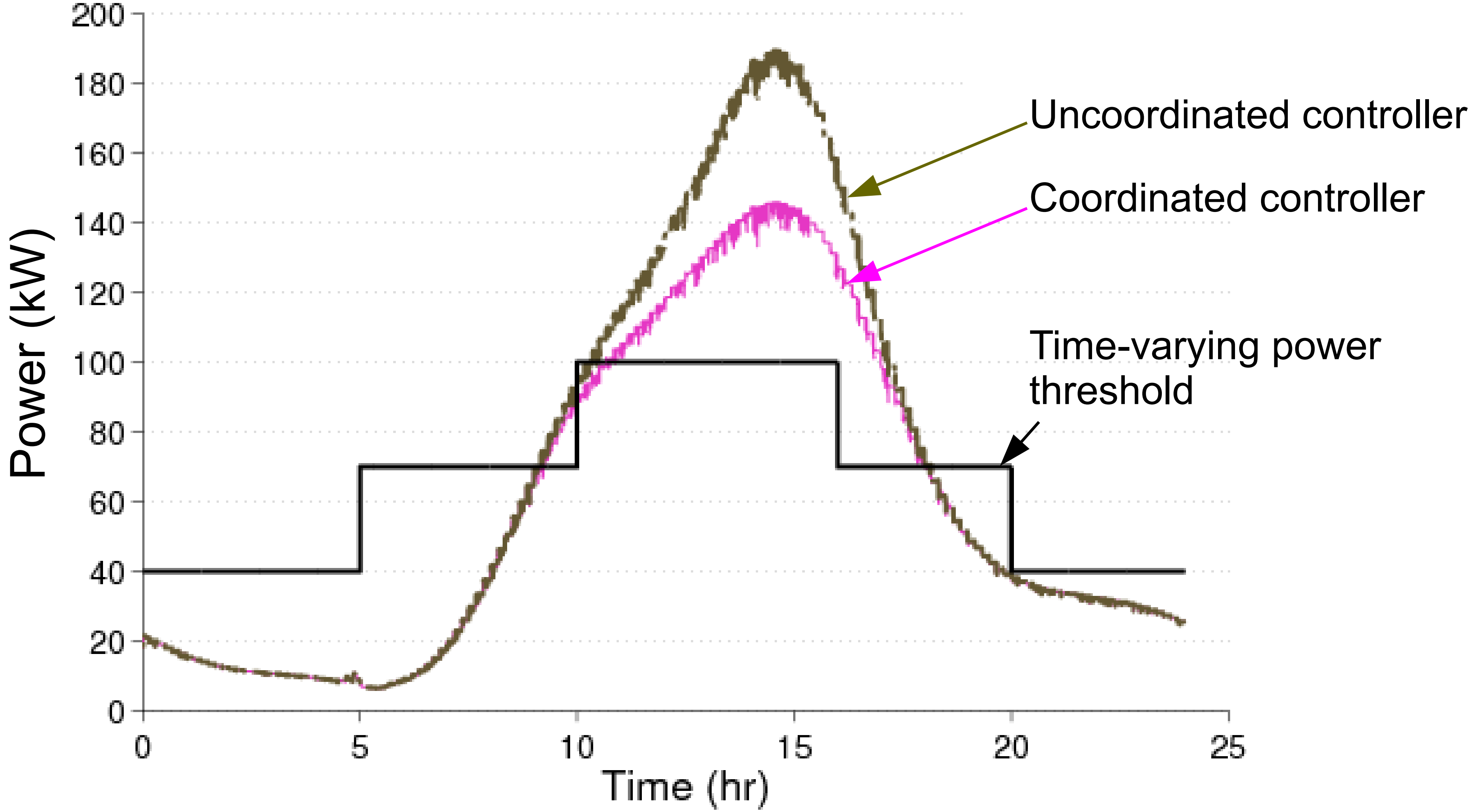
- **Parameters of the simulation**
  - Job arrival rate, 2 classes of jobs



  - Electricity costs

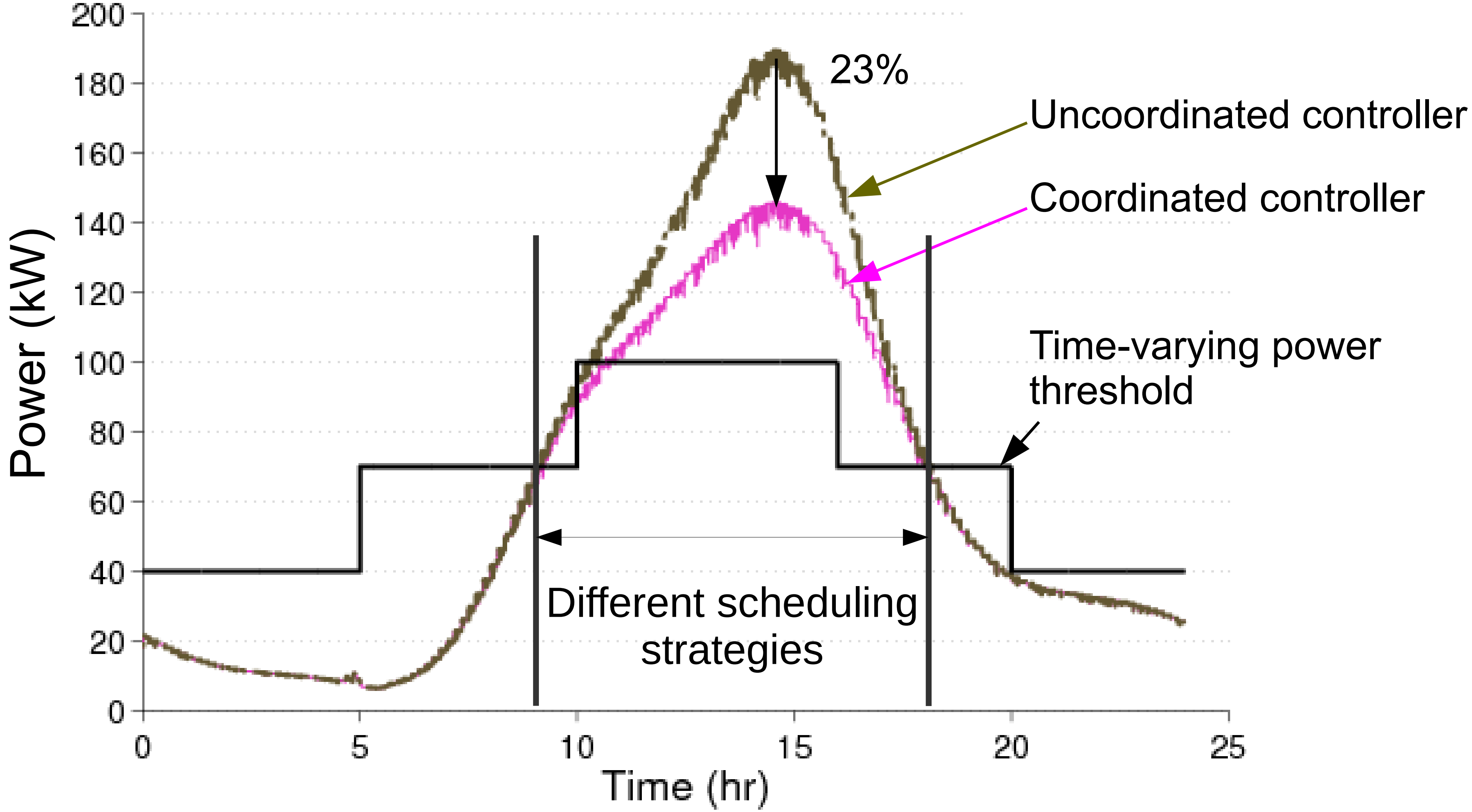


Total data center power consumption



Parolini *et al.* Proc. of IFAC 2011

Total data center power consumption



23%

Uncoordinated controller

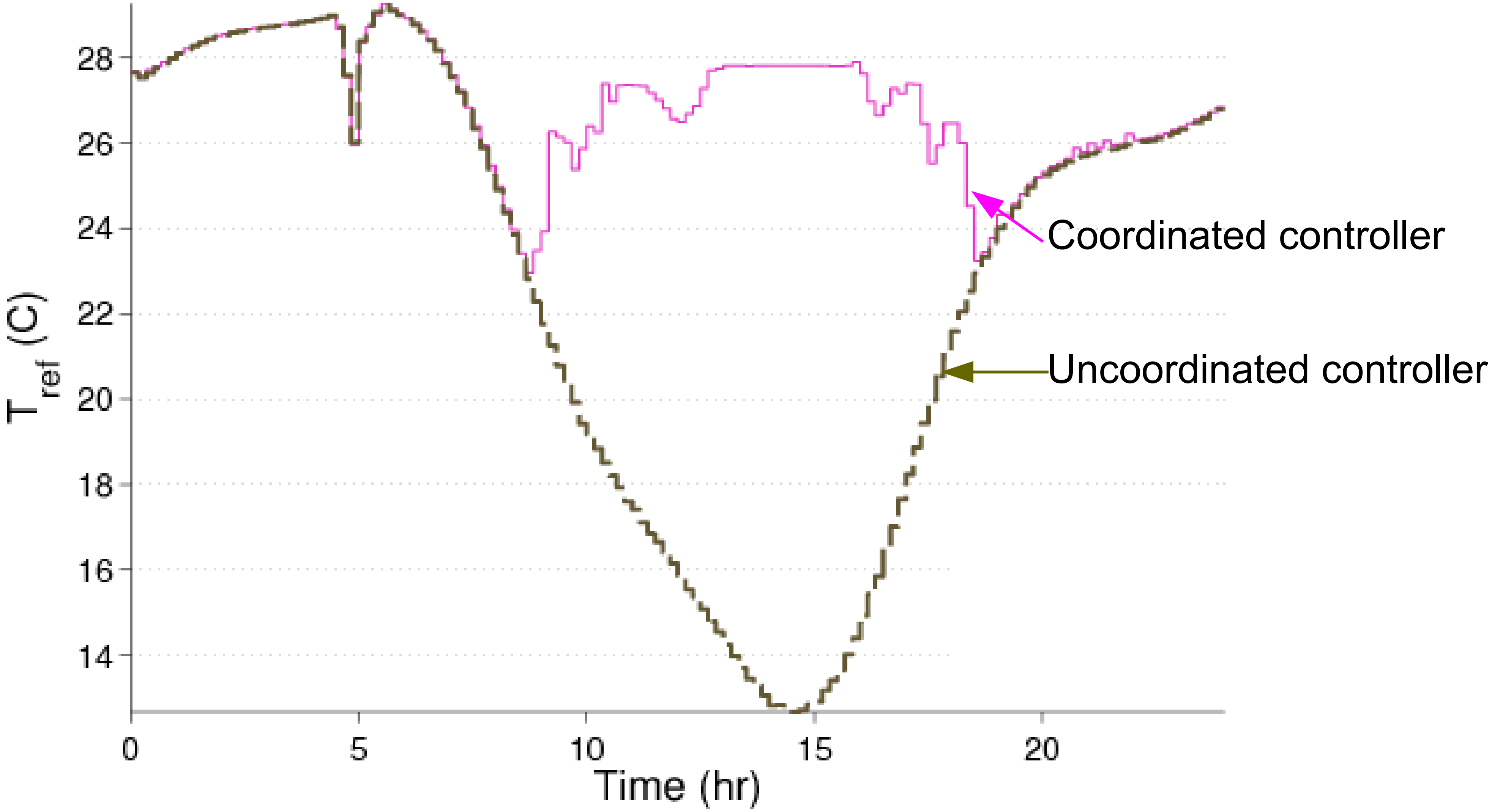
Coordinated controller

Time-varying power threshold

Different scheduling strategies

Parolini *et al.* Proc. of IFAC 2011

Average reference temperatures

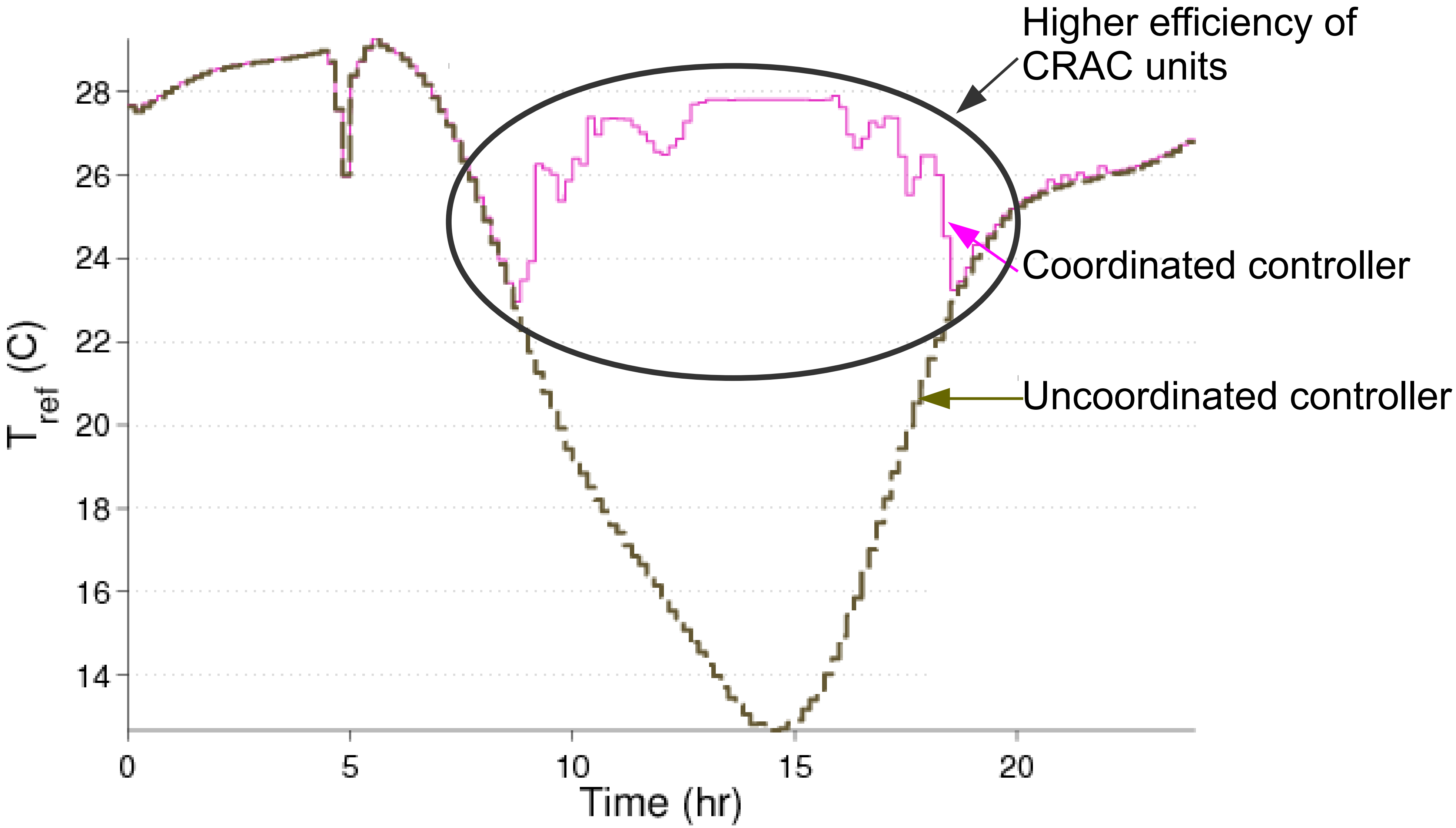


Coordinated controller

Uncoordinated controller

Parolini *et al.* Proc. of IFAC 2011

Average reference temperatures



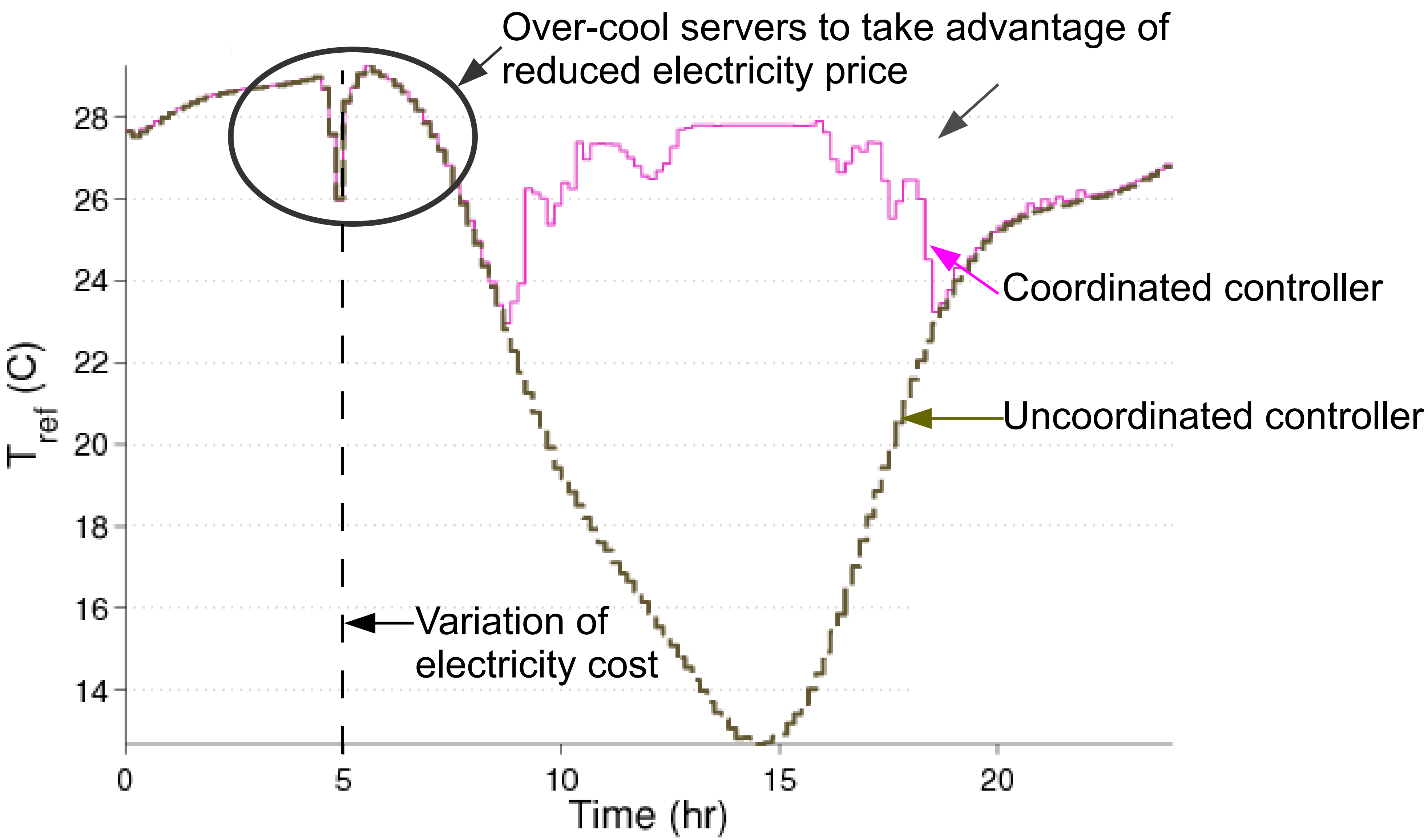
Higher efficiency of CRAC units

Coordinated controller

Uncoordinated controller

Parolini *et al.* Proc. of IFAC 2011

Average reference temperatures



Over-cool servers to take advantage of reduced electricity price

Coordinated controller

Uncoordinated controller

Variation of electricity cost

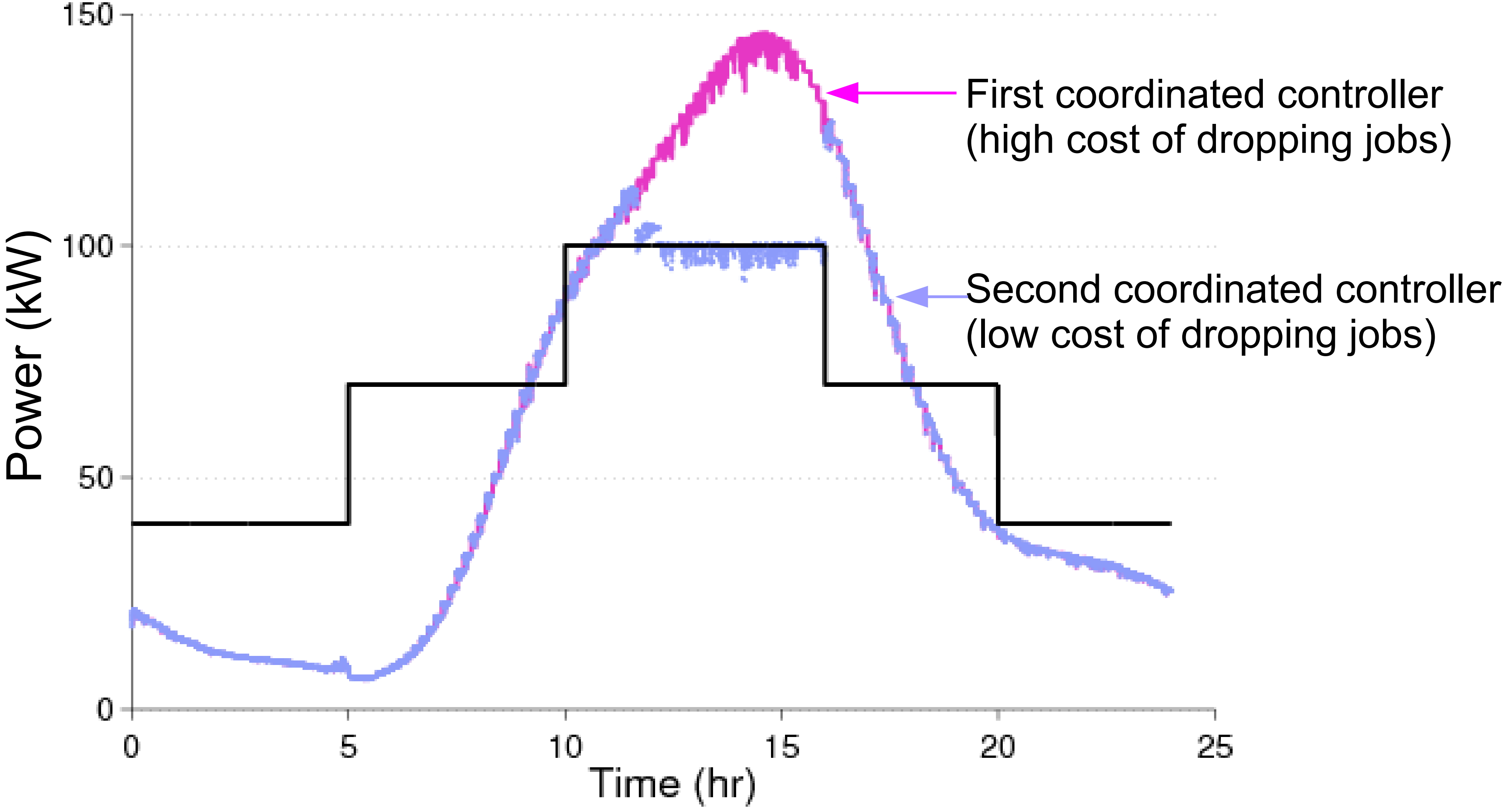
Parolini *et al.* Proc. of IFAC 2011

Simulation set-up

- **Neglects the baseline controller**
  - Baseline controller does not consider a cost function in the synthesis of its control strategy

- **Two cases are considered**
  - 1) A coordinated and an uncoordinated controller
    - Controllers are forced to execute jobs with the highest QoS, e.g., neglect SLA_U
    - The uncoordinated controller only considers the reduced electricity price ($\alpha_e(k)$)
  - 2) Two coordinated controllers
    - The first coordinated controller is forced to execute jobs at the highest QoS
    - The parameters used in the cost function of the second coordinated controller make it more favorable to drop jobs

Total data center power consumption

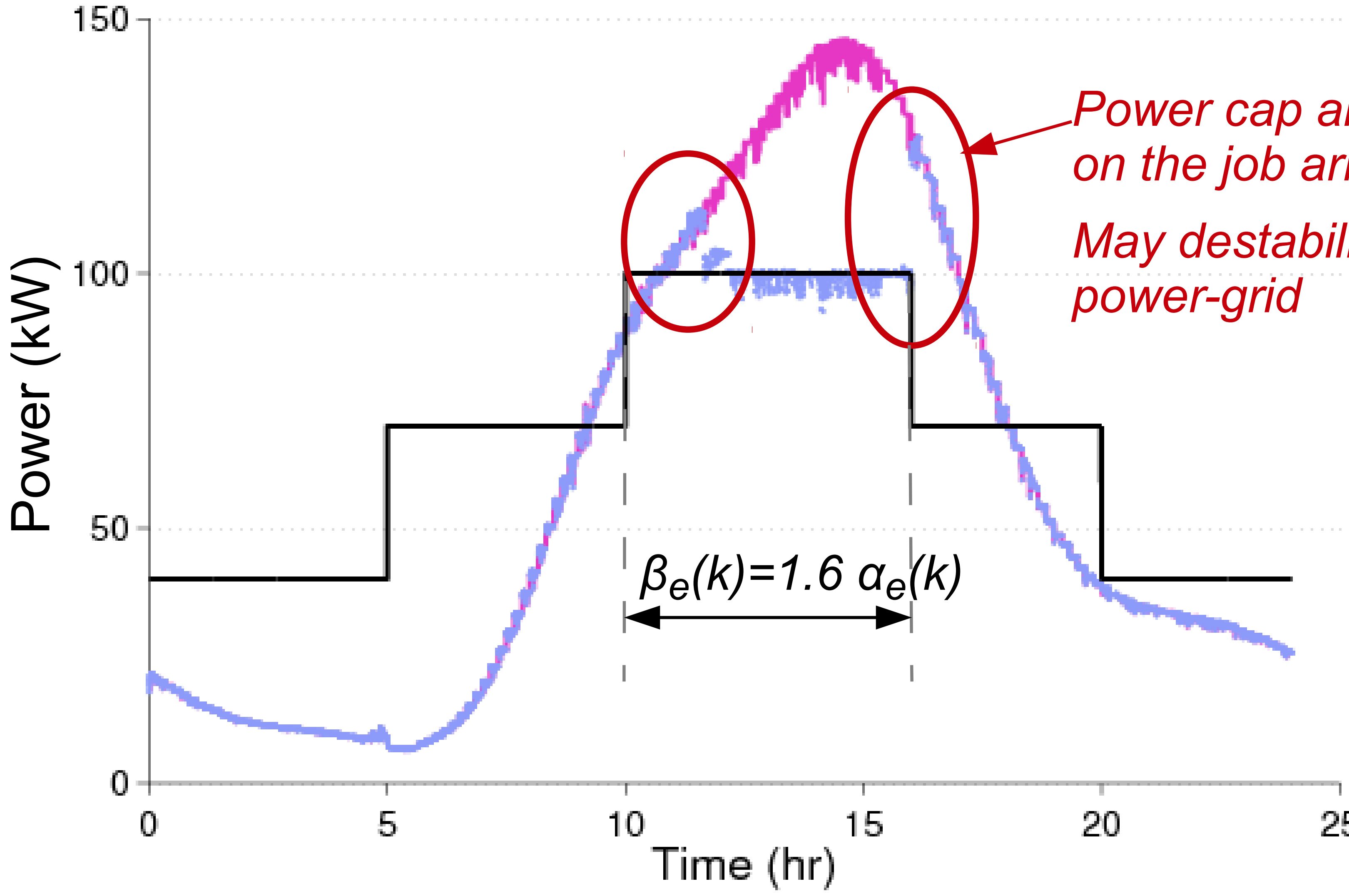


First coordinated controller (high cost of dropping jobs)

Second coordinated controller (low cost of dropping jobs)

Parolini *et al.* IFAC 2011

Total data center power consumption



Power cap also depends on the job arrival rate

May destabilize the power-grid

Parolini *et al.* IFAC 2011

Outline

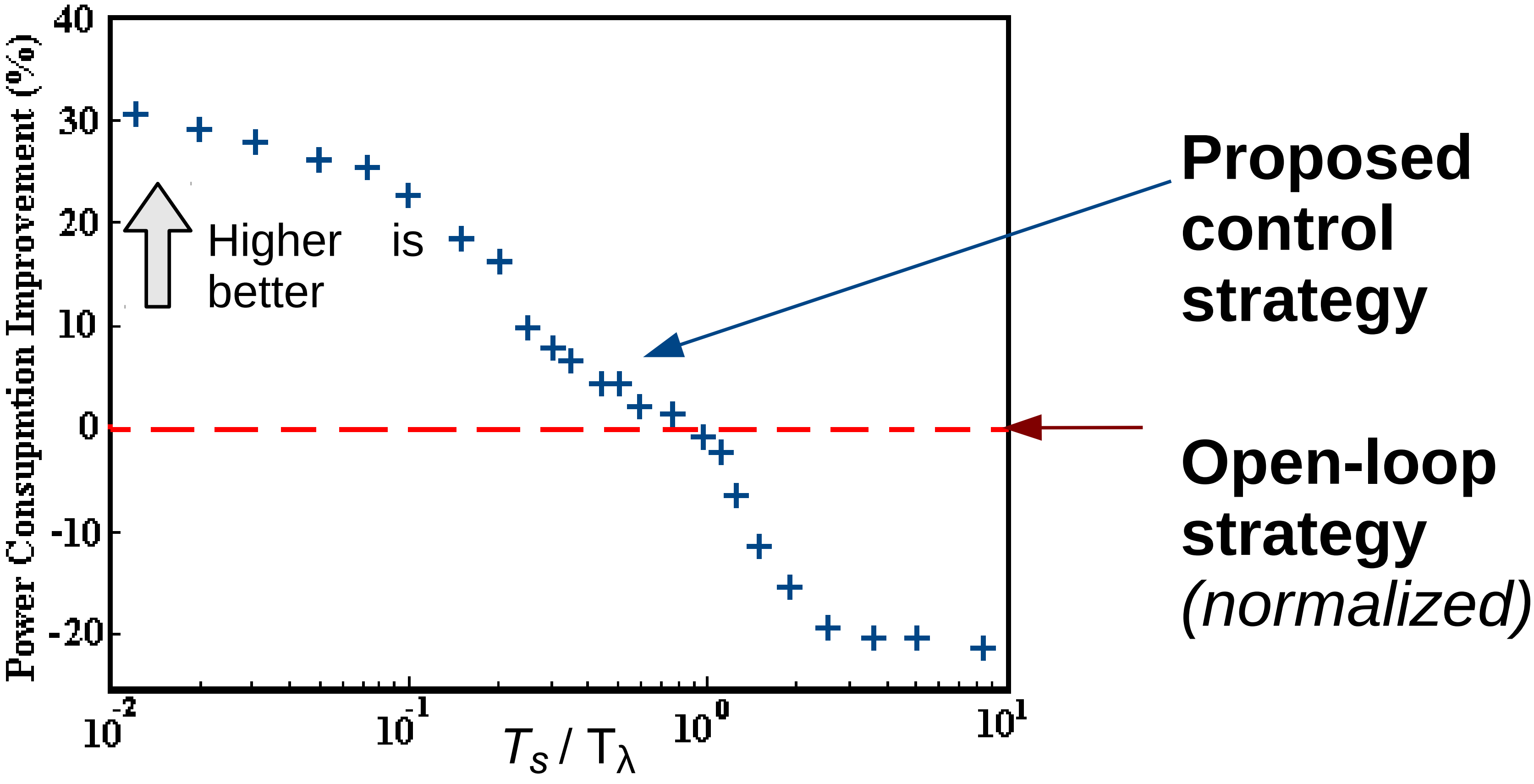
- Introduction
- Proposed control strategy
- Performance analysis for constant job arrival rate
- Interaction with the smart-grid
- **Zone-level control**
- Conclusion and future work

Zone-level controller

- **Operates on the minutes time-scale**
- **Decides how many servers in the zone should be turned on**
  - Serves take time to turn on
  - When turning on, servers consume power and do not execute jobs
- **Control actions based on**
  - Desired jobs execution rate (predictive control)
  - Current use of resources in the zone (reactive control)
- **Considers**
  - Time to turn servers on: T_S
  - Variability of workload arrival rate: $1/T_\lambda$
- **Controller details in the dissertation**

Simulation results

- **Compare the proposed control approach against optimal open-loop strategy**
  - The number of ON servers is *fixed* and based on the long-term average job arrival rate



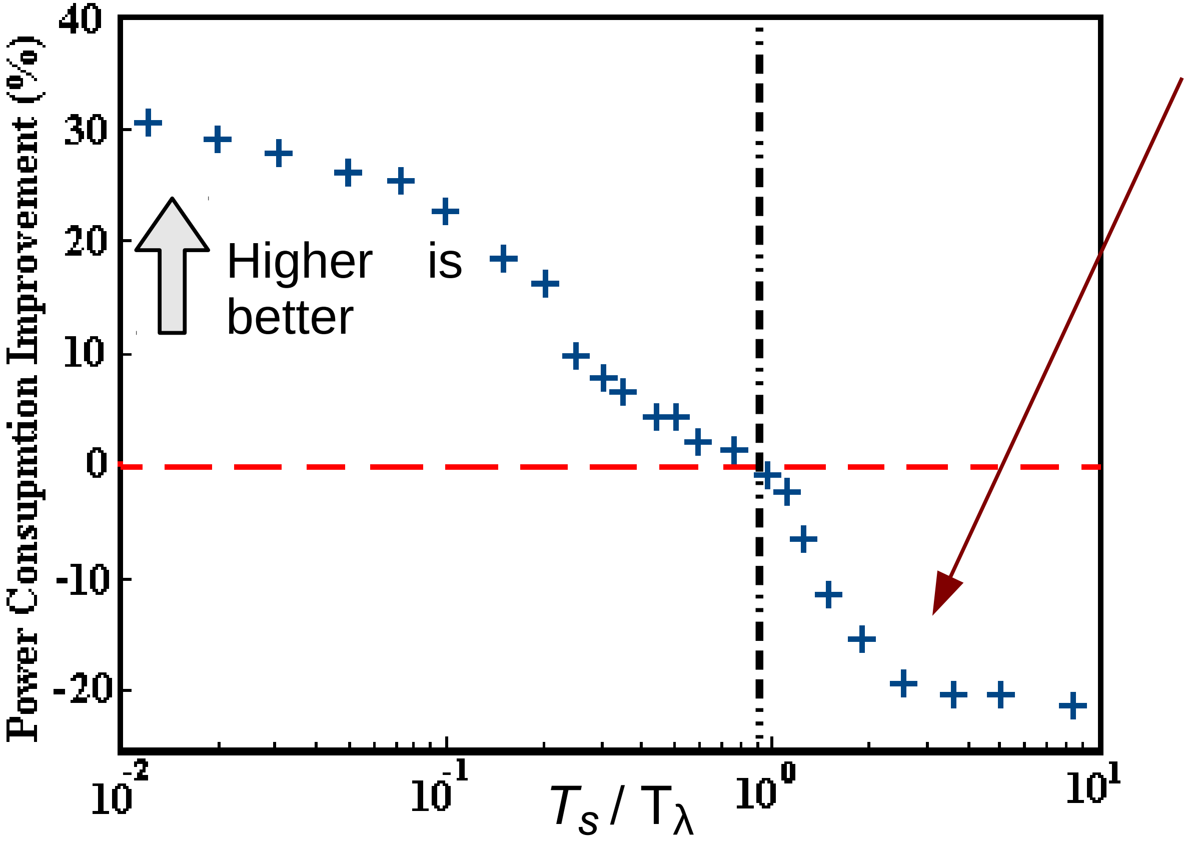
Proposed control strategy

Open-loop strategy *(normalized)*

Aghajani, Parolini, Sinopoli. CDC 2010

Simulation results

- **Compare the proposed control approach against optimal open-loop strategy**
  - The number of ON servers is *fixed* and based on the long-term average job arrival rate



- **Workload arrival rate varies too rapidly with respect to the time to turn servers on**
- **Best solution is to never turn servers on and off**

Aghajani, Parolini, Sinopoli. CDC 2010

Outline

- Introduction
- Proposed control strategy
- Performance analysis for constant job arrival rate
- Interaction with the smart-grid
- Zone-level control
- **Conclusion and future work**

Conclusion

- **Data centers have operating regimes for which the baseline and the uncoordinated controller are as optimal as the coordinated controller**
- **There are cases where the uncoordinated controller is *always* as optimal as the coordinated controller**
- **Proposed the cyber-phyiscal index**
  - First attempt to characterize the thermal and computational characteristics of the data center with a single index
- **Data centers can take advantage of SLAs with the power grid**
  - Depending on the SLA, the data center may destabilize the power-grid
- **If the job arrival rate changes too quickly, then never turn servers off is the best strategy**

Future work

- **Model validation**
  - Multiple data center cases have to be considered in order to understand which model is useful in which condition
- **Data collection and sharing with the research community**
  - The lack of public available data is one of the major issue related to data center modeling and control
- **Interaction with the power-grid**
  - How should the SLA with the power-grid be stipulated, in order to avoid sudden changes in the power demand?
- **Management of data storage**
  - Including management of data storage in the optimal control problem may improve the overall performance
- **Further development of the data center simulator**
  - A data center simulator used and recognized by the whole research community would allow the comparison of different control algorithms under the same test cases

Publications

- *L. Parolini*, B. Sinopoli, B. H. Krogh. Models and Control Strategies for Data Centers in the Smart Grid. In Control and Optimization Theory for Electric Smart Grids, Springer
- *L. Parolini*, B. Sinopoli, B. H. Krogh, Z. Wang. A Cyber-Physical-System Approach to Data Center Modeling and Control for Energy Efficiency. Proceedings of the IEEE, Special Issue on Cyber-Physical Systems
- *L. Parolini*, B. Sinopoli, B. H. Krogh. Model predictive control of data centers in the smart grid scenario. 18*th* World Congress of the International Federation of Automatic Control. August 2011
- *L. Parolini*, E. Garone, B. Sinopoli, B. H. Krogh. A Hierarchical Approach to Energy Management in Data Centers. 49th IEEE Conference on Decision and Control. December 2010
- M. Aghajani, *L. Parolini*, B. Sinopoli. Dynamic Power Allocation in Server Farms: a Real Time Optimization Approach. 49*th* IEEE Conference on Decision and Control. December 2010
- *L. Parolini*, N. Tolia, B. Sinopoli, B. H. Krogh. A Cyber-Physical Systems Approach to Energy Management in Data Centers. First International Conference on Cyber-Physical Systems. April 2010
- *L. Parolini*, B. Sinopoli, B. H. Krogh. A Unified Thermal-Computational Approach to Data Center Energy Management. Fourth International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks. April 2009
- *L. Parolini*, B. Sinopoli, B. H. Krogh. Reducing Data Center Energy Consumption via Coordinated Cooling and Load Management. HotPower. December 2008

Thank you!

Additional slides

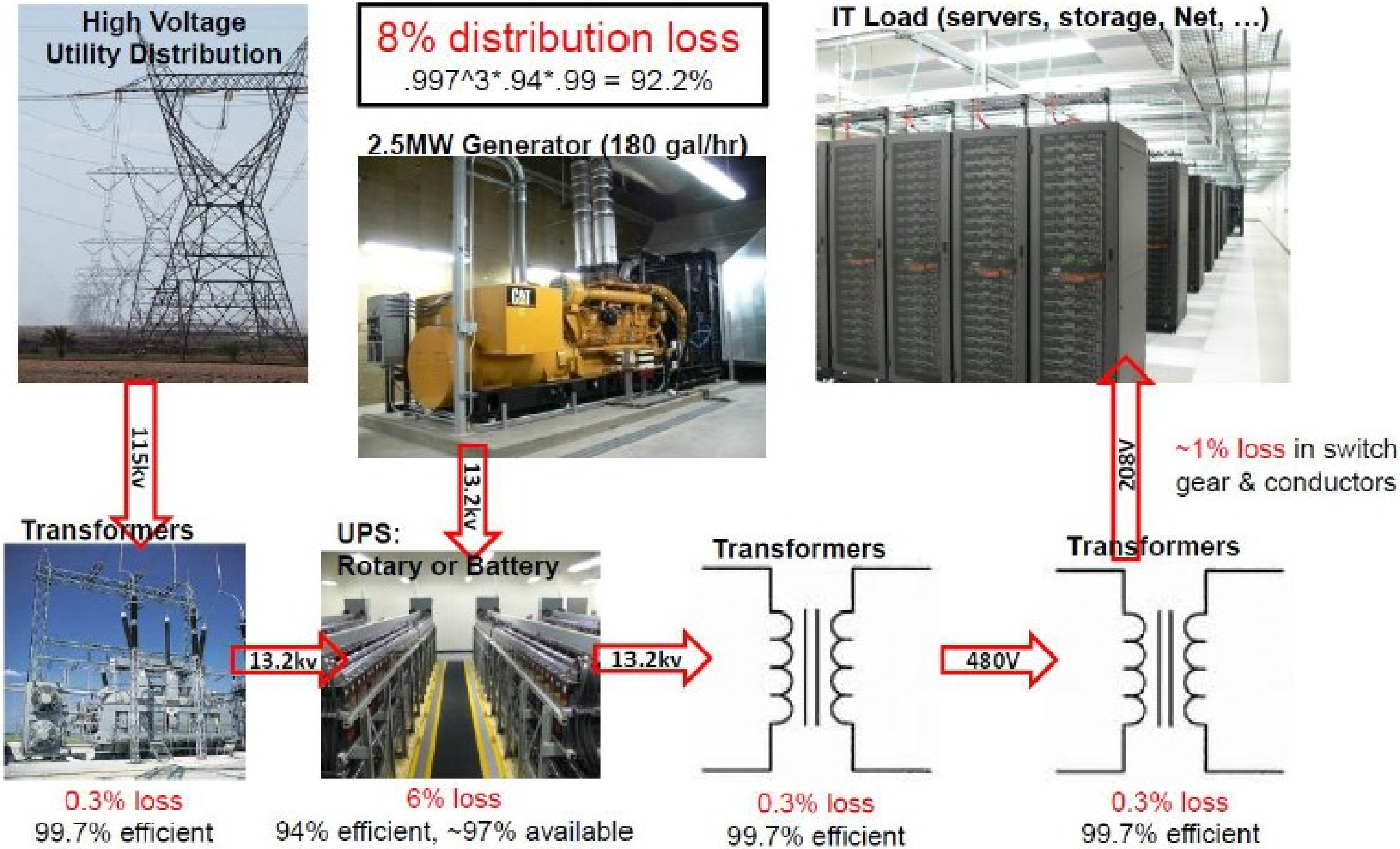
Introduction to Data centers

What is a data center?

- A data center is a facility used to house computer systems and associated components

[http://en.wikipedia.org/wiki/Data_center]

Power distribution



High Voltage Utility Distribution

8% distribution loss

.997^3*.94*.99 = 92.2%

2.5MW Generator (180 gal/hr)

IT Load (servers, storage, Net, …)

115kv

13.2kv

208V

~1% loss in switch gear & conductors

Transformers

UPS:
Rotary or Battery

13.2kv

13.2kv

Transformers

480V

Transformers

0.3% loss
99.7% efficient

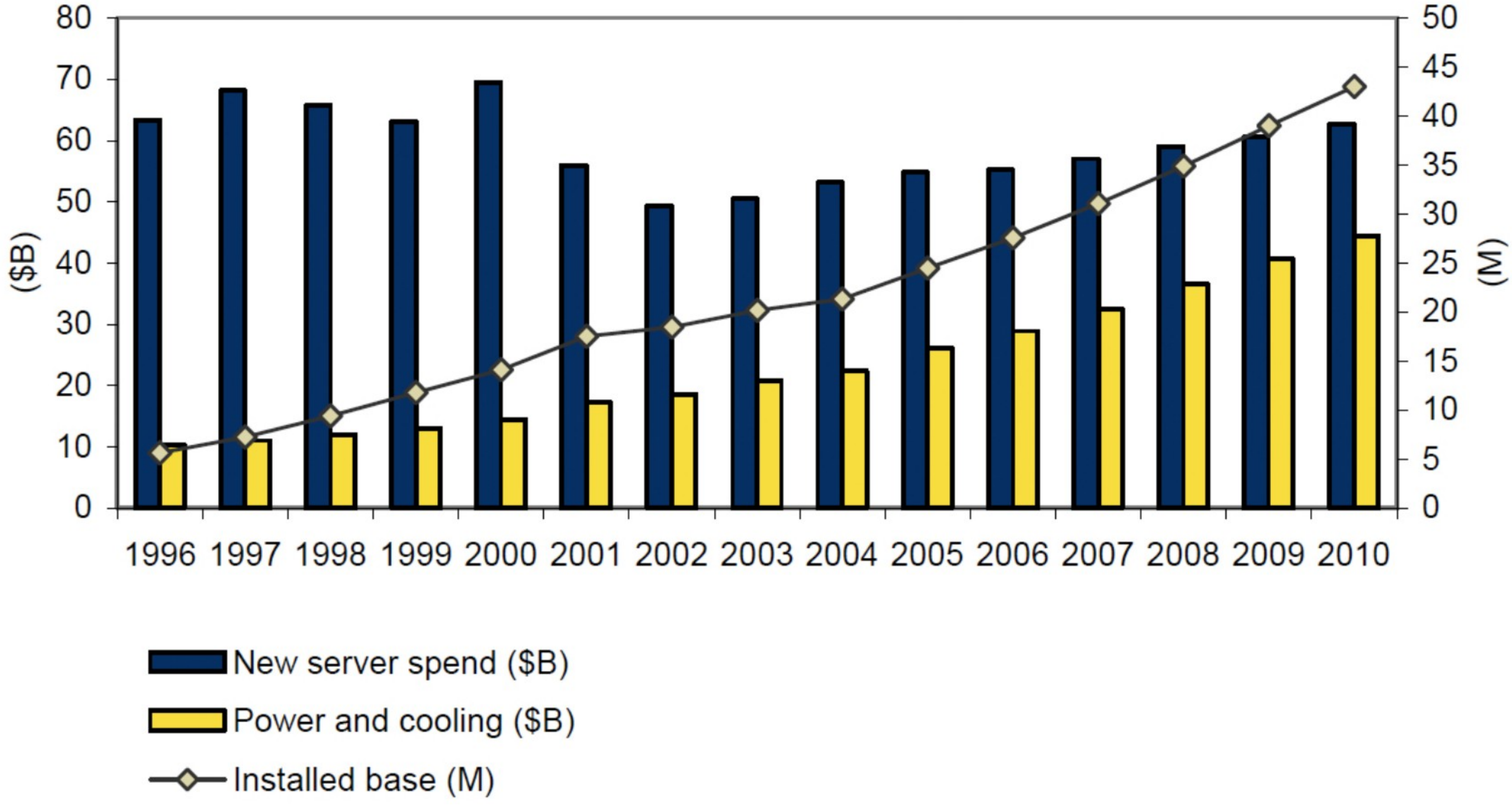
6% loss
94% efficient, ~97% available

0.3% loss
99.7% efficient

0.3% loss
99.7% efficient

J. Hamilton. Data Center Efficiency Best Practice. Slides. Apr. 09

Spending trend



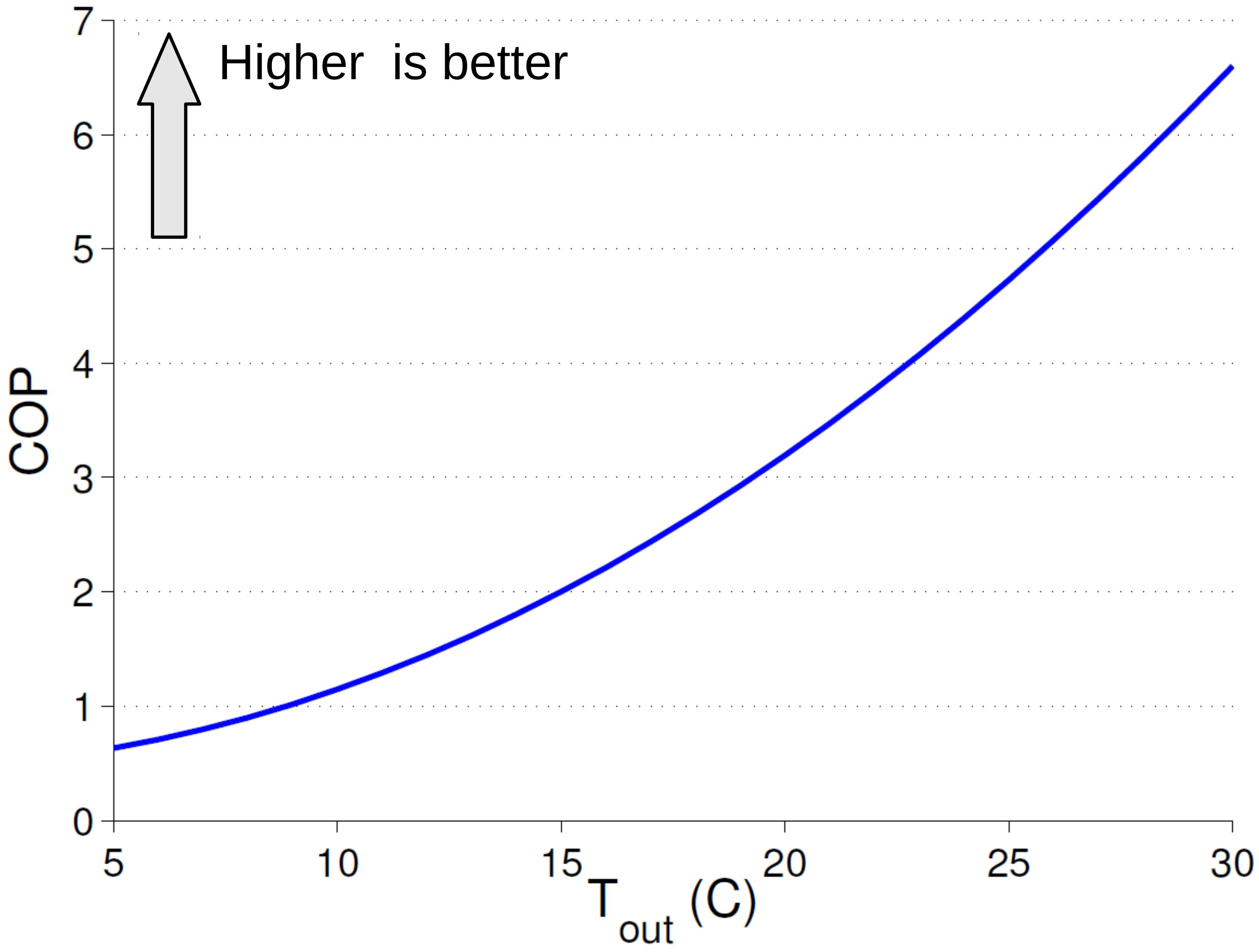
International Data Corporation, Tech. Rep. 2007

Data centers and cyber-physical systems

- **We have designed CPS since the advent of the computer**
  - A digitally controlled system is a CPS
- **Is it any different now? If so what is different?**
  - No separation of concerns
  - Coupling
  - Sensitivity
- **Data Centers are a prime example of a CPS**

Parolini et al. "A Cyber–Physical Systems Approach to Data Center Modeling and Control for Energy Efficiency", Proceedings of the IEEE, to appear

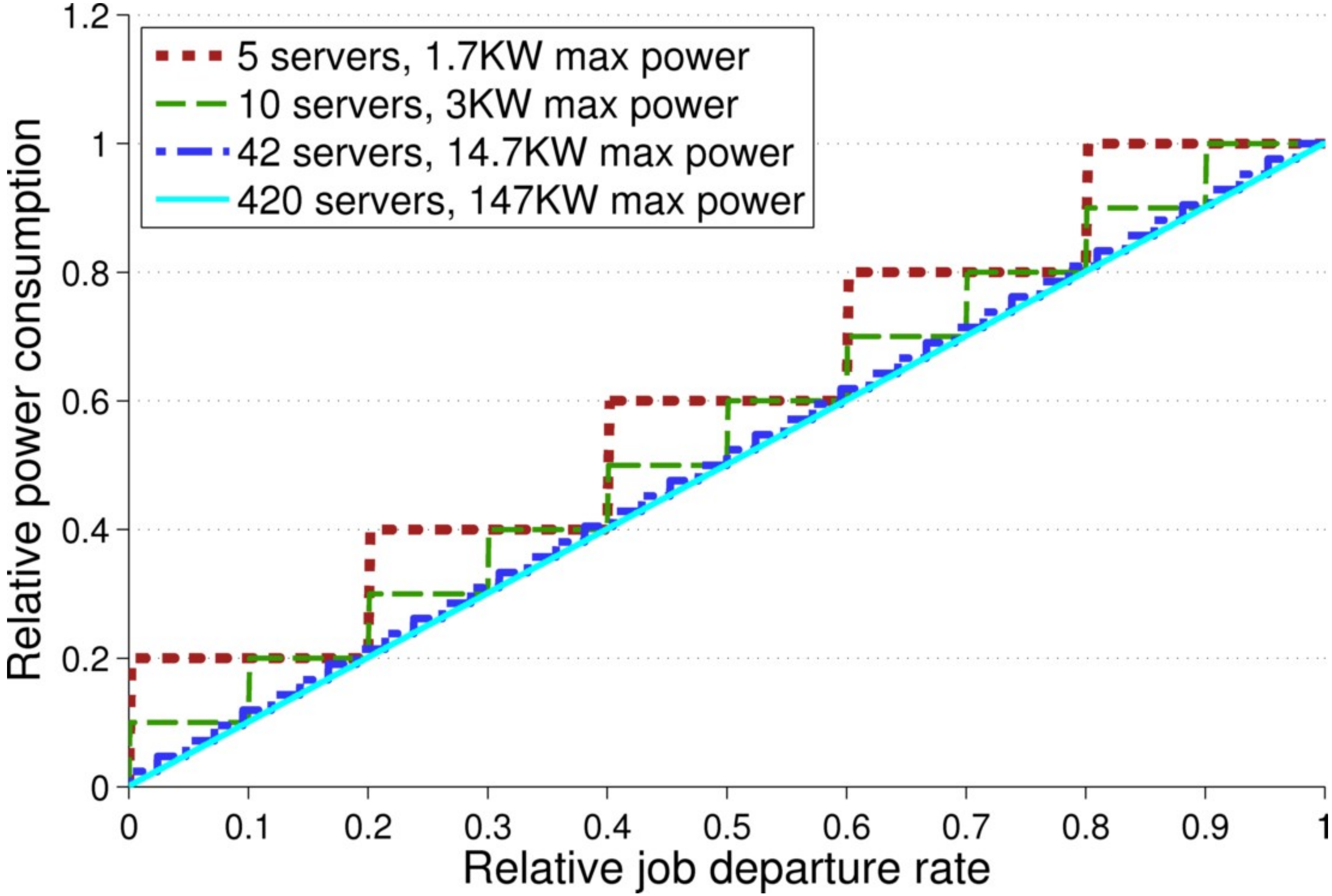
Efficiency of CRAC units



J. Moore et al. "Making scheduling ”cool”: temperature-aware workload placement in data centers." 2005

Ensemble power consumption

- Single server max power consumption: *350 W*



Examples of SLAs

- **Amazon Elastic Compute Cloud (EC2)**
  - “[...] If the Annual Uptime Percentage for a customer drops below 99.95% for the Service Year, that customer is eligible to receive a Service Credit equal to 10% of their bill [...]”
- **Google 500ms additional delay**
  - +20% reduction in search
- **Amazon 100ms additional delay**
  - +1% loss in sales

http://aws.amazon.com/ec2-sla/

Examples of SLAs

- **Amazon Elastic Compute Cloud (EC2)**
  - “[...] If the Annual Uptime Percentage for a customer drops below 99.95% for the Service Year, that customer is eligible to receive a Service Credit equal to 10% of their bill [...]”
- **Google 500ms additional delay**
  - +20% reduction in search
- **Amazon 100ms additional delay**
  - +1% loss in sales

http://aws.amazon.com/ec2-sla/

Issues in controlling a data center

- Large-scale nonlinear system
- Time scale of the controlled processes
- Cyber-physical coupling
- Lack of a widely accepted model

Issues in controlling a data center

- **Large-scale nonlinear system**
    - Example of nonlinearities
        - Response time with respect to the amount of available hardware resources
        - Server idle power consumption ~60 % of peak power
        - Cooling efficiency increases with temperature
        - Power conversion loss decreases with power consumption
        - Fan power consumption increases cubically with the speed of the fan
    - Example of the system scale
        - Control the voltage and frequency of every CPU (e.g., 4 CPU per server, 50k servers → 200k CPUs)
        - On/Off state of every server
        - Speed of server and CRAC fan (100k fans)
        - Performance indicators, (e.g., 2 per server → ~100k)

Issues in controlling a data center

- **Time scale of the controlled processes**
  - Server response time to HTTP request ~100ms
  - Server turn-on time ~5min
  - Data center thermal dynamic in the hour time scale

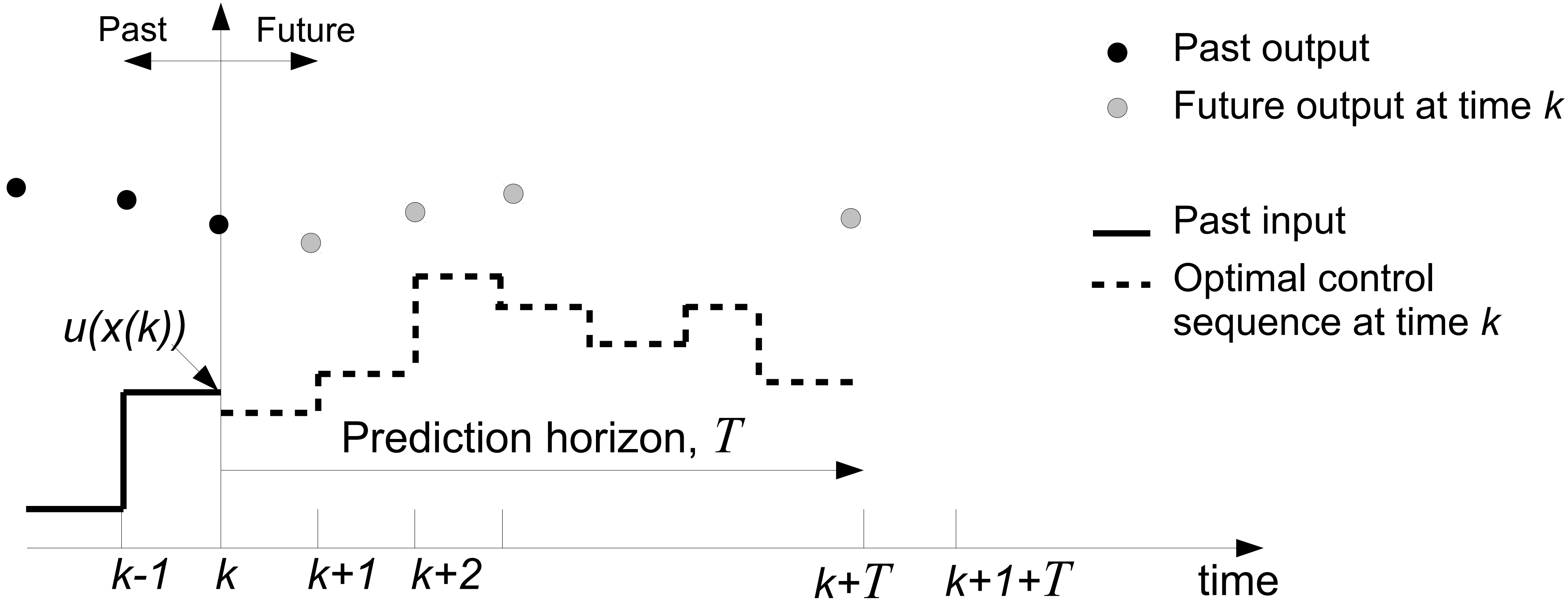
Issues in controlling a data center

- **Time scale of the controlled processes**
  - Server response time to HTTP request ~100ms
  - Server turn-on time ~5min
  - Data center thermal dynamic in the hour time scale
- **Cyber-physical coupling**
  - The way workload is executed and where it is executed affects the cooling power consumption
  - The capacities of the cooling and of the power distribution subsystem set limits to the computational subsystem

Brief introduction to Model Predictive Control (MPC)

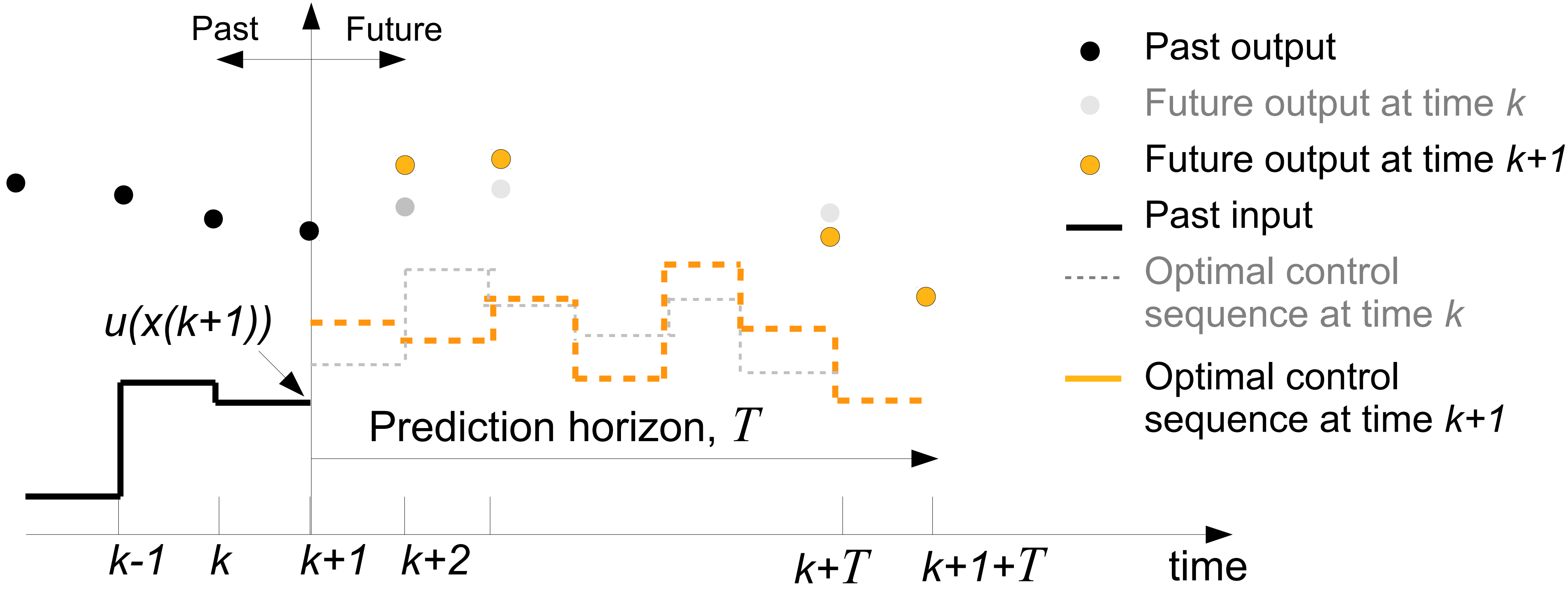
Model predictive control (MPC)

- The controller uses an *internal model of the plant* to predict future states and outputs of the plant
- *The optimal control sequence may change at every step*



Model predictive control (MPC)

- The controller uses an *internal model of the plant* to predict future states and outputs of the plant
- *The optimal control sequence may change at every step*



Thermal and computational network

Thermal network

- **Linear model**

Output temperature of zones and CRAC units

$$\hat{\boldsymbol{T}}_{\text{out}}(\nu+1|k) = A_{T,D}\hat{\boldsymbol{T}}_{\text{out}}(\nu|k) + B_{T,D}\left[\hat{\mathbf{p}}_{\mathcal{N}}(\nu|k)^T \; \hat{\boldsymbol{T}}_{\text{ref}}(\nu|k)^T\right]^T$$

Power consumption of zones

Reference temperature of CRAC units

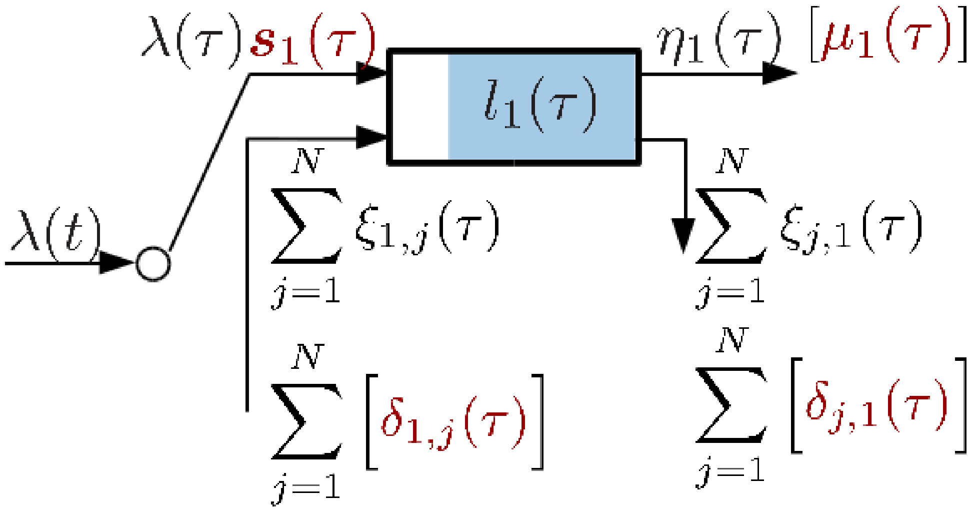
$$\hat{\boldsymbol{T}}_{\text{in}}(\nu|k) = \Psi\hat{\boldsymbol{T}}_{\text{out}}(\nu|k) \leq \overline{\boldsymbol{T}_{\text{in}}}$$

- **CRAC power consumption depends on the coefficient of performance (COP)**

$$\mathsf{p}_i(t) = \frac{\dot{Q}_i(t)}{COP_i(T_{\text{out},i}(t))}$$

Heat removed rate (W)

Computational network



$$l(\tau) = a_1(\tau) - d_1(\tau)$$

$$d_1(\tau) = \eta_1(\tau) + \sum_{j} \xi_{j,i}(\tau)$$

$$a_1(\tau) = \lambda(\tau)\boldsymbol{s}_1(\tau) + \sum_{j=1}^{N} \xi_{i,j}(\tau)$$

$$\eta_1(\tau) = \begin{cases} \boldsymbol{\mu}_1(\tau) & \text{if } l_1(\tau) > 0 \\ & \text{or } a_1(\tau) > \boldsymbol{\mu}_1(\tau) \\ a_1(\tau) & \text{otherwise} \end{cases}$$

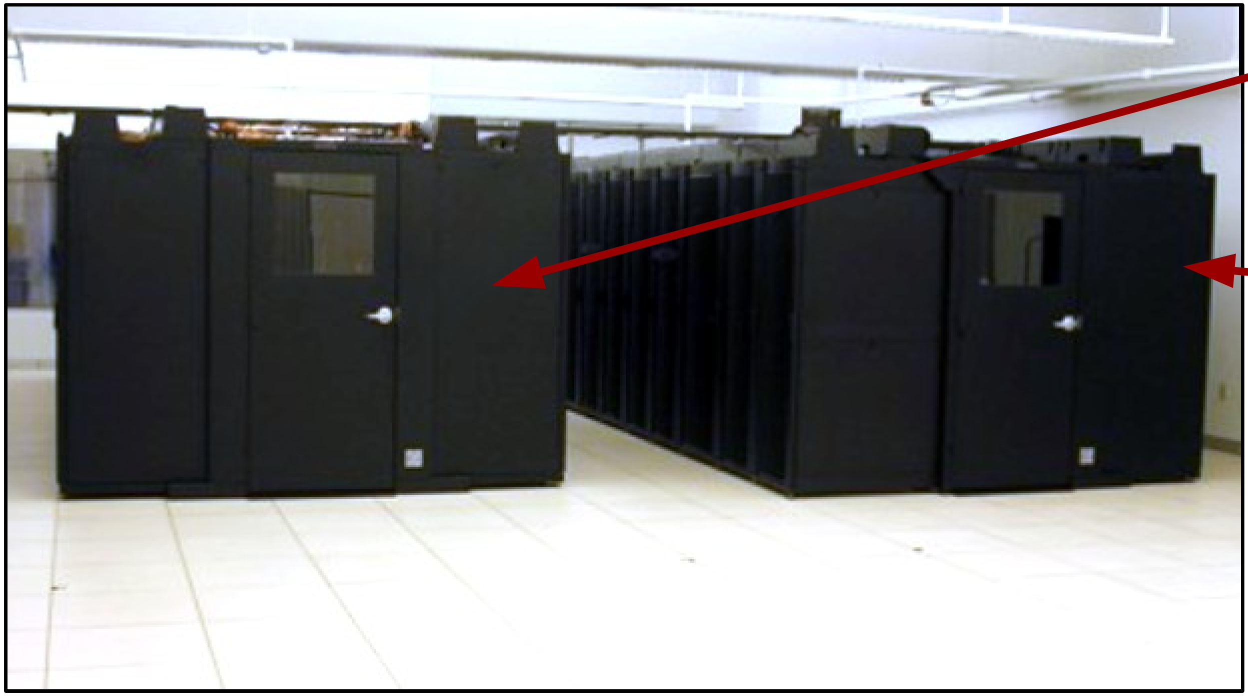
$$\nu_1(\tau) = \mu_1(\tau) + \sum_{j=1}^{N} \delta_{j,1}(\tau)$$

$$\xi_{j,1}(\tau) = \begin{cases} \delta_{j,1}(\tau) & \text{if } l_1(\tau) > 0 \\ & \text{or } a_1(\tau) > \nu_1(\tau) \\ \dfrac{\delta_{j,1}(\tau)}{\sum_{j=1}^{N} \delta_{j,1}}(a_1(\tau) - \nu_1(\tau)) & \text{otherwise} \end{cases}$$

Data collection in the
Data Center Observatory (DCO)
Carnegie Mellon University
Pittsburgh, Pa

Data collection in the DCO

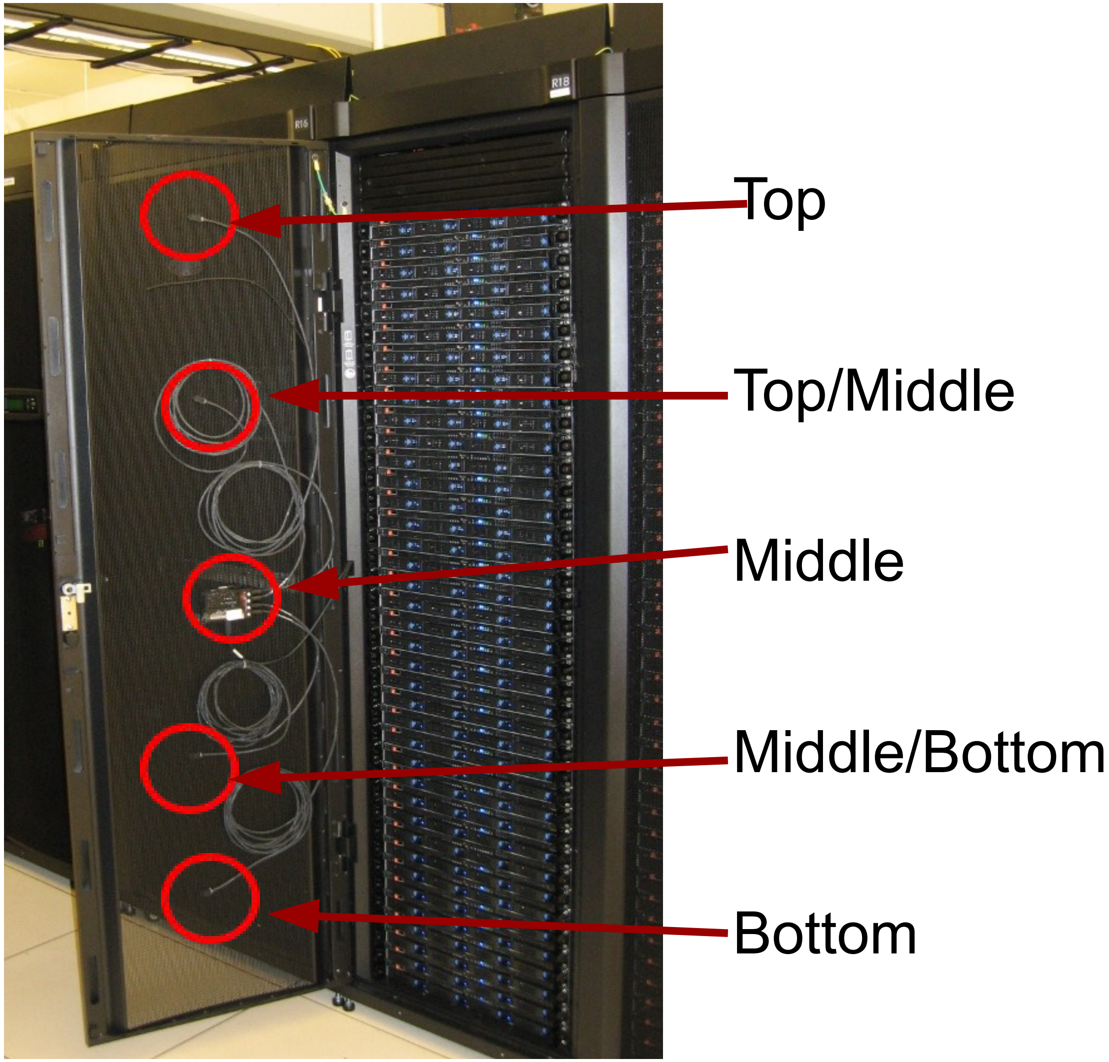
- **Data center observatory (DCO)**
  - Testbed for research on data-intensive workload



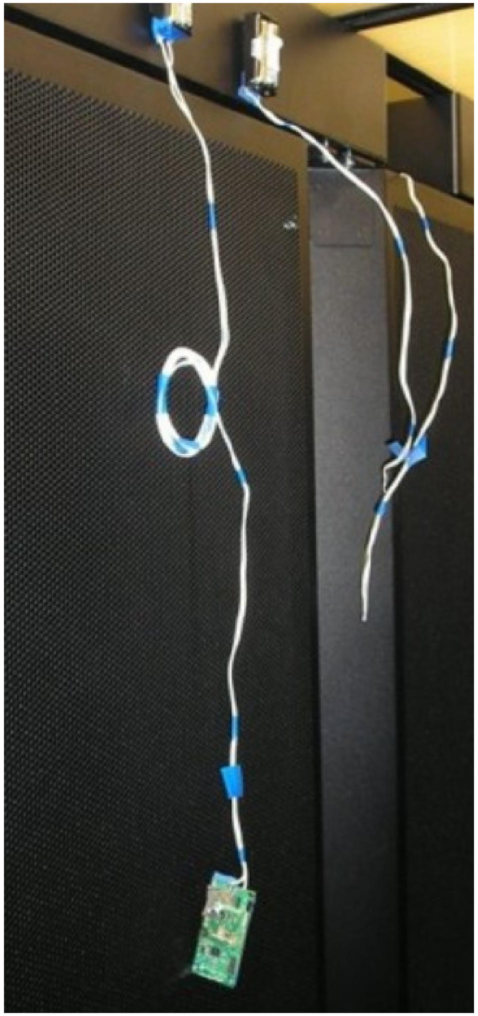
- Zone 1
  - 325 computers
- Zone 2

Temperature and humidity sensors

- Wired sensors

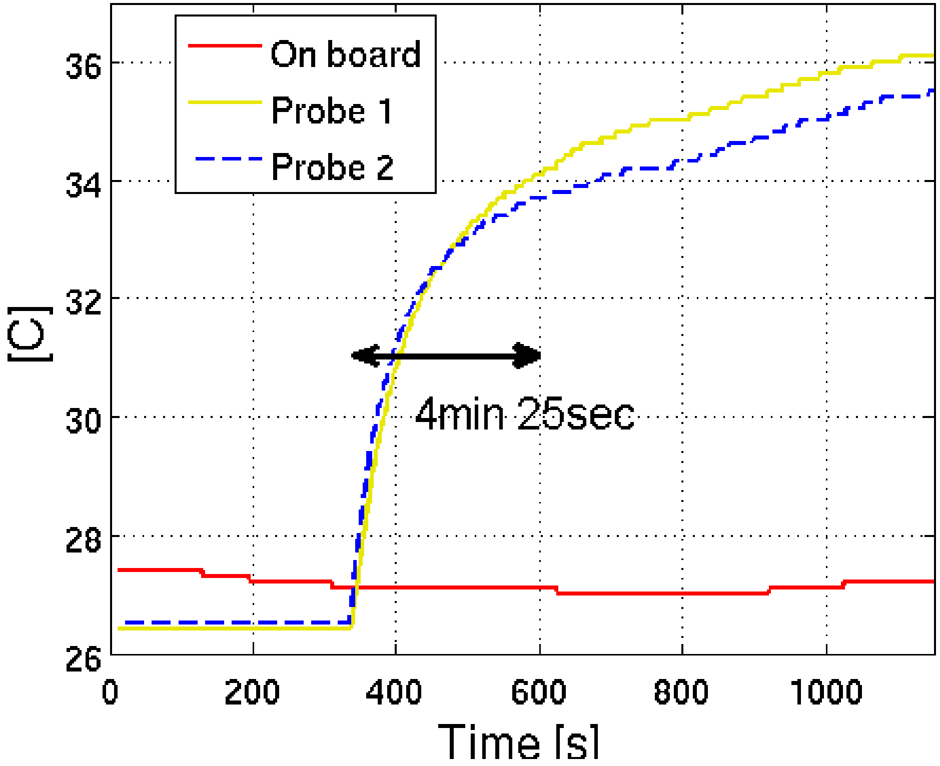


- Wireless sensors



Wired sensors

- **APC NetBotz sensor**
- **Time constant ~53 sec**
  - Wireless sensors have time constant ~13sec



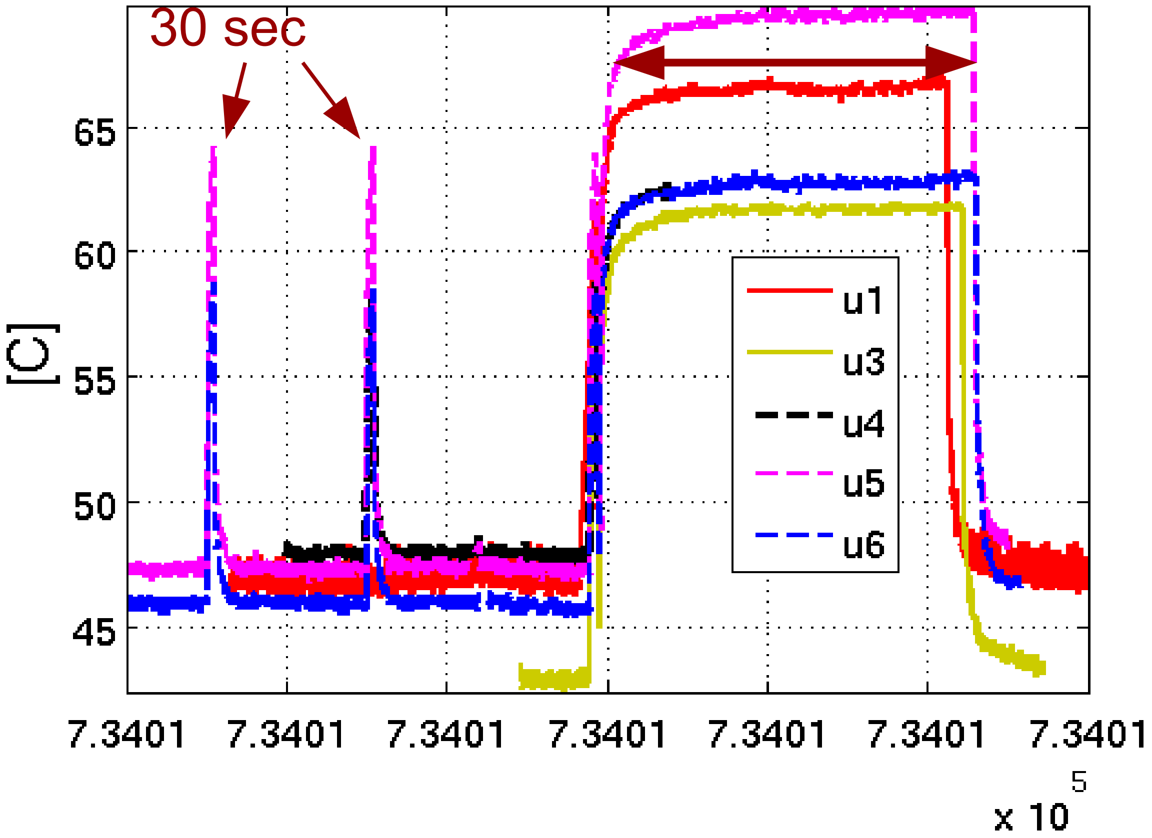
Thermal test on 6 servers

- **Time varying workload**
  - Measure server power consumption
  - CPU temperature (on-board sensors)
  - Inlet and outlet temperatures

CPU temperature

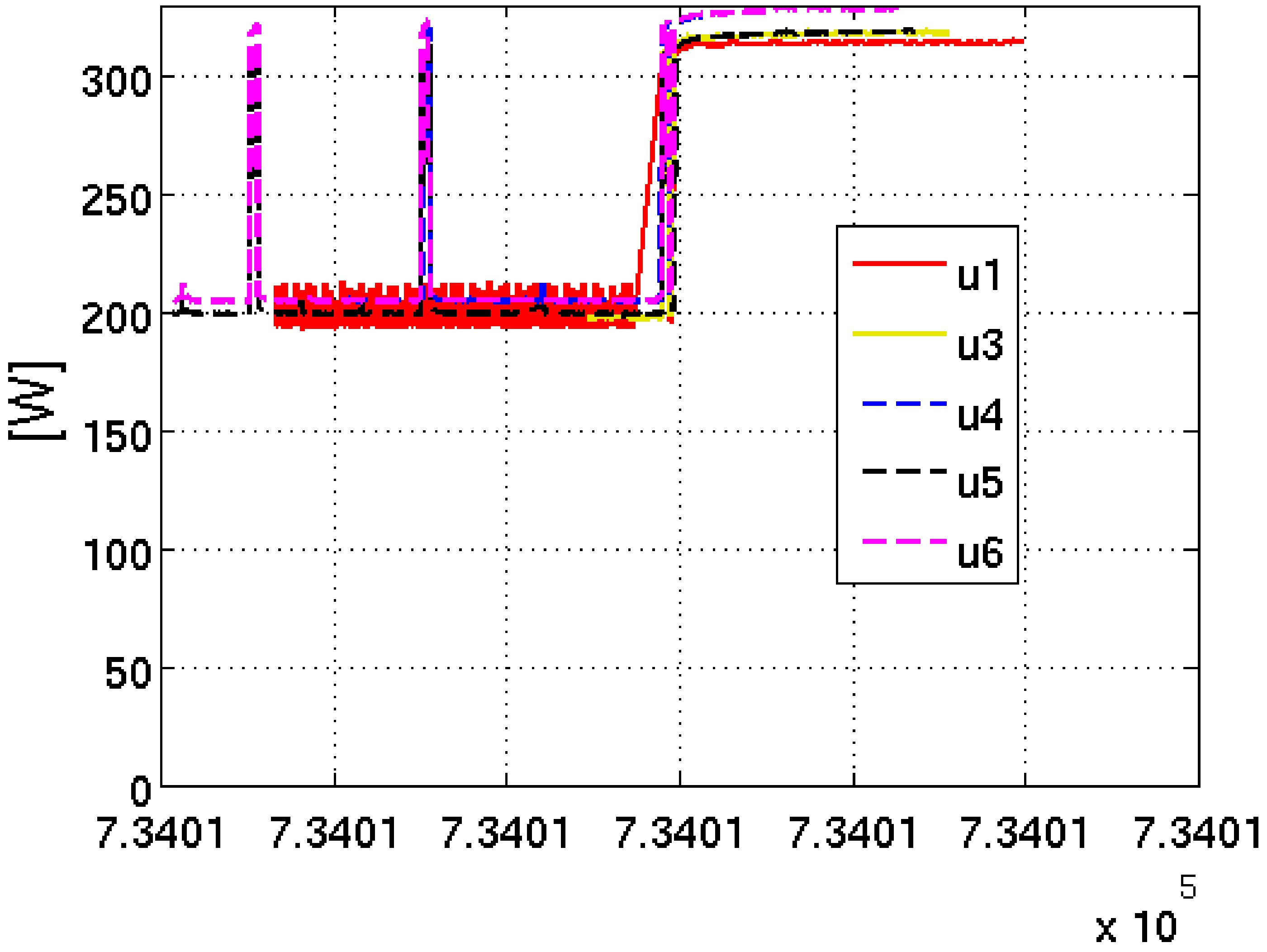
- **8 Cores per CPU**
  - We consider the average value of the 8 cores

Peak values

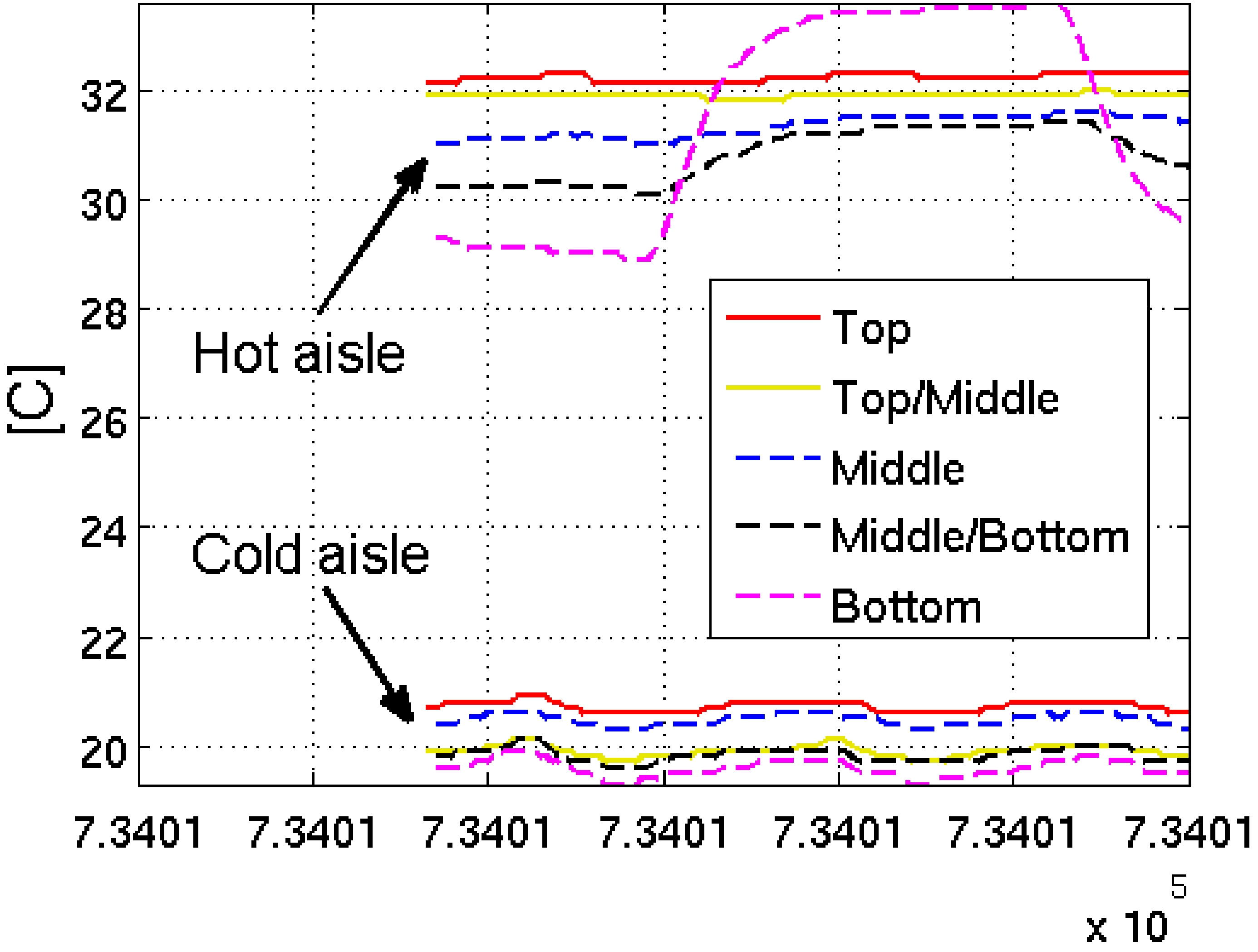


- Test lasts for 34 min

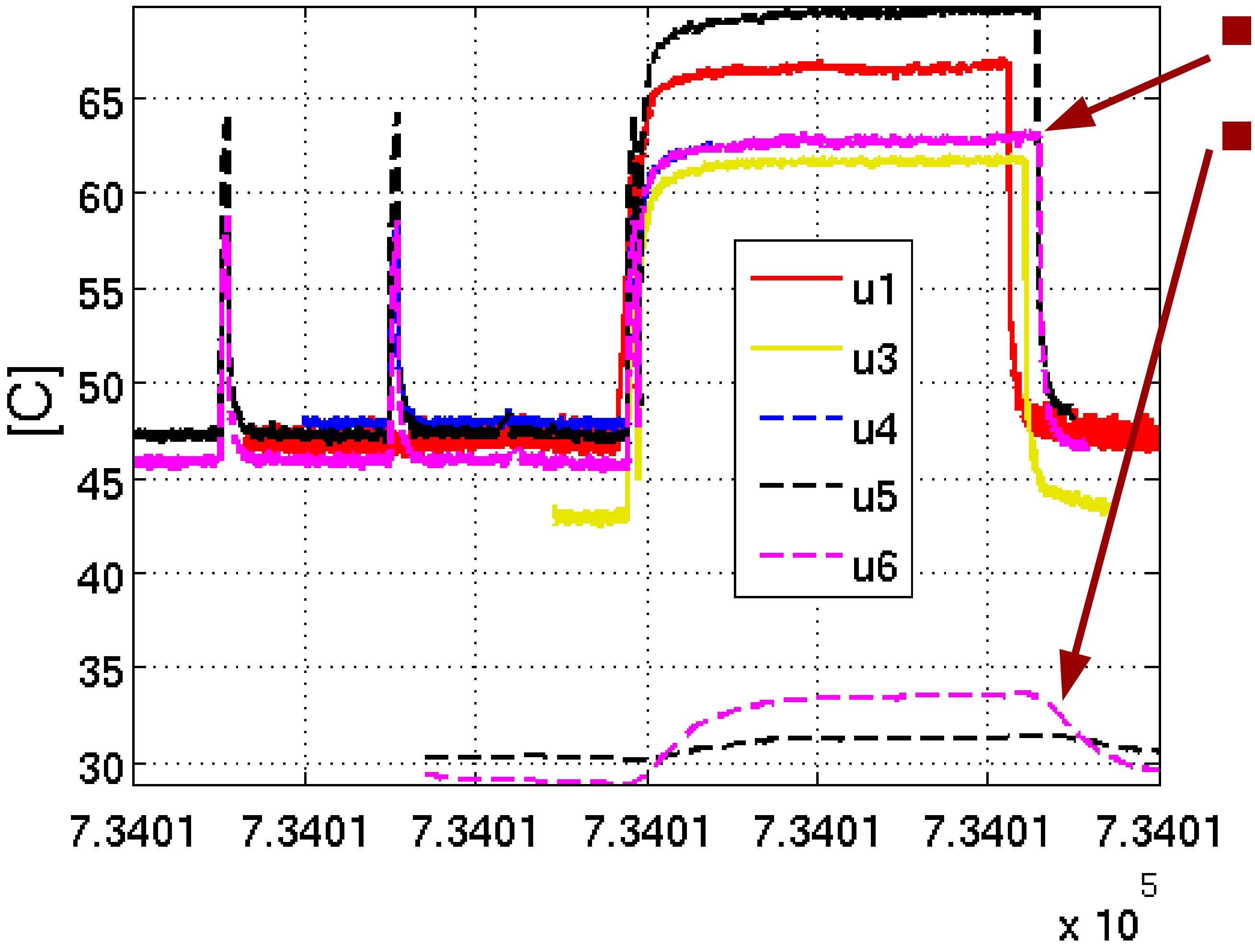
Server power consumption



Inlet and outlet temperatures



Outlet and CPU temperatures



- CPU temperature
- Outlet temperature

Estimation of server power consumption

Joint work with
Anshul Gandhi and Kevin Woo

Servers

- **Dave's cn007**
    - 8 core Intel Xeon X5355 2.67 GHz
- **PH server**
    - 2 core Intel Xeon 3.00 GHz

Tools

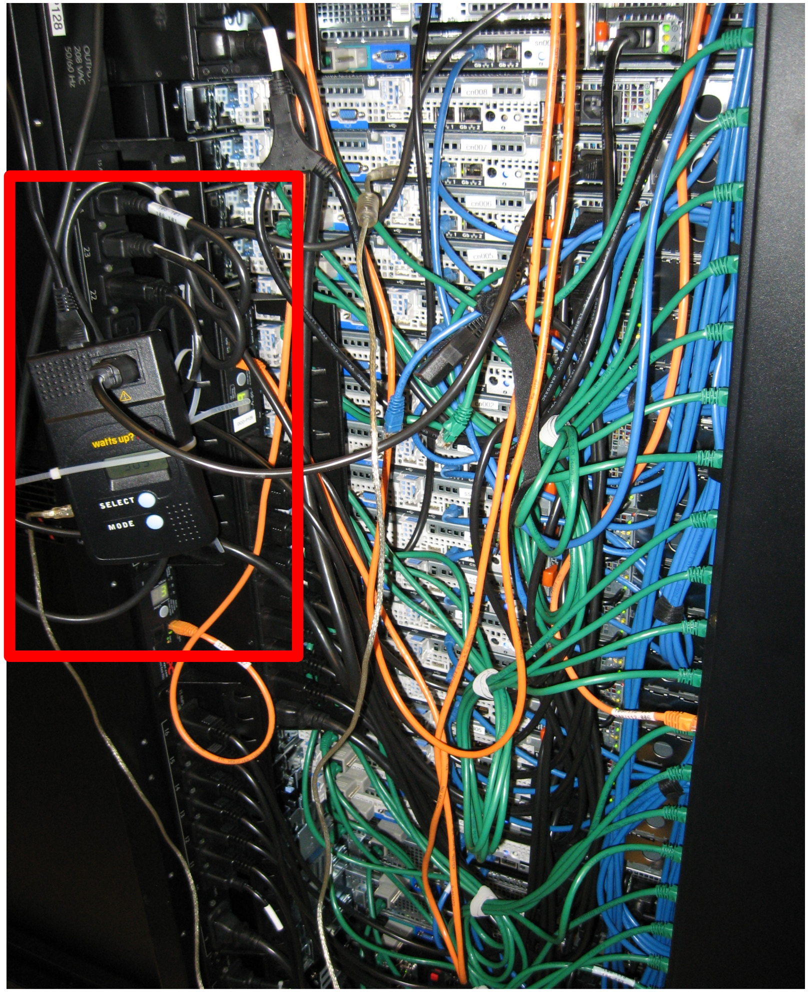
- **Software**
  - SAR
    - CPU, network, disk
  - Perfmon
    - Cache misses
  - Powertop
    - P-states and C-states
- **Hardware**
  - Watts-Up Pro
    - Power consumption

Perfmon

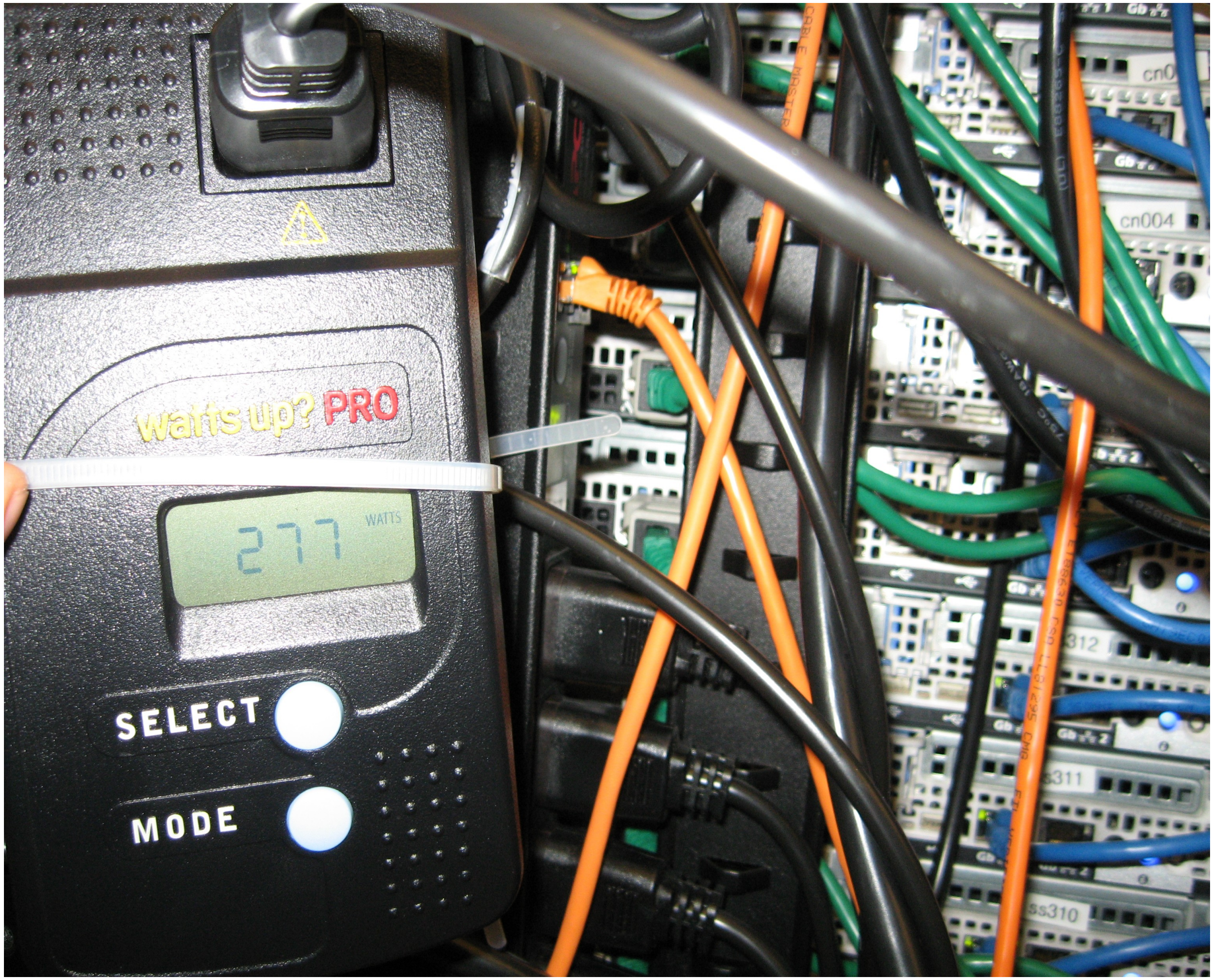
- **Tool developed by HP**
  - Exposes hardware performance counters
- **Composed of 3 components**
  - Kernel Patch
  - Kernel Module
  - User Interface
- **Requires kernel recompilation**
  - Unstable
- **Not tried on cn007, tested on PH**

cn007

Watts-Up Pro Installation

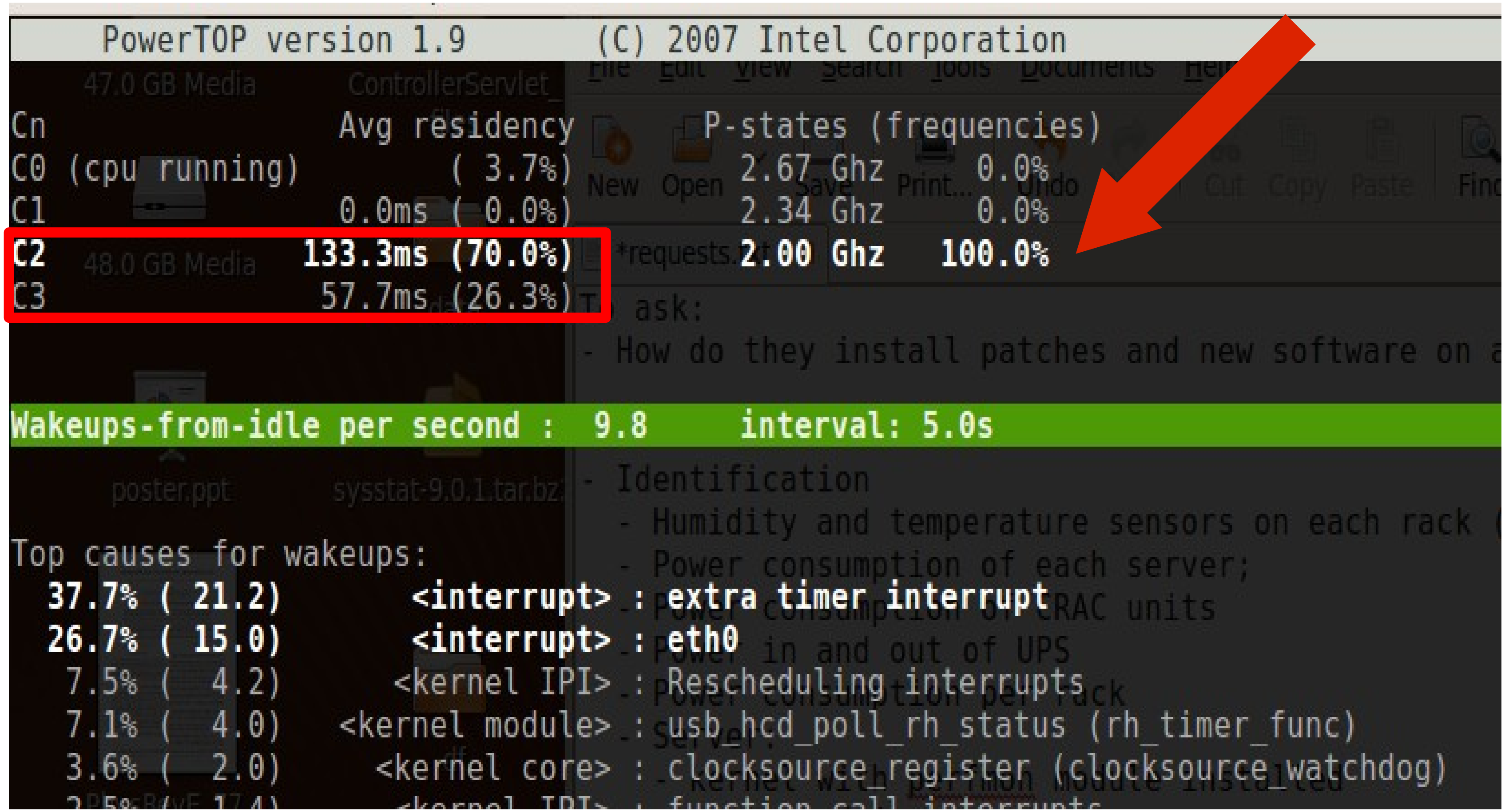


Watts-Up Pro Monitoring



Script Overhead

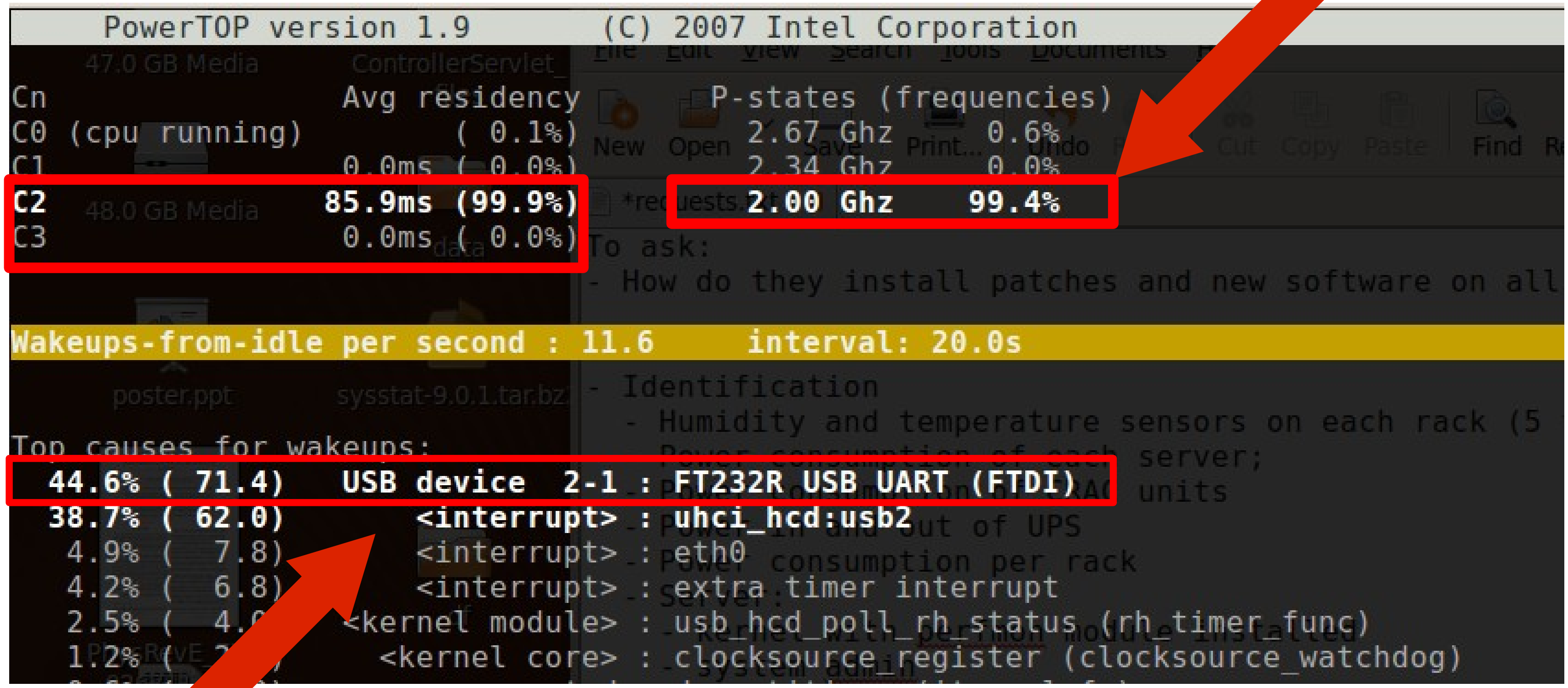
Idle System



```
     PowerTOP version 1.9       (C) 2007 Intel Corporation

Cn                Avg residency       P-states (frequencies)
C0 (cpu running)        ( 3.7%)         2.67 Ghz     0.0%
C1                  0.0ms ( 0.0%)       2.34 Ghz     0.0%
C2                133.3ms (70.0%)       2.00 Ghz   100.0%
C3                 57.7ms (26.3%)

Wakeups-from-idle per second :  9.8      interval: 5.0s

Top causes for wakeups:
  37.7% ( 21.2)        <interrupt> : extra timer interrupt
  26.7% ( 15.0)        <interrupt> : eth0
   7.5% (  4.2)       <kernel IPI> : Rescheduling interrupts
   7.1% (  4.0)    <kernel module> : usb_hcd_poll_rh_status (rh_timer_func)
   3.6% (  2.0)      <kernel core> : clocksource_register (clocksource_watchdog)
```

Script running

Low script overhead



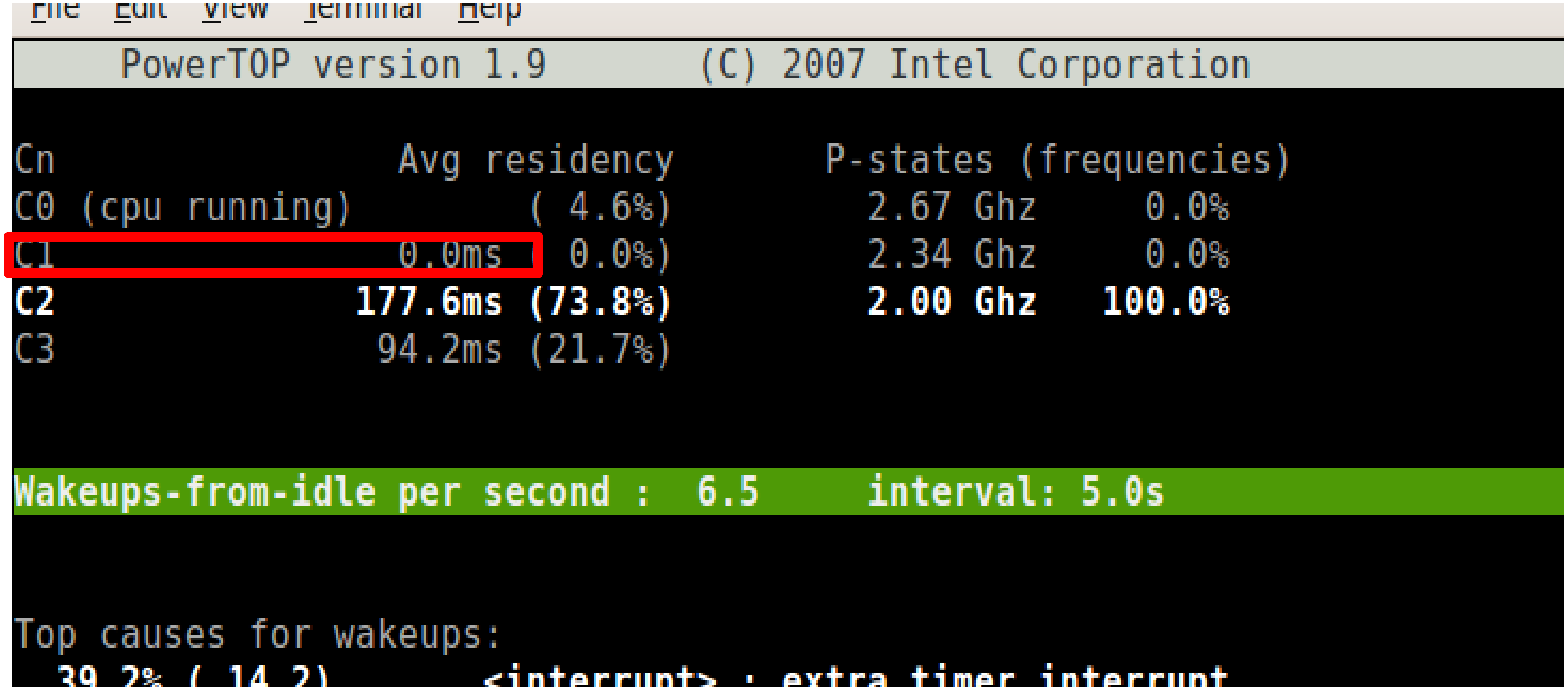
```
     PowerTOP version 1.9       (C) 2007 Intel Corporation

Cn                   Avg residency       P-states (frequencies)
C0 (cpu running)           ( 0.1%)         2.67 Ghz     0.6%
C1                   0.0ms ( 0.0%)         2.34 Ghz     0.0%
C2                  85.9ms (99.9%)         2.00 Ghz    99.4%
C3                   0.0ms ( 0.0%)


Wakeups-from-idle per second : 11.6     interval: 20.0s


Top causes for wakeups:
  44.6% ( 71.4)   USB device  2-1 : FT232R USB UART (FTDI)
  38.7% ( 62.0)       <interrupt> : uhci_hcd:usb2
   4.9% (   7.8)      <interrupt> : eth0
   4.2% (   6.8)      <interrupt> : extra timer interrupt
   2.5% (   4.   <kernel module> : usb_hcd_poll_rh_status (rh_timer_func)
   1.2% (   2.     <kernel core> : clocksource_register (clocksource_watchdog)
```

To ask:
- How do they install patches and new software on all

- Identification
  - Humidity and temperature sensors on each rack (5

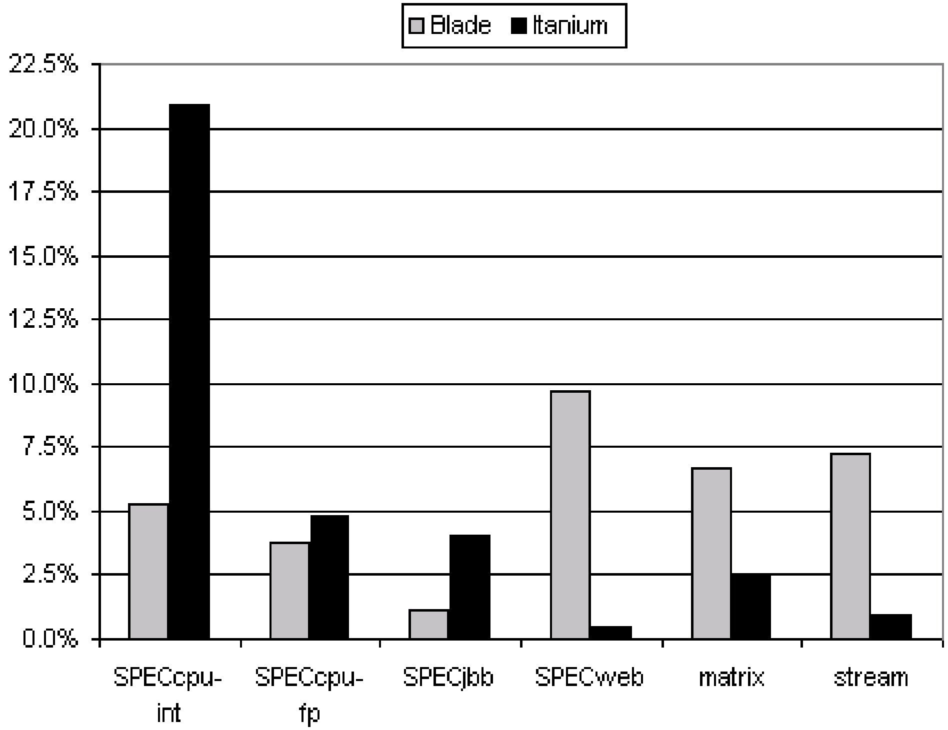
Watts-Up Communication

No Polling



```
     PowerTOP version 1.9        (C) 2007 Intel Corporation

Cn                  Avg residency       P-states (frequencies)
C0 (cpu running)        ( 4.6%)          2.67 Ghz      0.0%
C1                  0.0ms ( 0.0%)          2.34 Ghz      0.0%
C2                177.6ms (73.8%)          2.00 Ghz    100.0%
C3                 94.2ms (21.7%)


Wakeups-from-idle per second :  6.5     interval: 5.0s


Top causes for wakeups:
  39.2% ( 14.2)       <interrupt> : extra timer interrupt
```

Previous work



90th percentile estimation error. [Economou06]

- **Server power consumption is a (affine) linear combination of:**
  - CPU utilization
  - Off-chip memory access
  - HD I/O rate
  - Network I/O rate

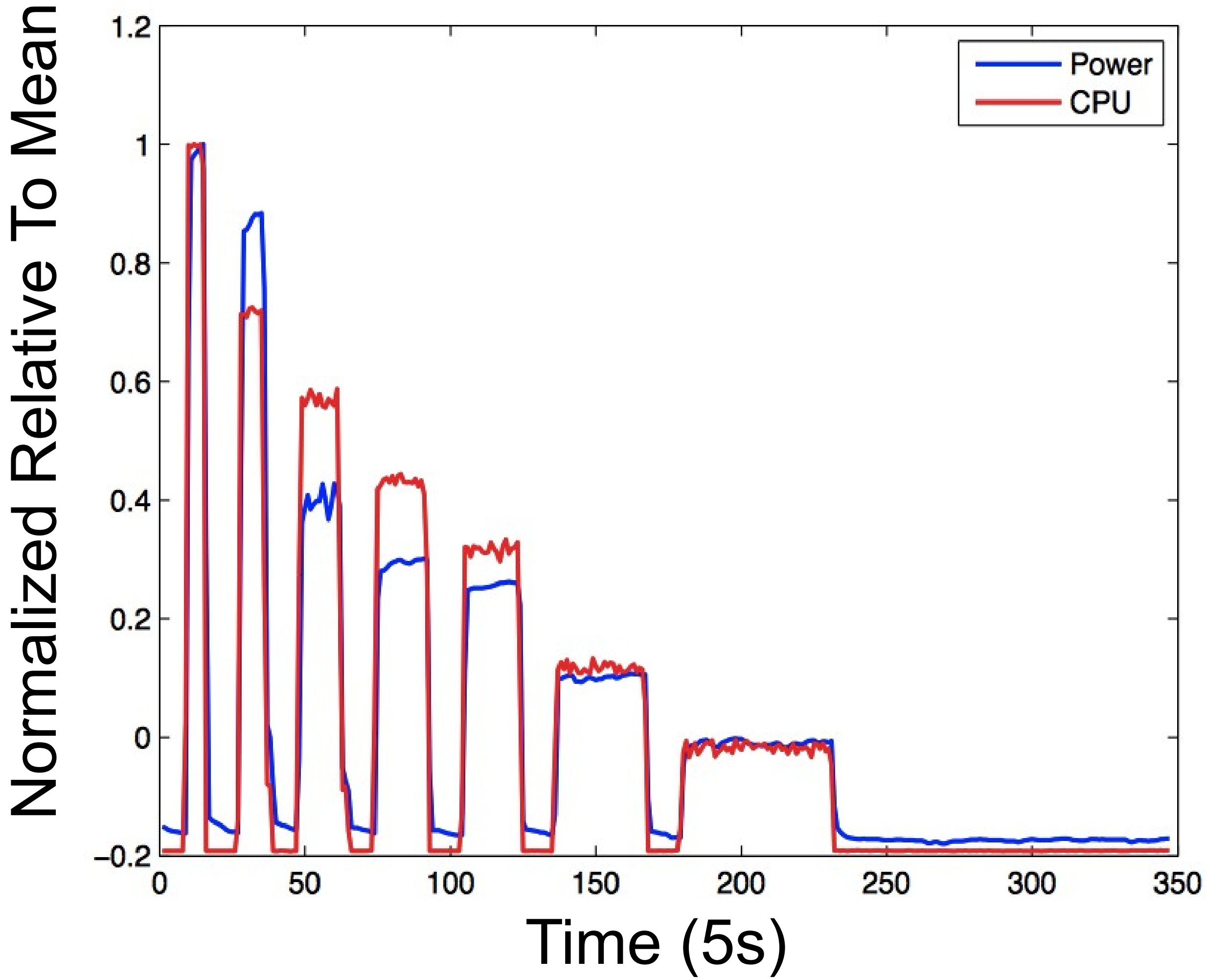
Model

- **Hypothesis**
    - Server power consumption only depends on current data values
    - Server power consumption is a (affine) linear combination of ?

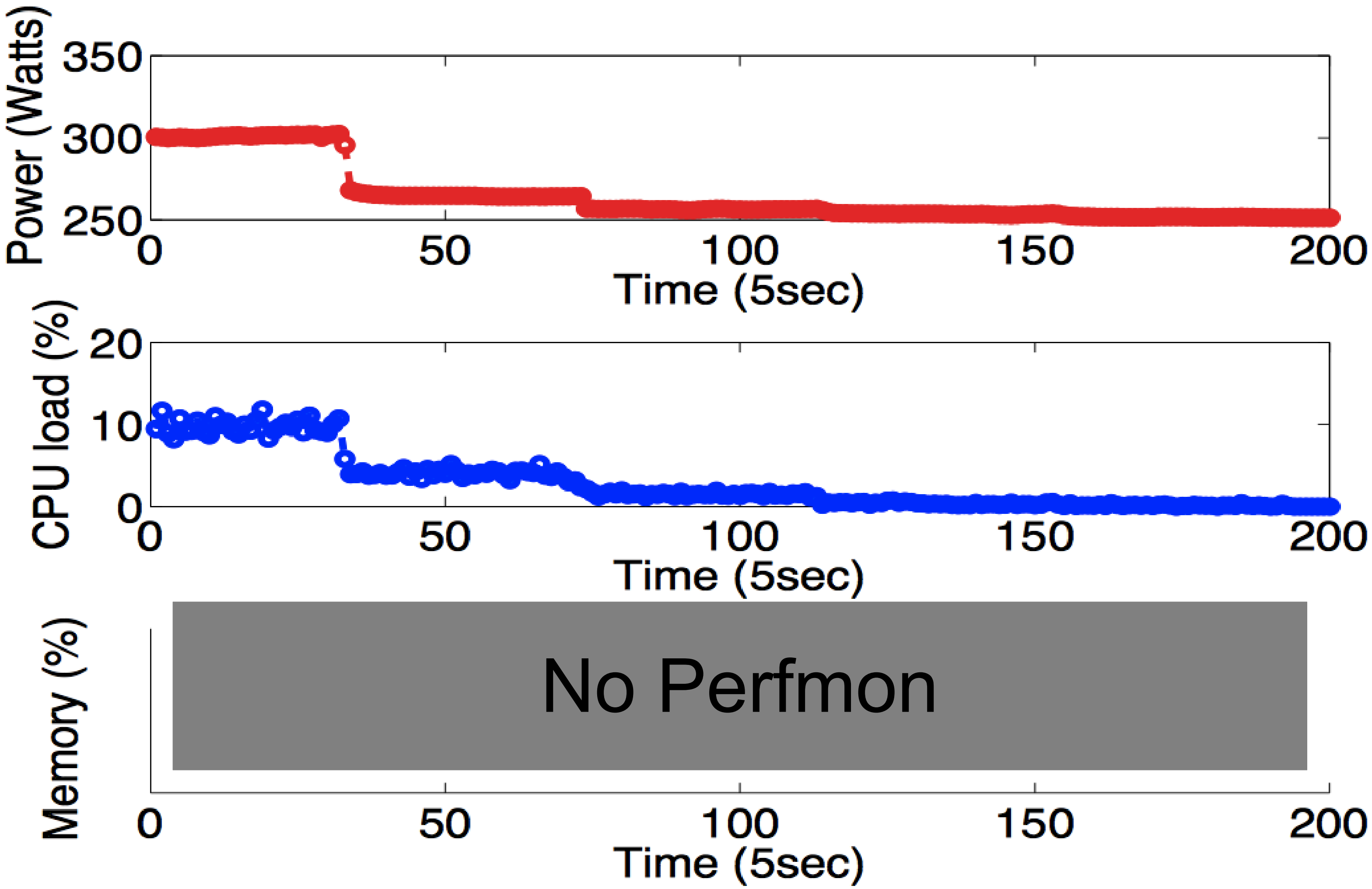
$Ax = b$

$$\begin{bmatrix} ? \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} \text{power} \end{bmatrix}$$

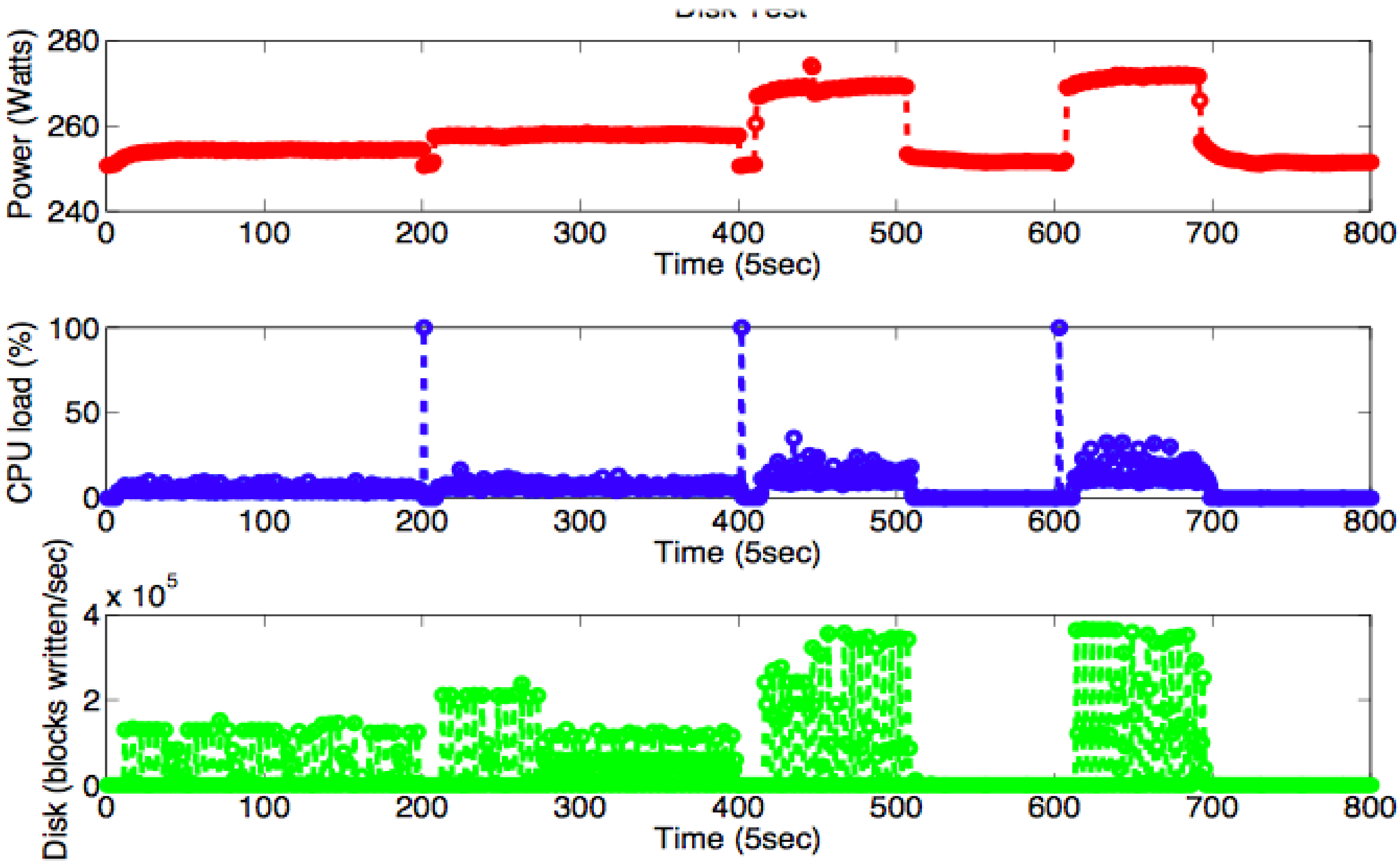
CPU Test


Power
CPU
Normalized Relative To Mean
Time (5s)
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
0
50
100
150
200
250
300
350

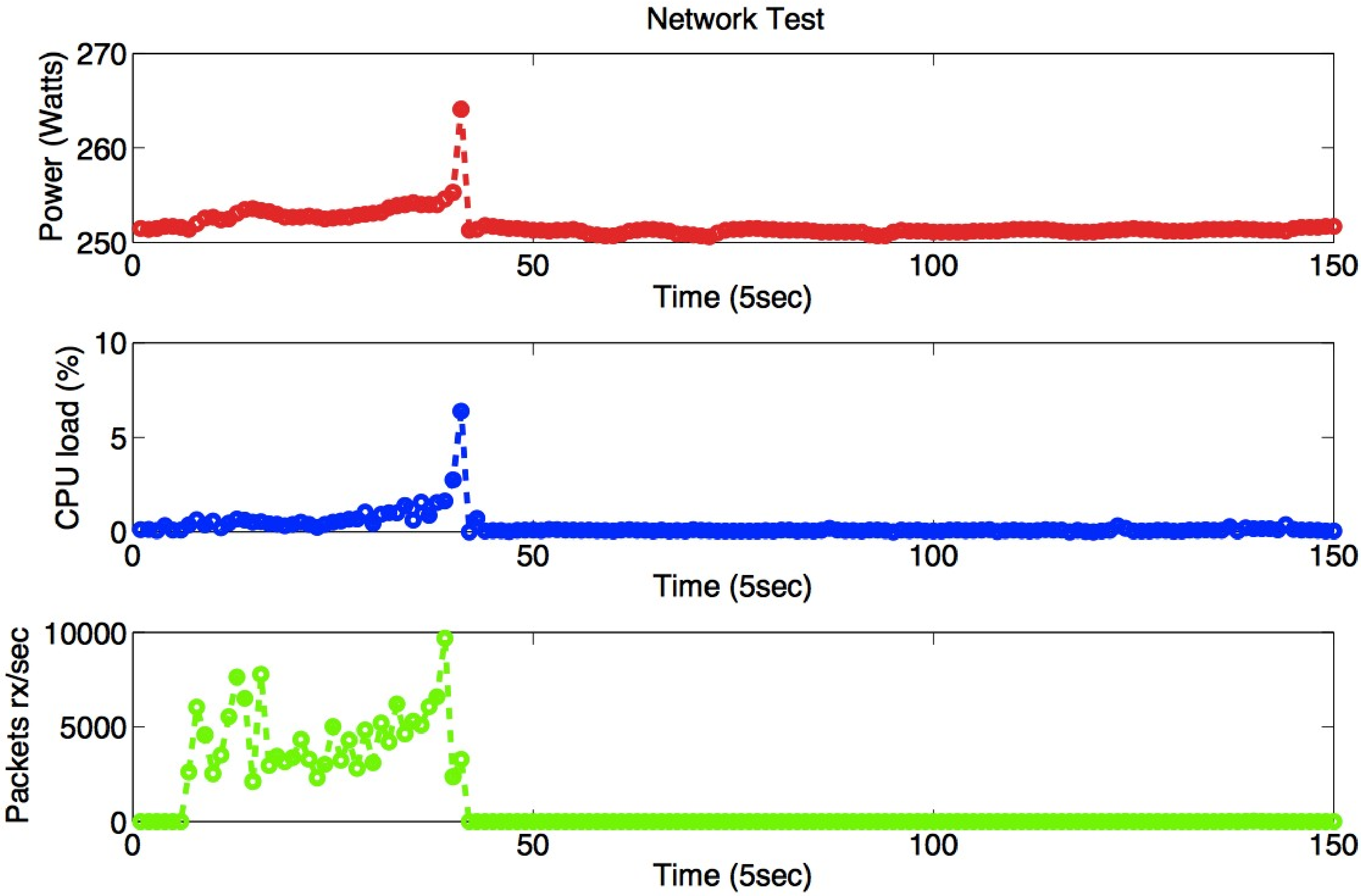
Memory Test


Power (Watts)
350
300
250
0
50
100
150
200
Time (5sec)
CPU load (%)
20
10
0
0
50
100
150
200
Time (5sec)
Memory (%)
No Perfmon
0
50
100
150
200
Time (5sec)

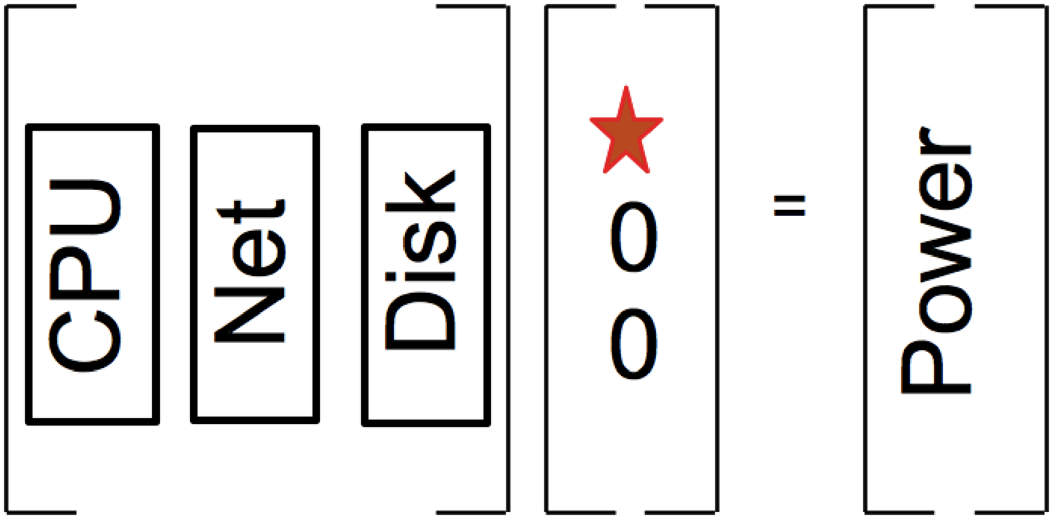
Disk Test


Disk Test
Power (Watts)
280
260
240
0
100
200
300
400
500
600
700
800
Time (5sec)
CPU load (%)
100
50
0
0
100
200
300
400
500
600
700
800
Time (5sec)
Disk (blocks written/sec)
x 10^5
4
2
0
0
100
200
300
400
500
600
700
800
Time (5sec)

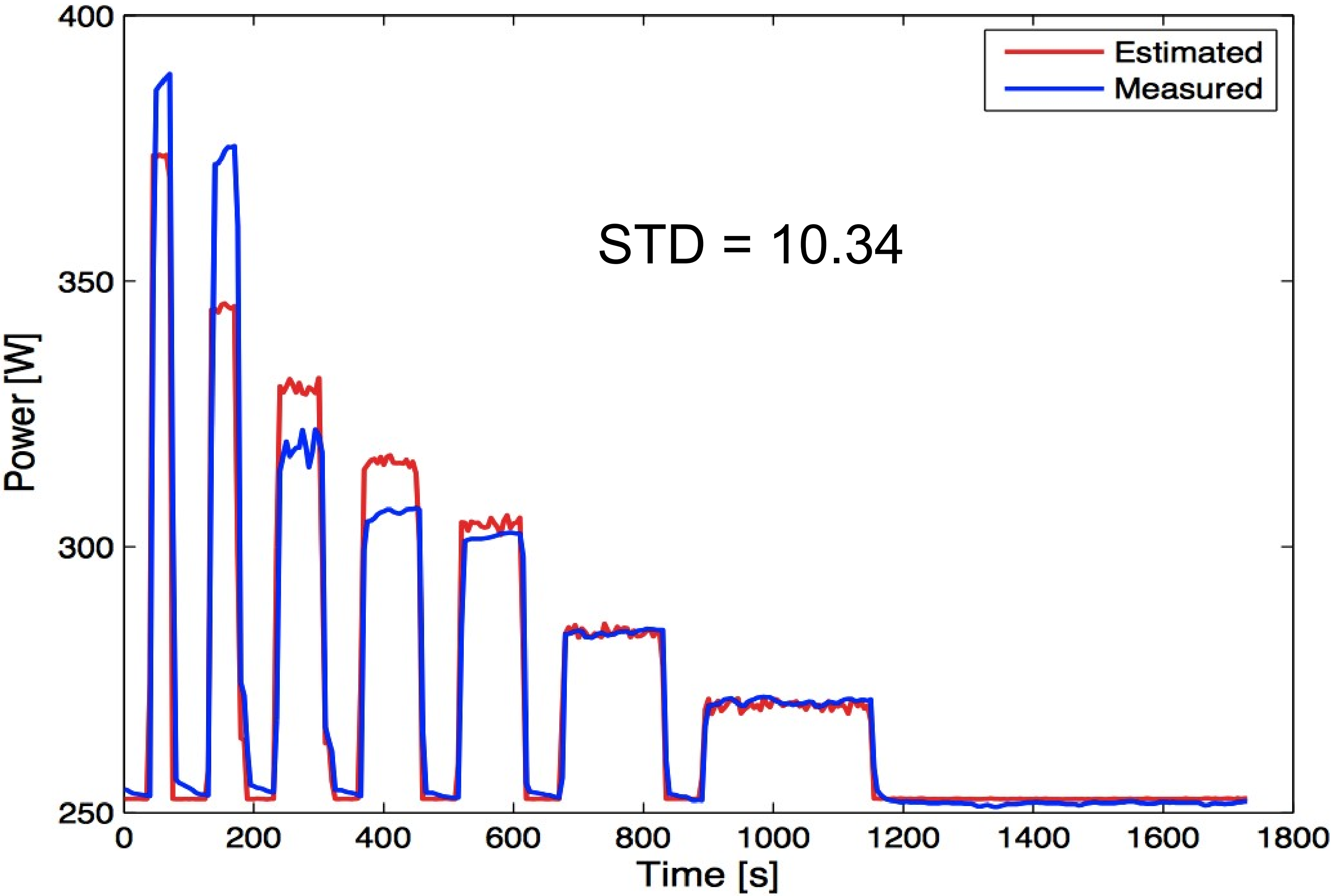
Network Test



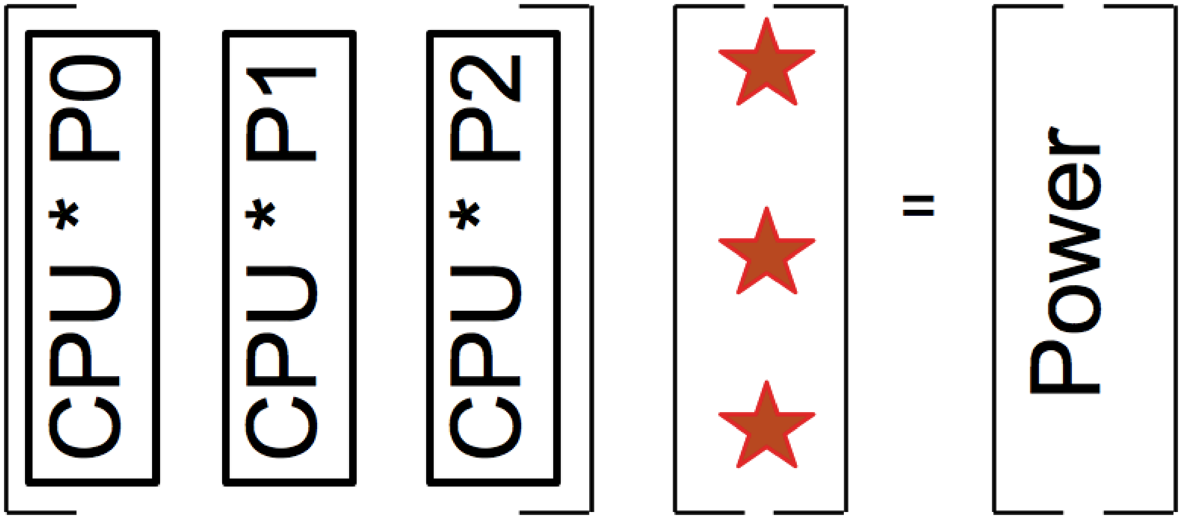
Model Analysis



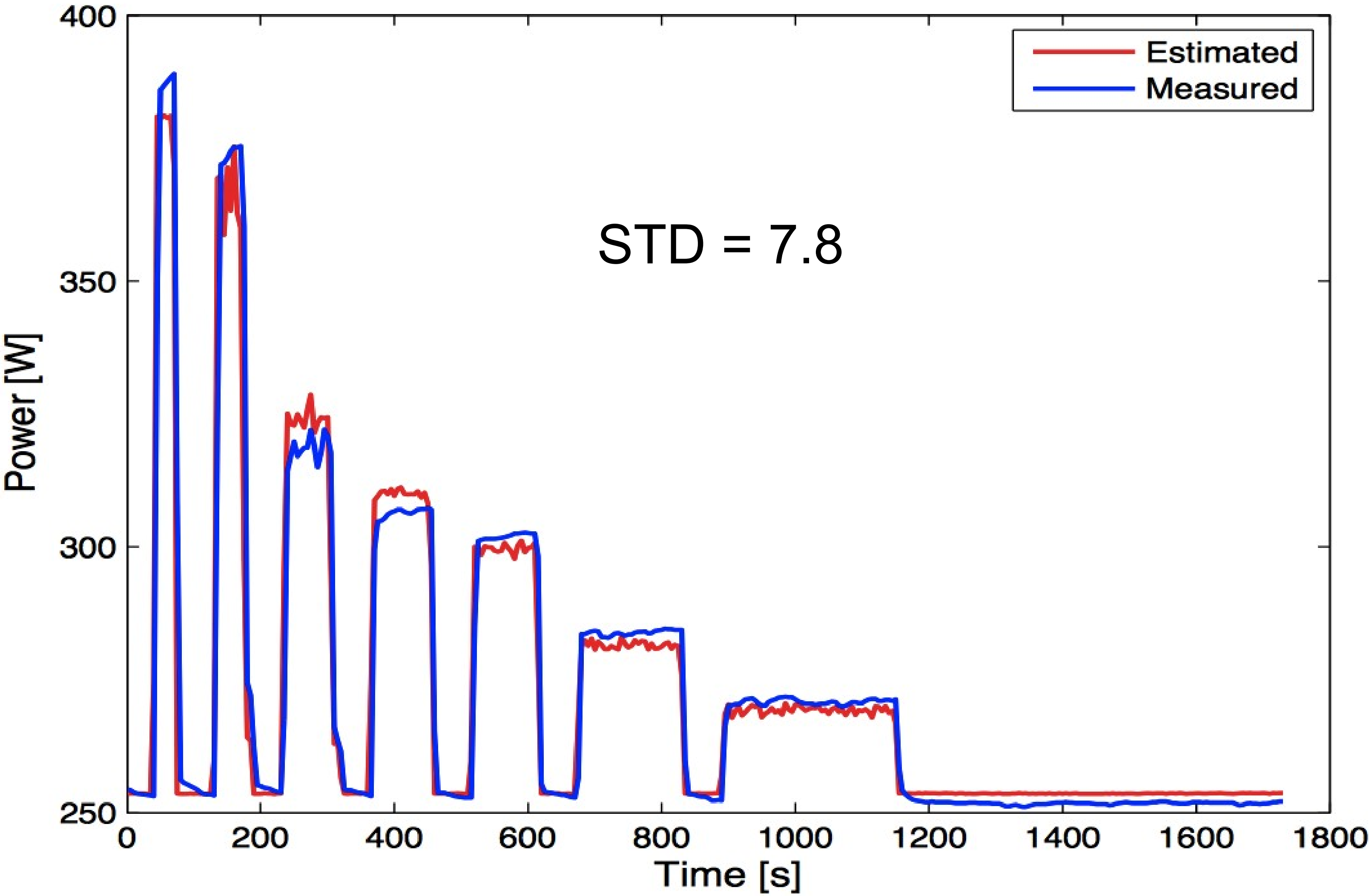
Estimation - CPU data only



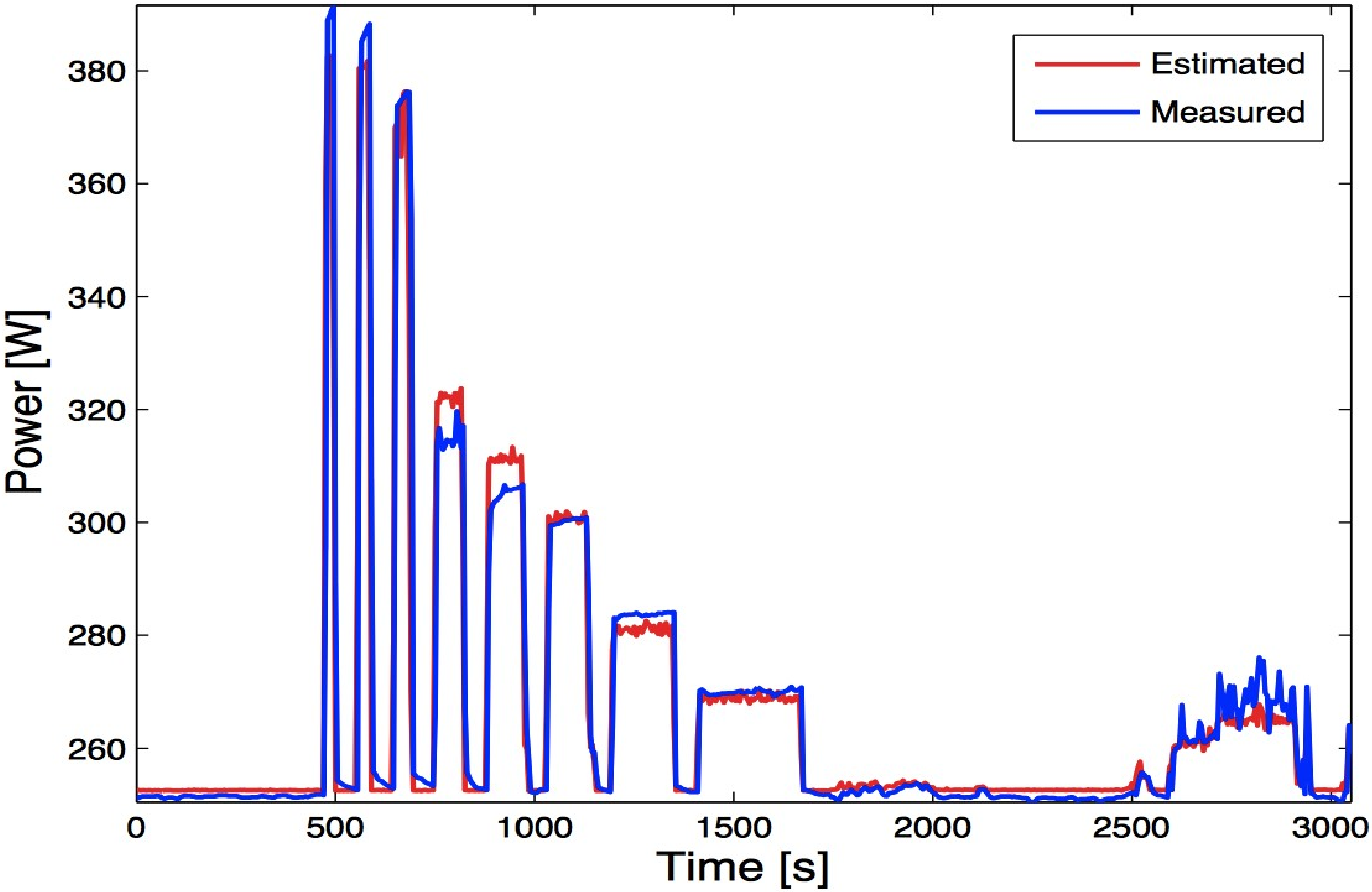
New Model



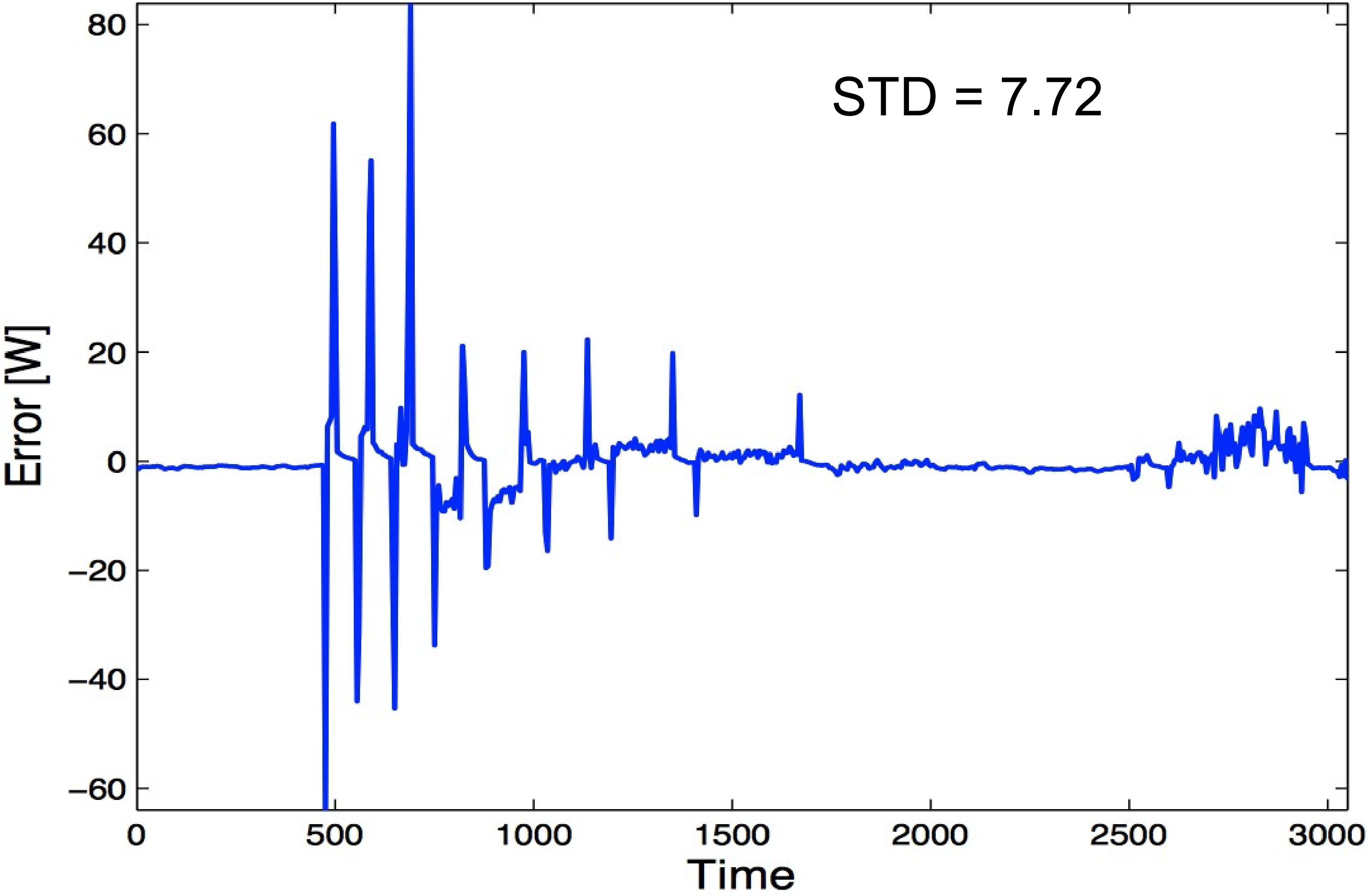
Estimation – P-States



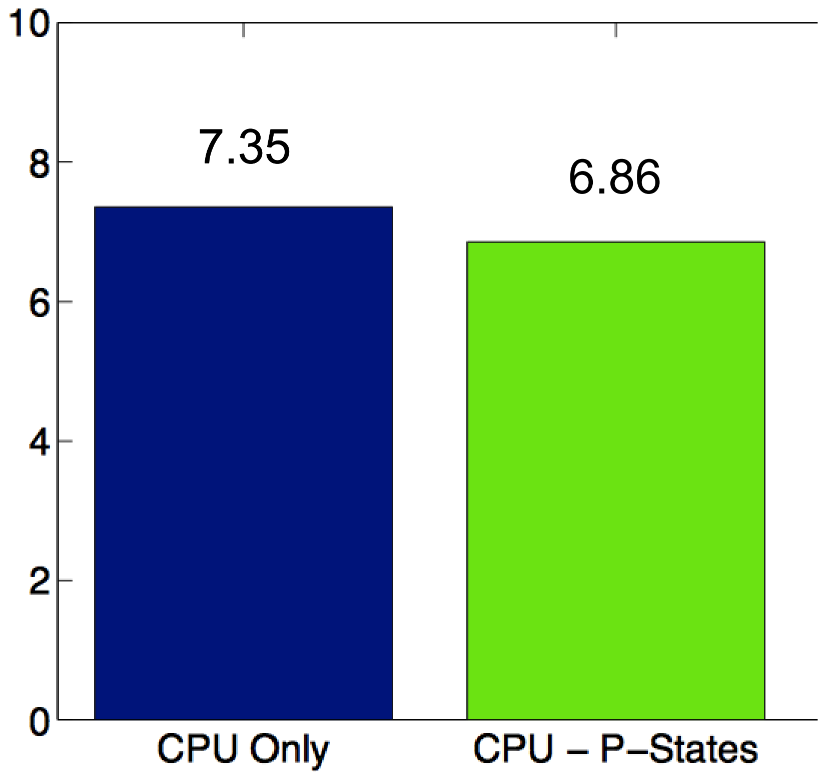
Model Validation



Error Analysis



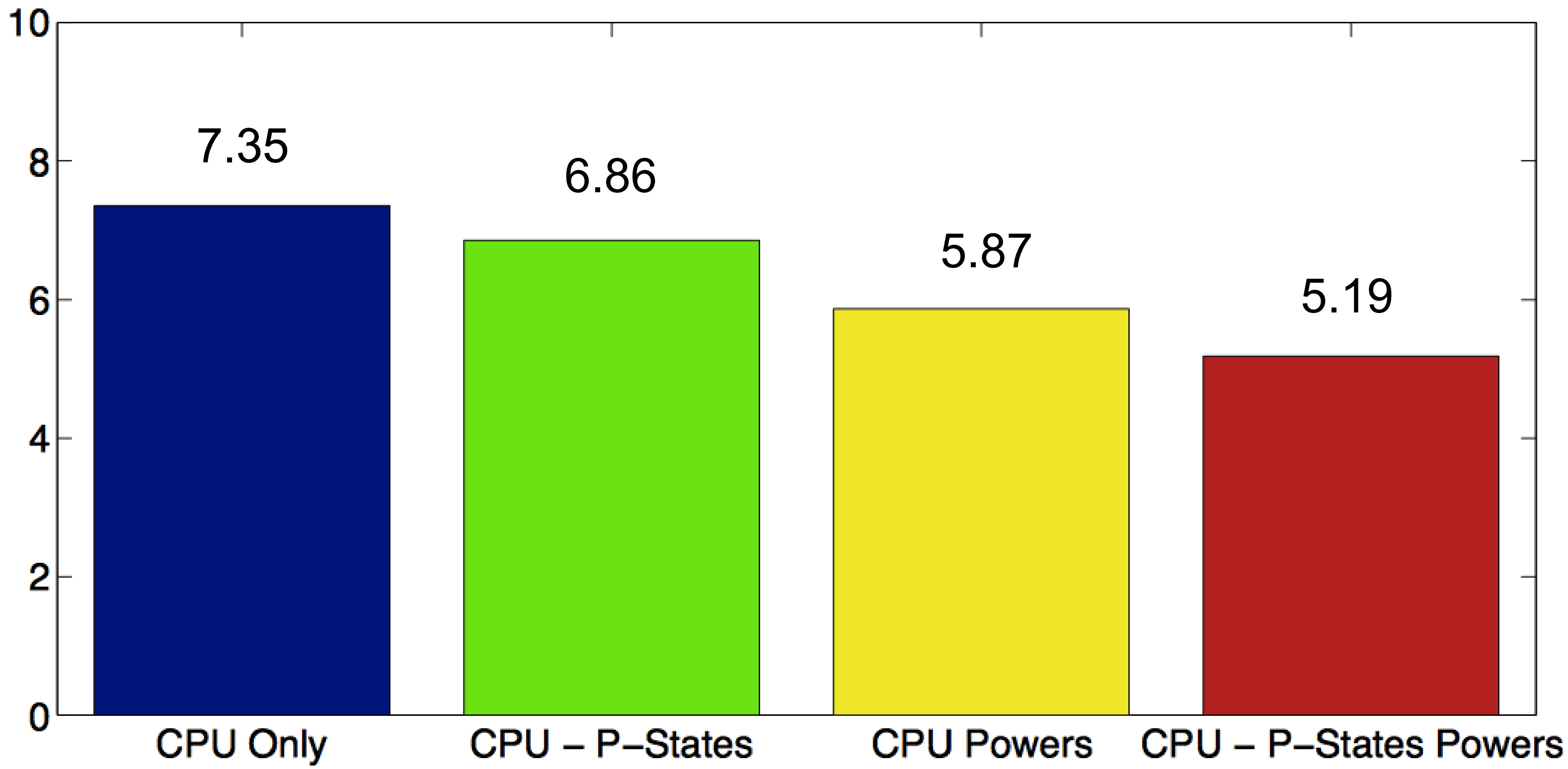
90^{th} Percentile



[CPU]

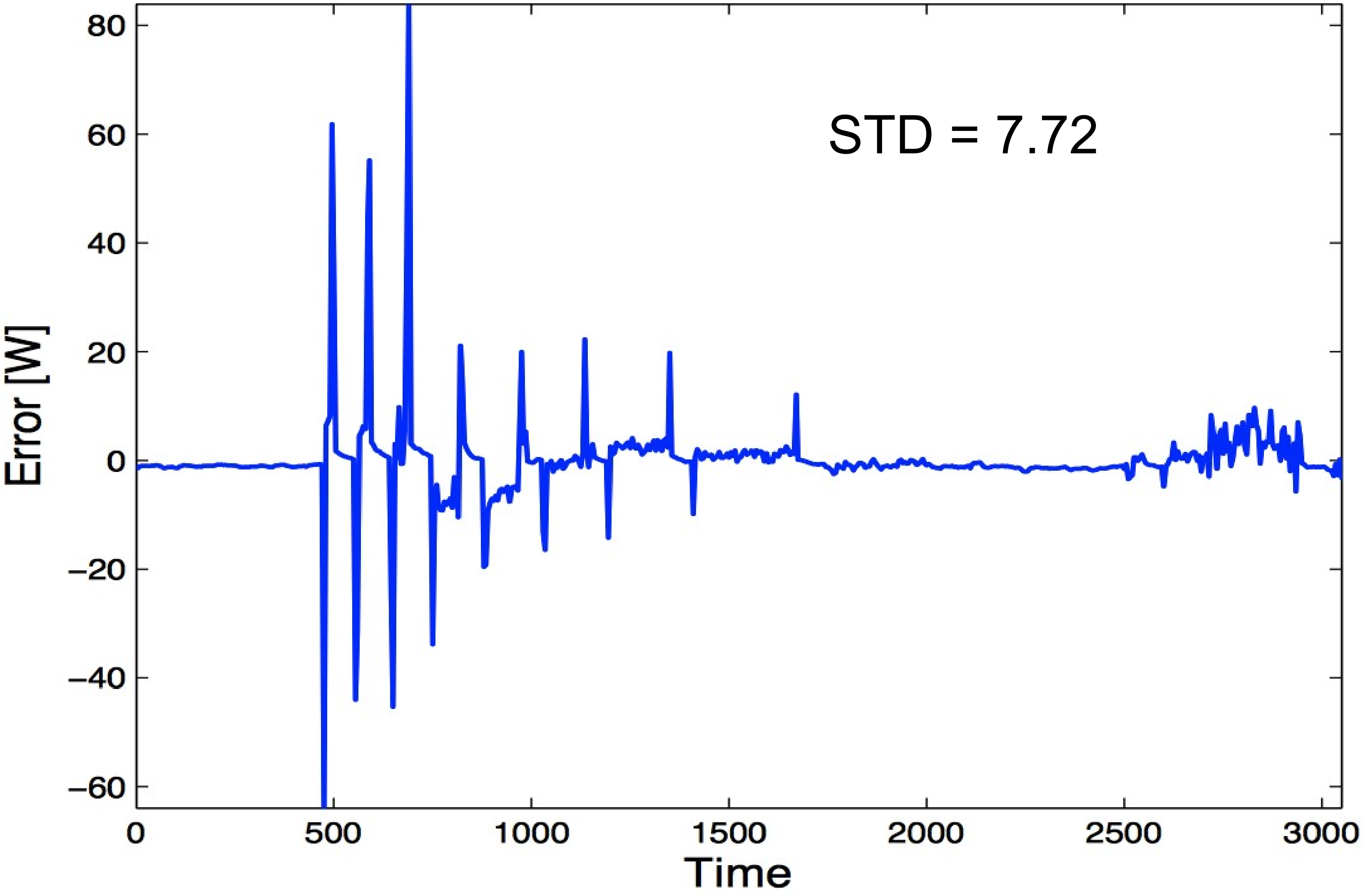
[CPU*P0]
[CPU*P1]
[CPU*P2]

90th Percentile



[CPU]	[CPU*P0] [CPU*P1] [CPU*P2]	[CPU^1] [CPU^2] .... [CPU^6]	[CPU*P0] [CPU^2*P0] .... [CPU^2*P3]

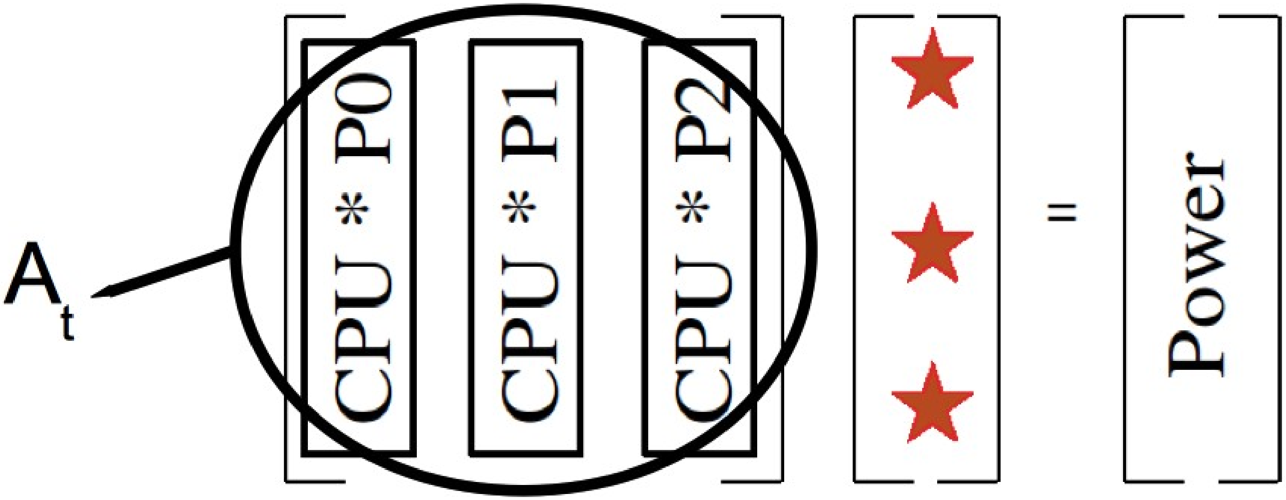
Error Analysis



Dynamic Model

- **Main error during CPU load transitions**
  - Possibly due to power dynamics caused by capacitors in power circuitry etc.

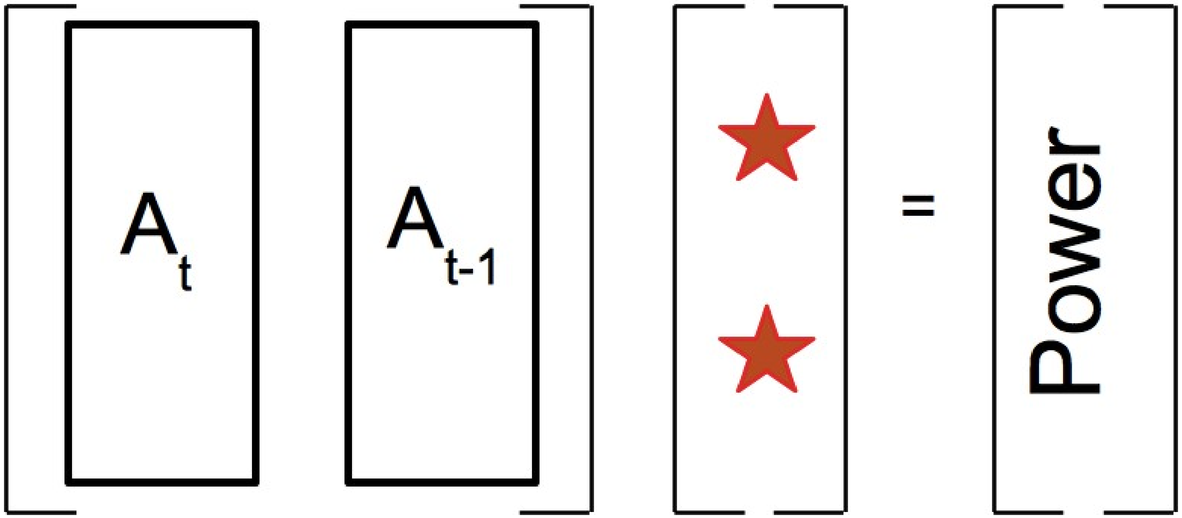
- **Solution: use history**



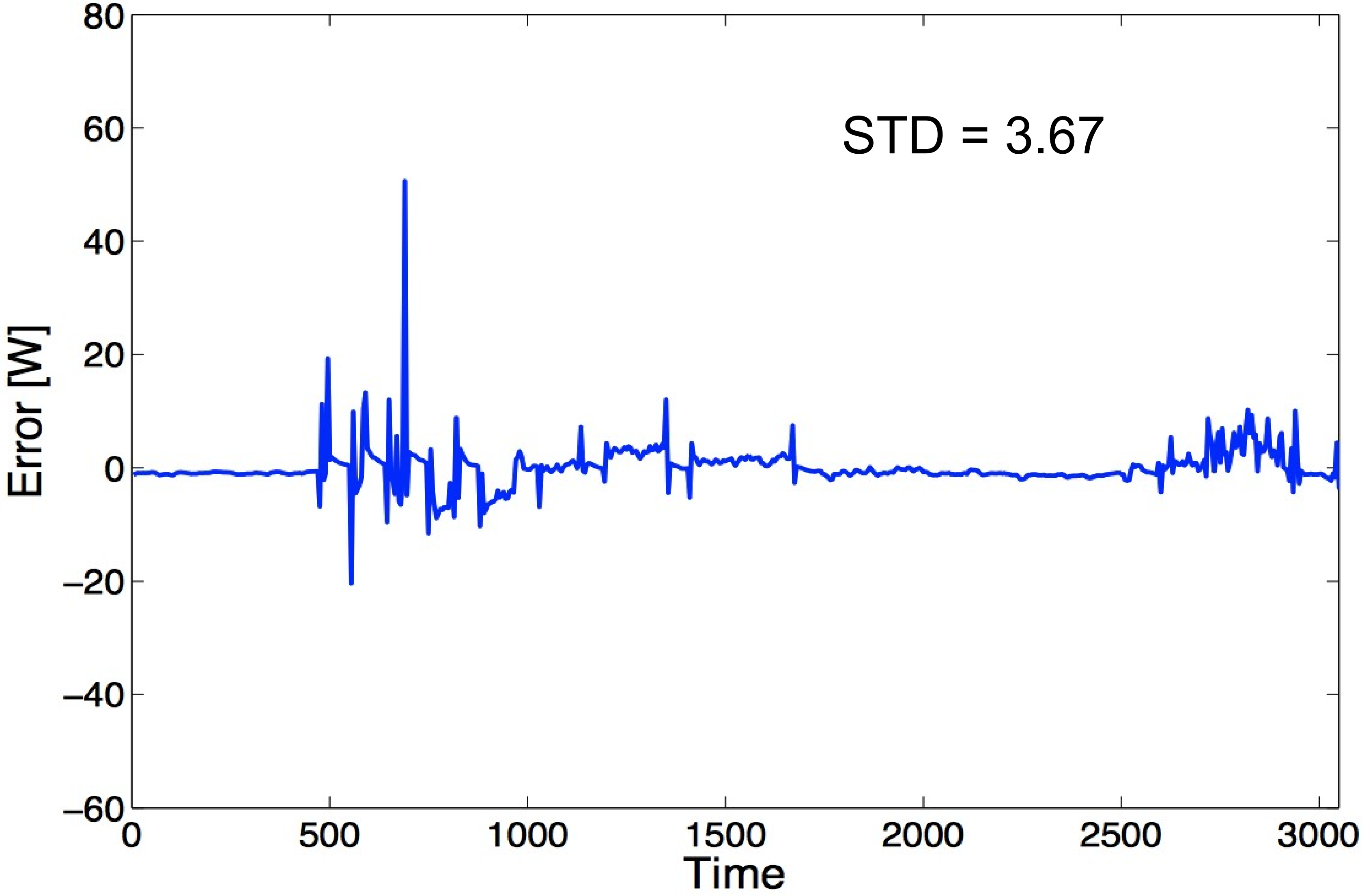
Dynamic Model

- **Main error during CPU load transitions**
  - Possibly due to power dynamics caused by capacitors in power circuitry etc.

- **Solution: use history**

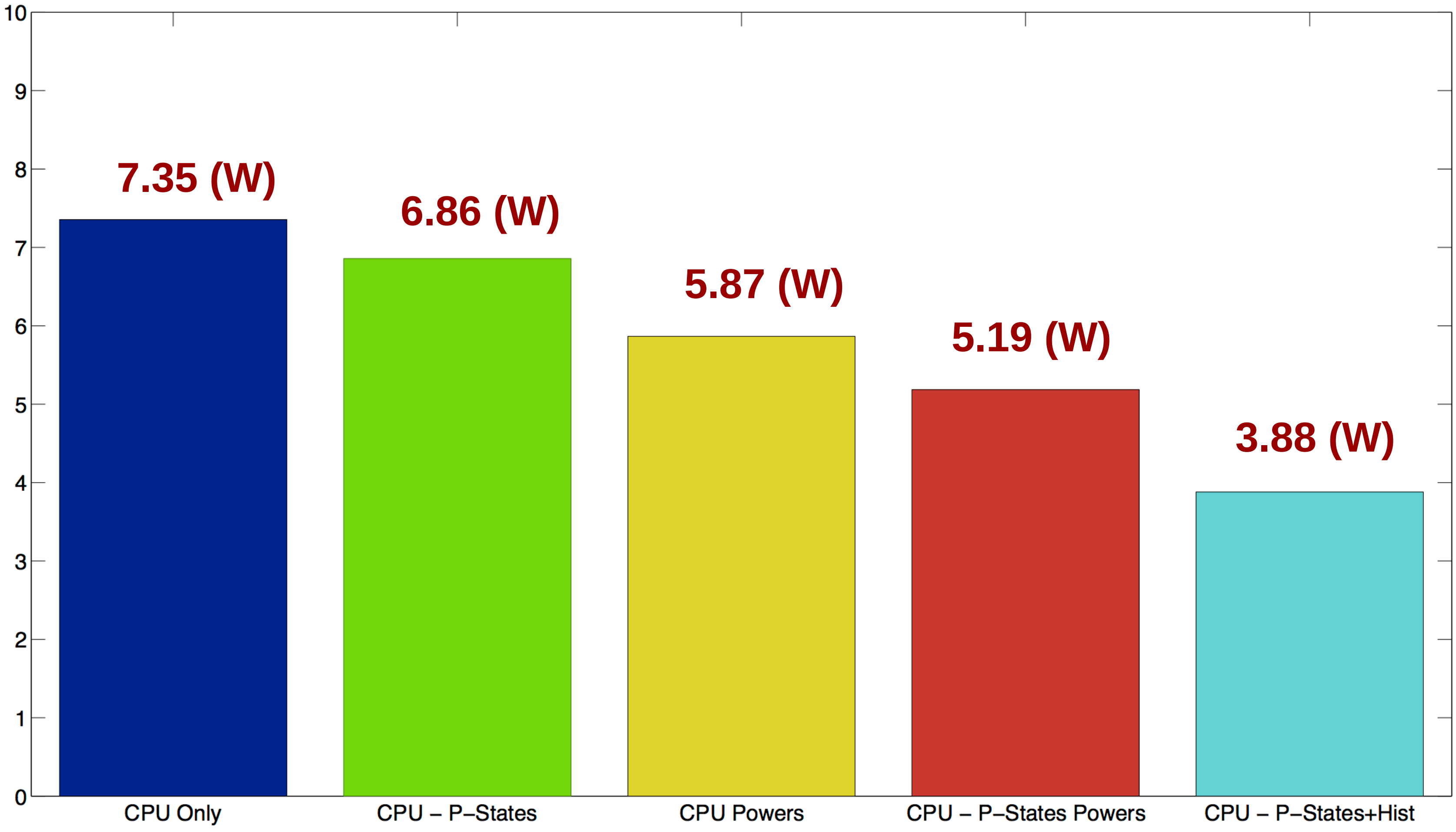


Error Analysis – Dynamic Model



Estimation of server power consumption

- Bars shows the 90th percentile of the estimation error

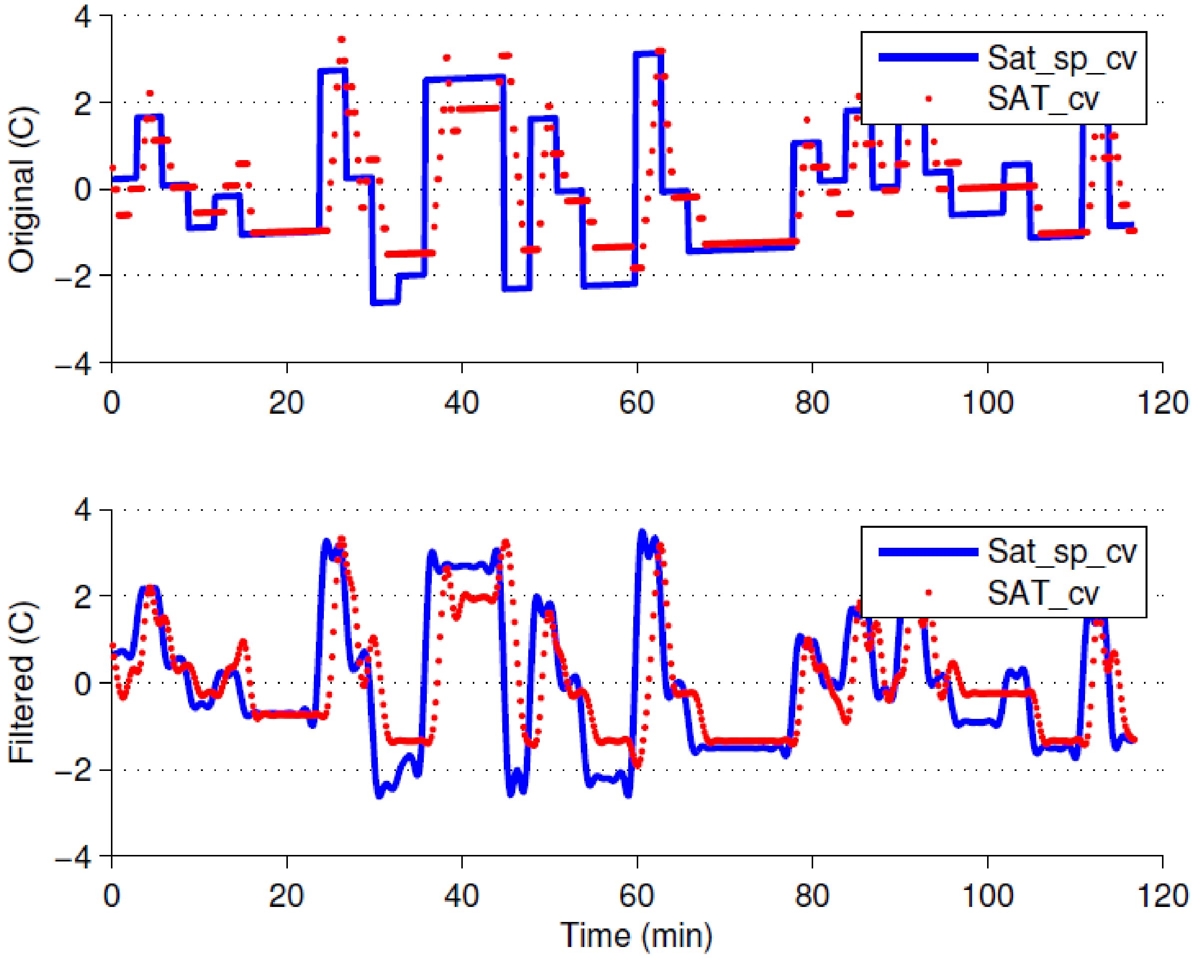


Conclusions

- **For cn007, CPU Load + P-State are the informative data**
  - Network and Disk are insignificant
  - Memory?
- **Main error during CPU load transition**
  - Possible Causes:
    - Synchronization between tools
    - Effects of power circuitry
  - Solution: use history (dynamic model)
- **Non-linear models seem to fit data better**

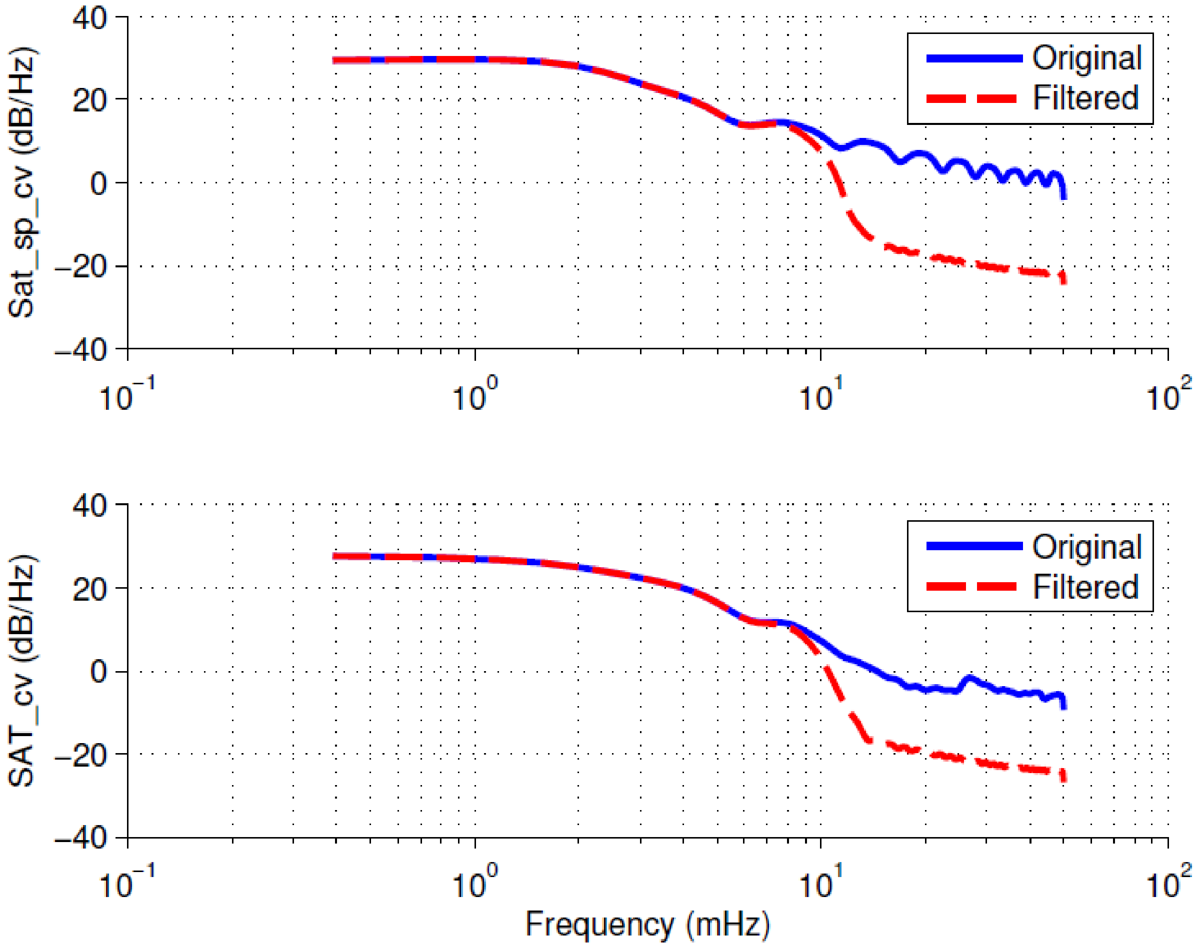
I/O models of CRAC units
Data collected by researchers at
HP Labs, Palo Alto, CA

Input and output data over time



- **Original data**
  - Set point (SAT_cv)
  - Output temperature (SAT)
- **Filtered data**
  - Low-pass filter at 10(mHZ)

Power spectral density of I/O data

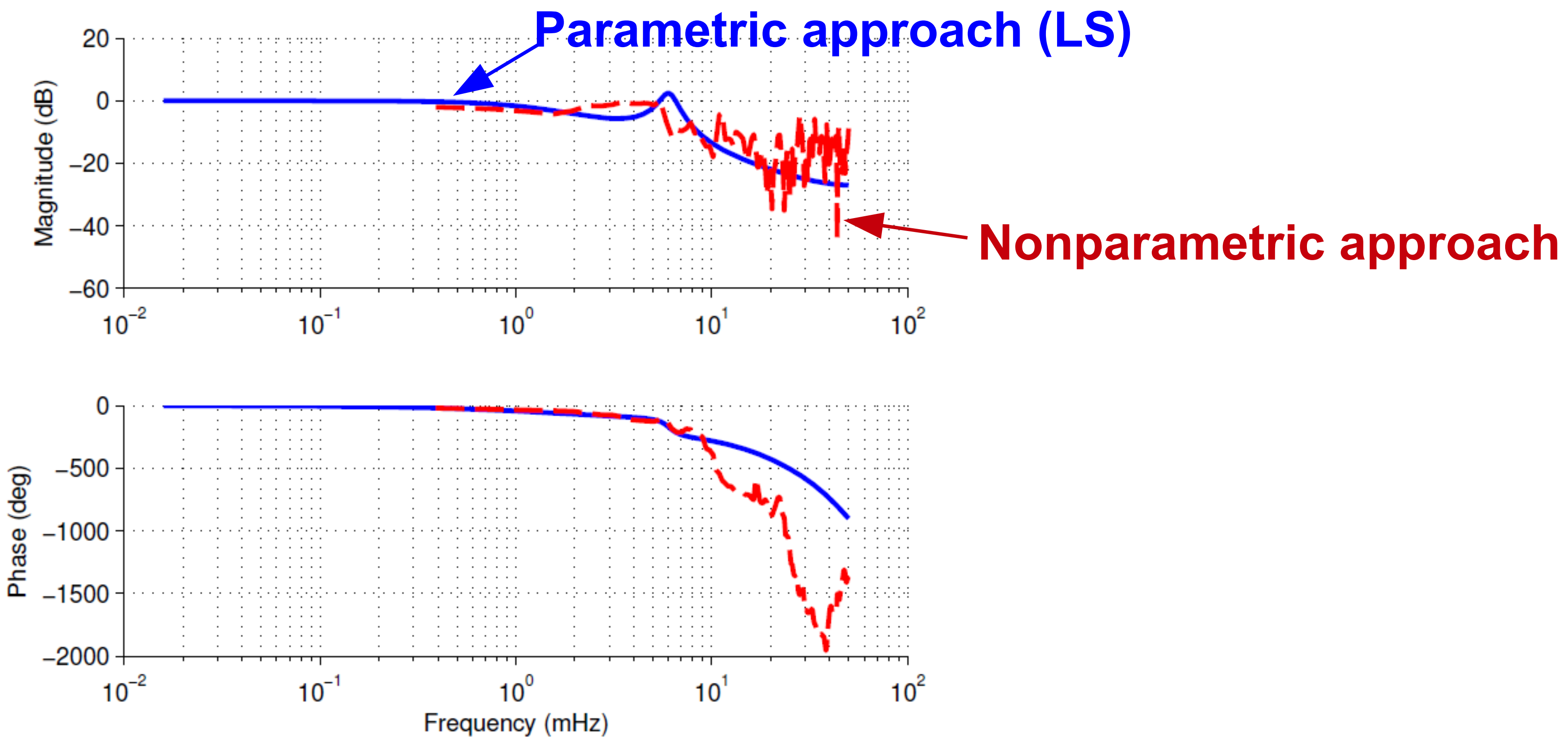


- **Reference temperature**
  - Set point
- **Output temperature**
  - Supply air temperature (SAT)

Estimated transfer function

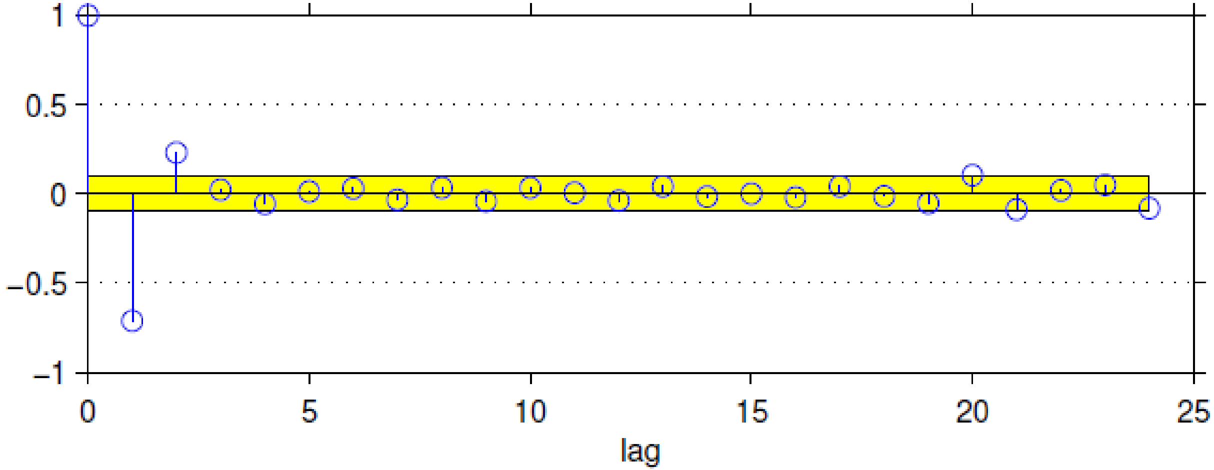
- **ARX model**

$$ARX433_5(s) = e^{-40.40s}\frac{0.0088743(s^2+2\cdot 0.6028\cdot 0.0375s+0.0375^2)}{(s+0.008609)(s^2+2\cdot 0.1014\cdot 0.0381s^2+0.0381^2)}$$

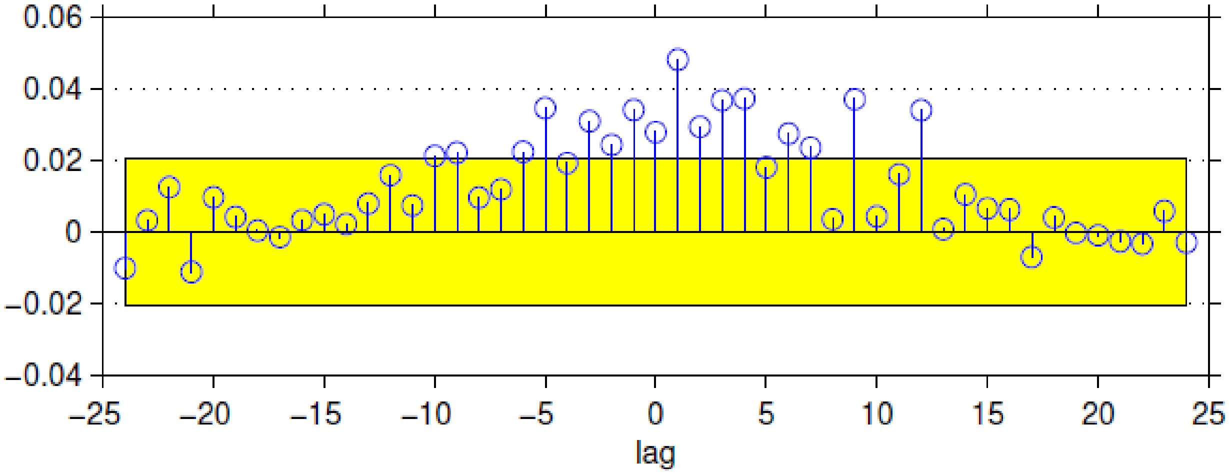


Residual analysis

- **Fit on the validation data-set: 84 %**



Cross corr. function between input SAT$_5$reference value and residuals from output SAT$_5$current value



Notes on the cyber-physical index (CPI)

Sensitivity index

- **Given a data center**
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?
- **Zone input temperature at the equilibrium**

$$\boldsymbol{T}_{\mathrm{in}\mathcal{N}} = \boldsymbol{\mathcal{L}}\boldsymbol{\eta} + \Psi_{[\mathcal{N},\mathcal{C}]}\boldsymbol{T}_{\mathrm{ref}}$$

  - Relative sensitivity index of the *i* th zone

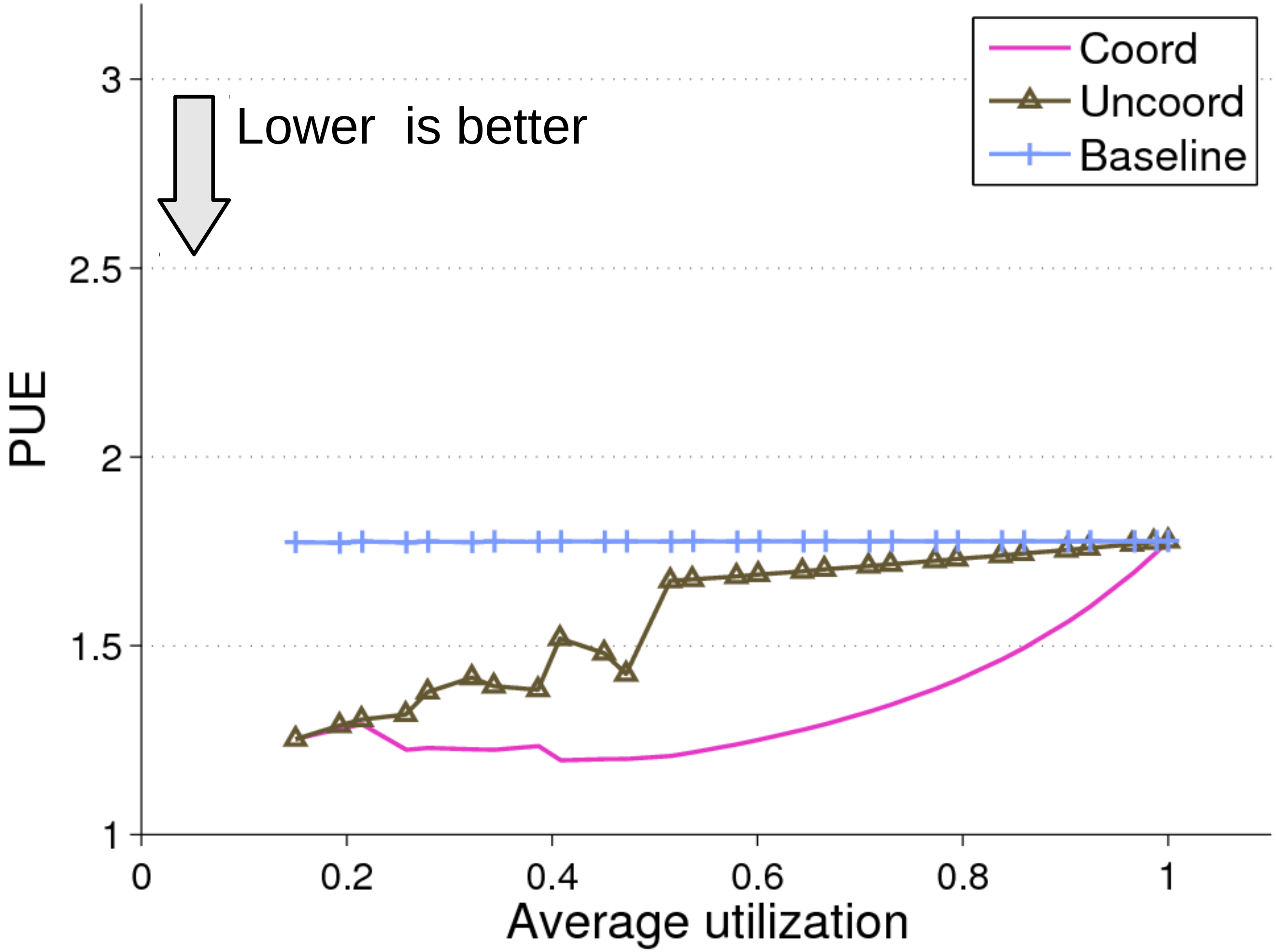
$$\mathcal{S}_i = \left|\left|\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}}\right|\right|_2 \Big/ \left|\left|\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}}\right|\right|_2 \qquad \boldsymbol{z} = \left[\ {\boldsymbol{T}_{\mathrm{ref}}}^T \boldsymbol{\eta}^T\ \right]^T$$

- **Data center *cyber-physical index***

$$\boldsymbol{CPI} = k\,\mathrm{std}\left(\left[\ S_1 \quad \ldots \quad S_N\ \right]\right)$$

Power usage effectiveness

- **PUE =** $\dfrac{\textbf{Total data center power consumption}}{\textbf{Server power consumption}}$



Thermal network

- **Linear model**

Output temperature of zones and CRAC units

$$\dot{\boldsymbol{T}}_{\text{out}}(\tau) = A_T \boldsymbol{T}_{\text{out}}(\tau) + B_T[\mathbf{p}_{\mathcal{N}}(\tau)^T \ \boldsymbol{T}_{\text{ref}}(\tau)^T]^T$$

Power consumption of zones

Reference temperature of CRAC units

$$\boldsymbol{T}_{\text{in}}(\tau) = \Psi \boldsymbol{T}_{\text{out}}(\tau) \leq \overline{\boldsymbol{T}_{\text{in}}}$$

Inlet temperature constraint

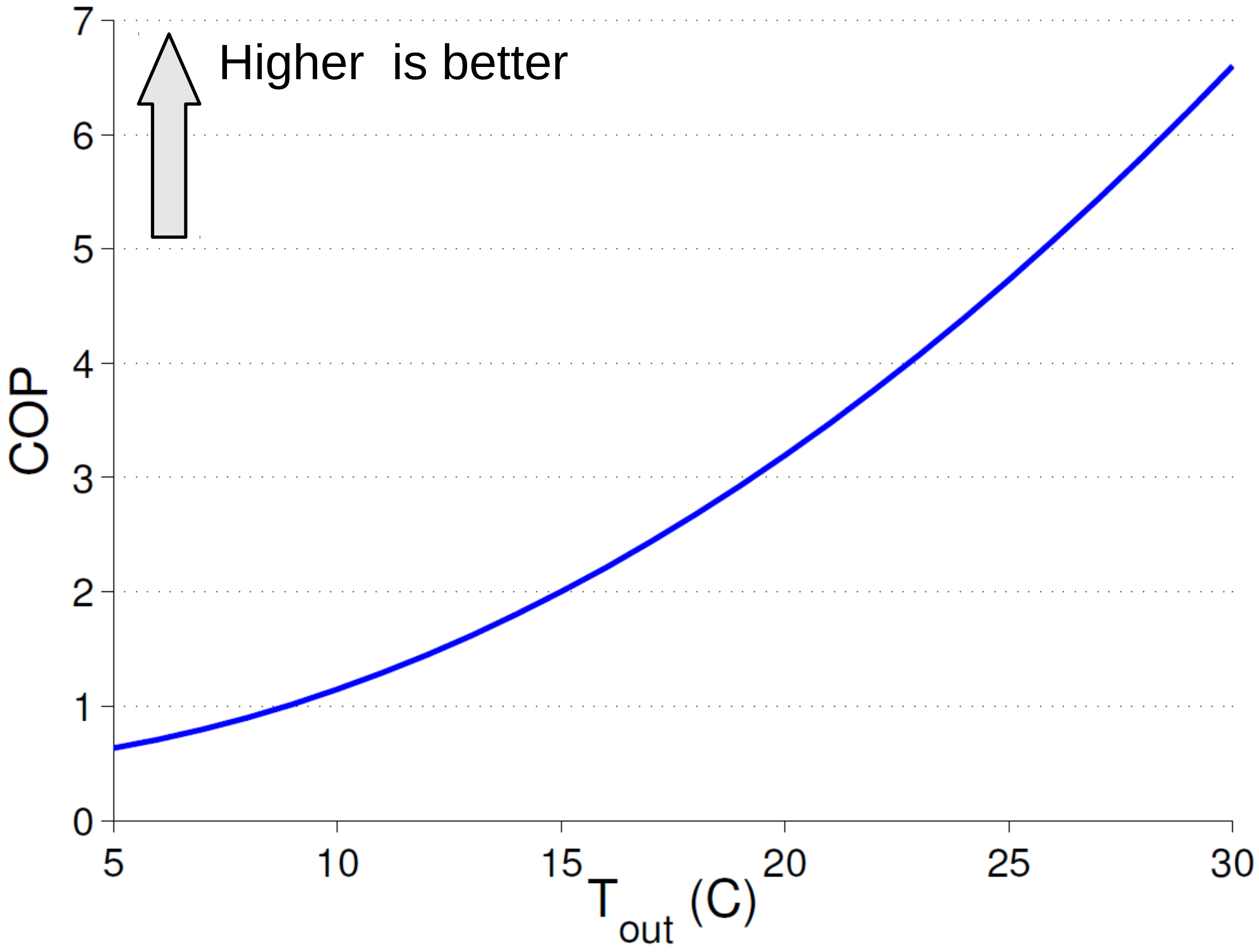
- **CRAC power consumption depends on the coefficient of performance (COP)**

$$\mathsf{p}_i(t) = \frac{\dot{Q}_i(t)}{COP_i(T_{\text{out},i}(t))}$$

Heat removed rate (W)

Q, Tang *et al.* "Sensor-based fast thermal evaluation model for energy efficient high-performance data centers", 2006

Efficiency of CRAC nodes



J. Moore et al. "Making scheduling ”cool”: temperature-aware workload placement in data centers." 2005

Sensitivity index

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

- **Relative sensitivity of the *i* th zone at the equilibrium**
  - $$\mathcal{S}_i = \left|\left| \frac{\partial T_{\text{in},i}}{\partial \boldsymbol{T}_{\text{ref}}} \right|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \right|\right|_2 \Big/ \left|\left| \frac{\partial T_{\text{in},i}}{\partial \boldsymbol{z}} \right|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \right|\right|_2 \qquad \boldsymbol{z} = \left[\ {\boldsymbol{T}_{\text{ref}}}^T \boldsymbol{\eta}^T\ \right]^T$$

Sensitivity index

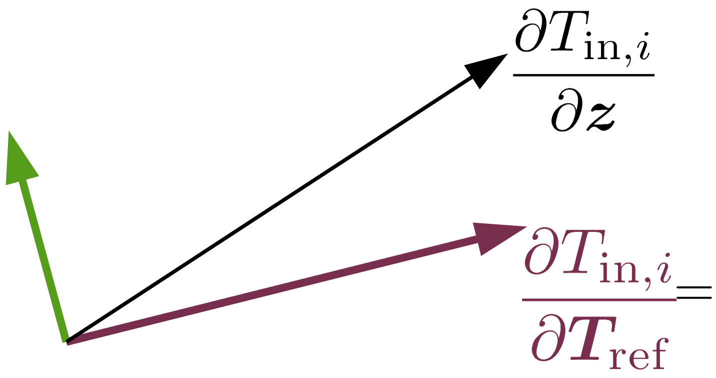
- **Given a data center**
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?
- **Relative sensitivity of the *i* th zone at the equilibrium**
  - $\mathcal{S}_i = \left|\left| \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}} \right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \right|\right|_2 \Big/ \left|\left| \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}} \right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \right|\right|_2 \qquad \boldsymbol{z} = \left[\ \boldsymbol{T}_{\mathrm{ref}}{}^{T} \boldsymbol{\eta}^{T}\ \right]^{T}$

$$\begin{bmatrix} \left.\frac{\partial T_{\mathrm{in},i}}{\partial \eta_1}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \\ \vdots \\ \left.\frac{\partial T_{\mathrm{in},i}}{\partial \eta_1}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \end{bmatrix} = \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{\eta}}$$

$$\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}} = \begin{bmatrix} \left.\frac{\partial T_{\mathrm{in},i}}{\partial T_{\mathrm{ref},1}}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \\ \vdots \\ \left.\frac{\partial T_{\mathrm{in},i}}{\partial T_{\mathrm{ref},C}}\right|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \end{bmatrix}$$

Sensitivity index

- **Given a data center**
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?
- **Relative sensitivity of the *i* th zone at the equilibrium**
  - $\mathcal{S}_i = \left|\left| \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \right|\right|_2 \Big/ \left|\left| \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \right|\right|_2 \qquad \boldsymbol{z} = \left[\ {\boldsymbol{T}_{\mathrm{ref}}}^T \boldsymbol{\eta}^T\ \right]^T$



$$\begin{bmatrix} \frac{\partial T_{\mathrm{in},i}}{\partial \eta_1}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \\ \vdots \\ \frac{\partial T_{\mathrm{in},i}}{\partial \eta_1}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \end{bmatrix} = \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{\eta}}$$

$$\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}}$$

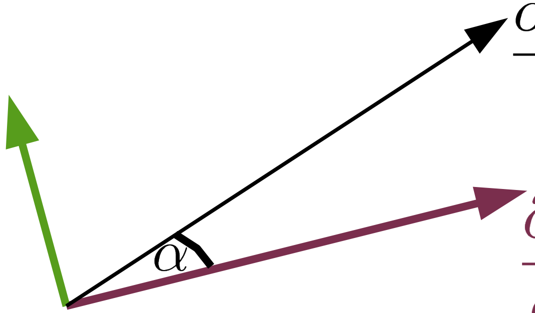
$$\frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}} = \begin{bmatrix} \frac{\partial T_{\mathrm{in},i}}{\partial T_{\mathrm{ref},1}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \\ \vdots \\ \frac{\partial T_{\mathrm{in},i}}{\partial T_{\mathrm{ref},C}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\mathrm{out}})} \end{bmatrix}$$

Sensitivity index

- **Given a data center**
  - How much energy can be saved by a coordinated controller, with respect to an uncoordinated controller?
- **Relative sensitivity of the *i* th zone at the equilibrium**

$$S_i = \left|\left| \frac{\partial T_{\text{in},i}}{\partial \boldsymbol{T}_{\text{ref}}}\bigg|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \right|\right|_2 \Big/ \left|\left| \frac{\partial T_{\text{in},i}}{\partial \boldsymbol{z}}\bigg|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \right|\right|_2 \qquad \boldsymbol{z} = \left[\ \boldsymbol{T}_{\text{ref}}{}^T \boldsymbol{\eta}^T\ \right]^T$$

$$\begin{bmatrix} \frac{\partial T_{\text{in},i}}{\partial \eta_1}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \\ \vdots \\ \frac{\partial T_{\text{in},i}}{\partial \eta_1}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \end{bmatrix} = \frac{\partial T_{\text{in},i}}{\partial \eta}$$



$$\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{z}}$$

$$\alpha$$

$$S_i = cos(\alpha)$$

$$\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{T}_{\text{ref}}} = \begin{bmatrix} \frac{\partial T_{\text{in},i}}{\partial T_{\text{ref},1}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \\ \vdots \\ \frac{\partial T_{\text{in},i}}{\partial T_{\text{ref},C}}\Big|_{(\boldsymbol{l},\boldsymbol{T}_{\text{out}})} \end{bmatrix}$$

Sensitivity index for the *i* th zone

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

- **Cyber-physical index (CPI)**
  - Relative sensitivity of the *i* th zone at the equilibrium

$$\mathcal{S}_i = \left\| \left. \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{T}_{\mathrm{ref}}} \right|_{(\boldsymbol{l}, \boldsymbol{T}_{\mathrm{out}})} \right\|_2 \Big/ \left\| \left. \frac{\partial T_{\mathrm{in},i}}{\partial \boldsymbol{z}} \right|_{(\boldsymbol{l}, \boldsymbol{T}_{\mathrm{out}})} \right\|_2 \qquad \boldsymbol{z} = \left[\, \boldsymbol{T}_{\mathrm{ref}}{}^T \boldsymbol{\eta}^T \,\right]^T$$

  - Easy to compute when using the proposed model

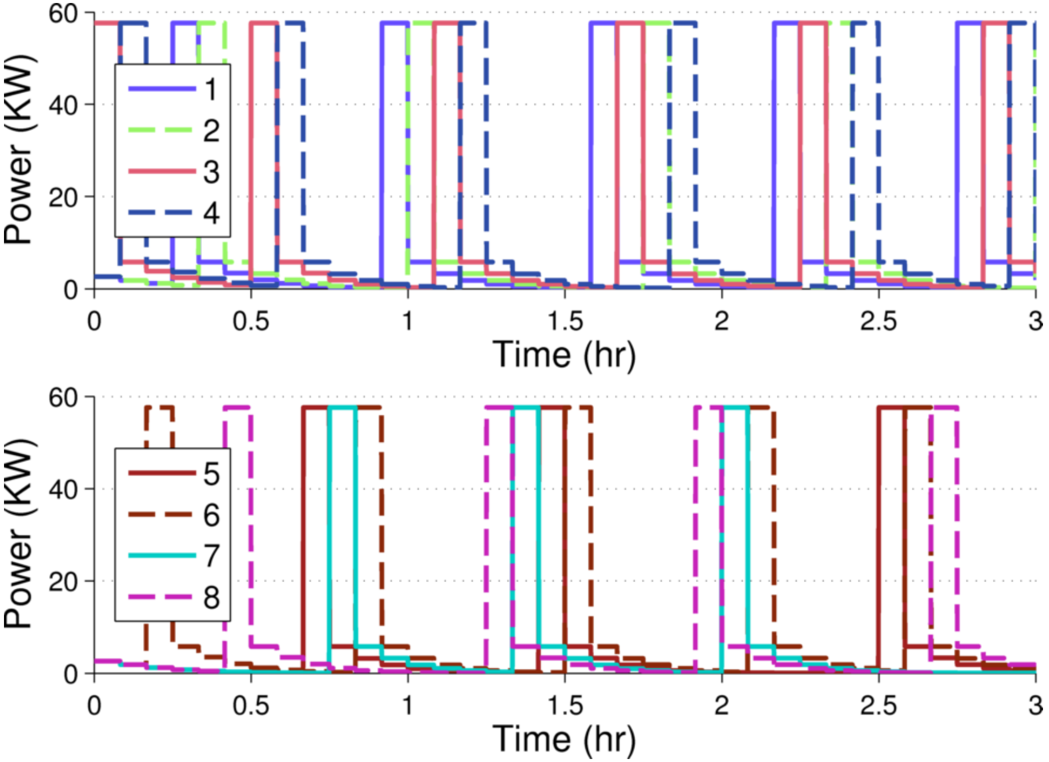
- **Data center *cyber-physical index***

$$\boldsymbol{CPI} = k \operatorname{std}\left(\left[\, S_1 \quad \ldots \quad S_N \,\right] \right)$$

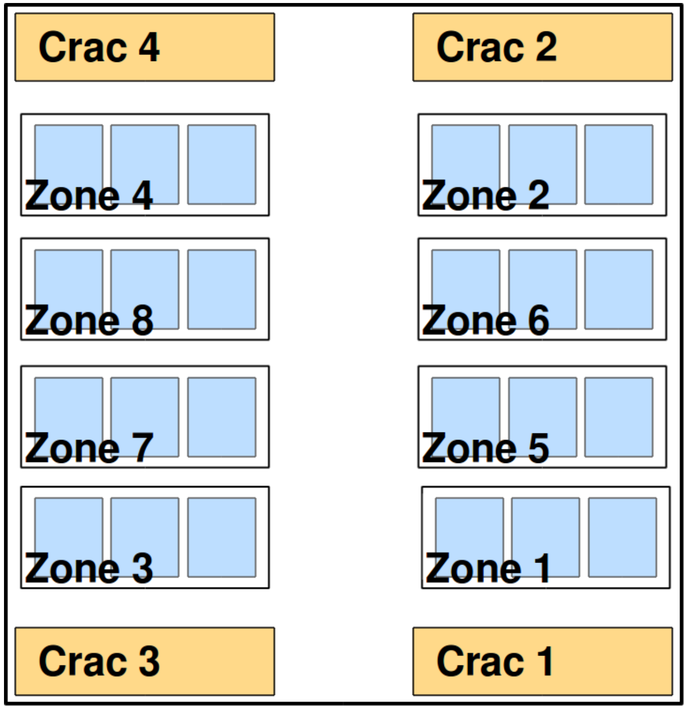
Stability of the coordinated controller

Economic MPC and data center

- **Coordinated control approach**
  - Perfect predictive model
  - Constant workload arrival rate
- **Zone power consumption**

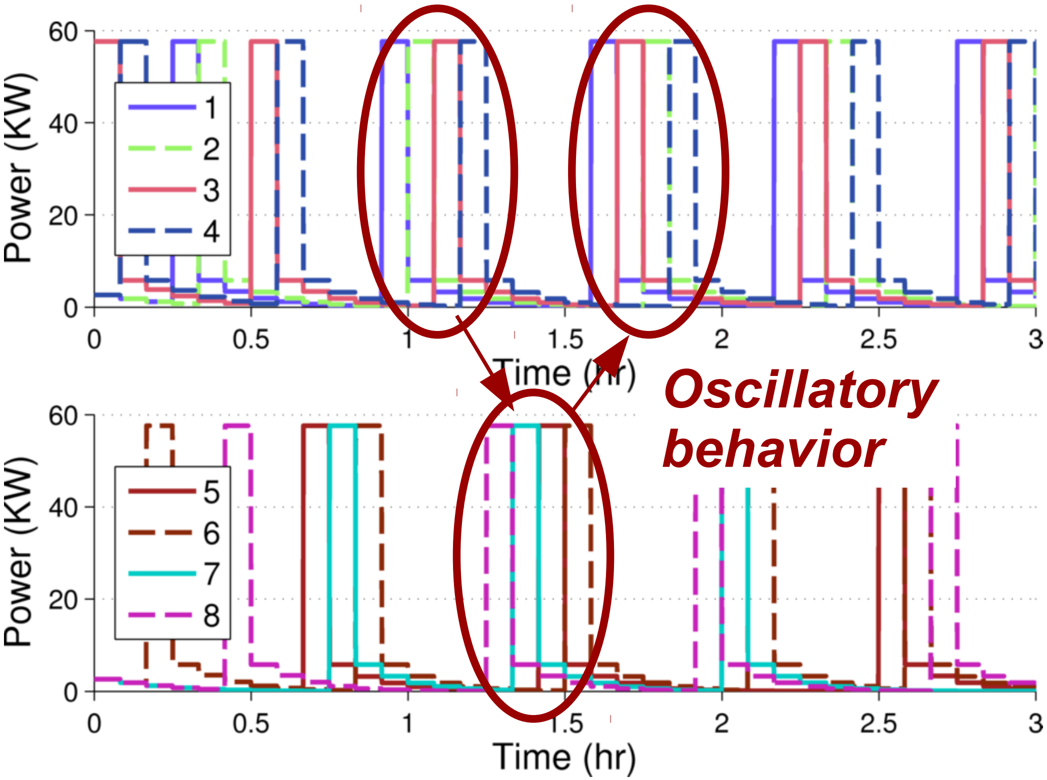


Data center layout



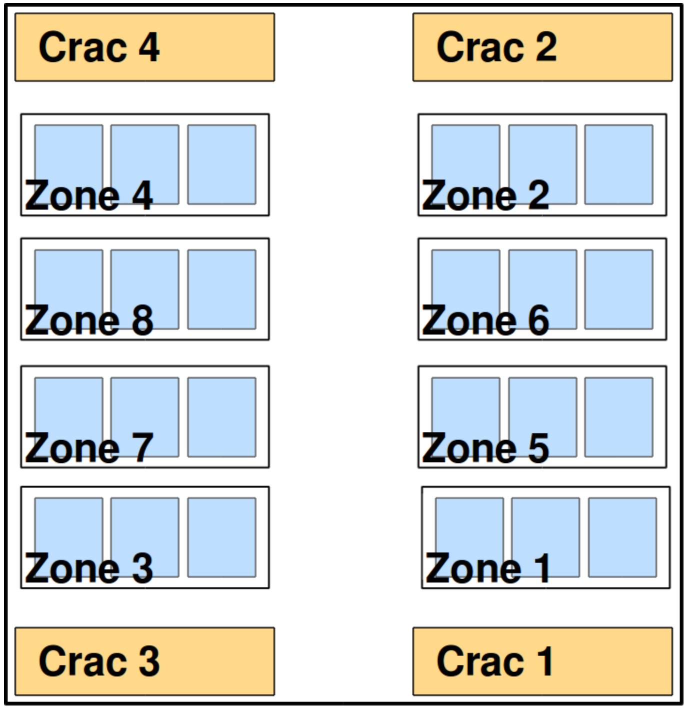
Economic MPC and data center

- **Coordinated control approach**
  - Perfect predictive model
  - Constant workload arrival rate
- **Zone power consumption**



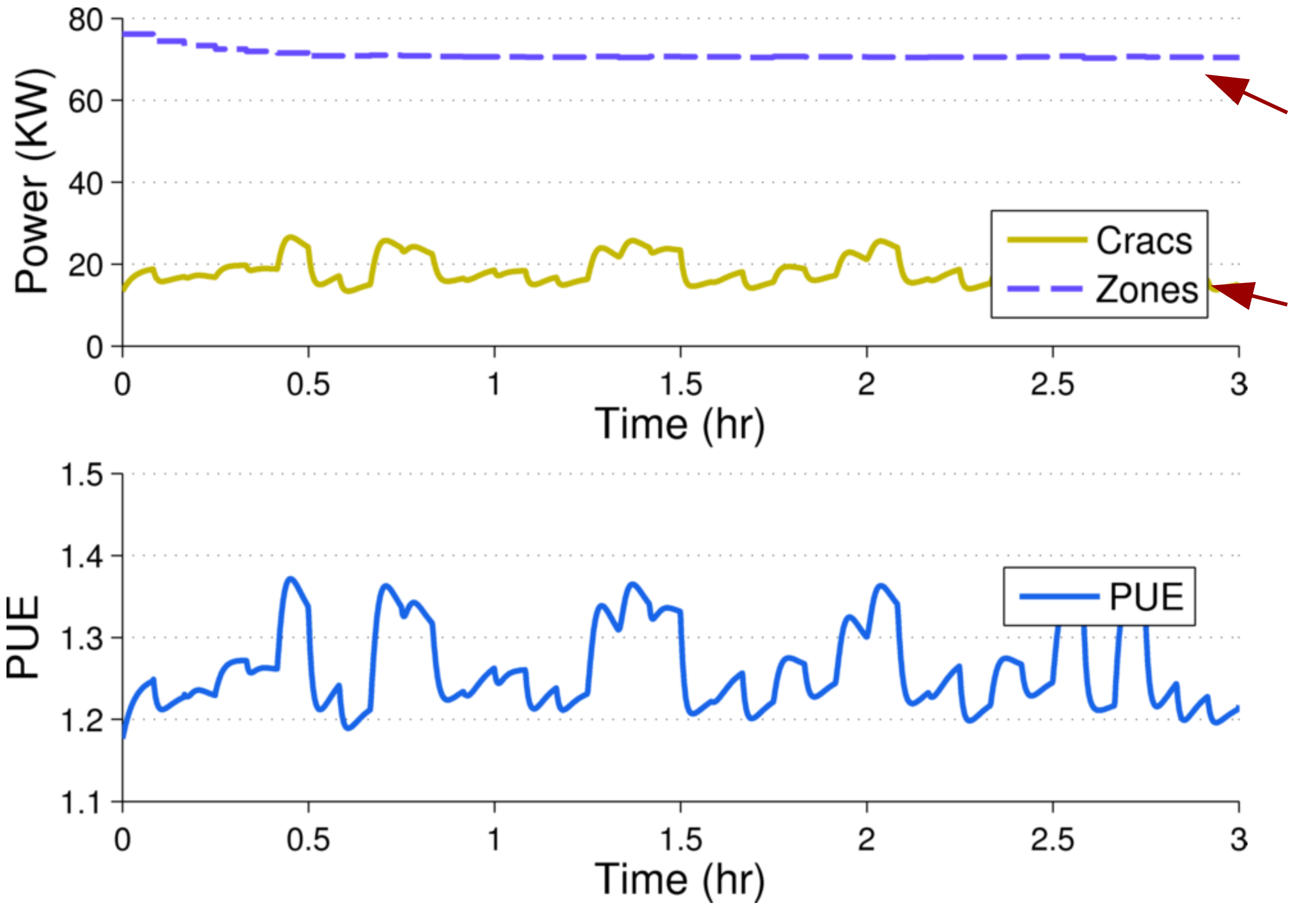
Oscillatory behavior

Data center layout



Economic MPC and data center

- Total server and CRAC power consumption



- All workload is executed
- Constant arrival rate induces constant total power consumption
- Variation on the location of workload execution and cooling leads to variations on the CRAC power consumption

Effects of model mismatch

Control-oriented model

- **Considers each rack and each CRAC as a dynamic subsystem**

$$\mathrm{x}_i(k+1) = A_i \mathrm{x}_i(k) + B_i(k)\mathrm{u}_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{R+C} B_{i,j}\mathrm{z}_j(k)$$

$$\mathrm{z}_i(k) = G_i \mathrm{x}_i(k)$$

	$x_i(k)$	***$u_i(k)$***	***$z_i(k)$***
Racks	Outlet temperature of every server in the rack *(vector)*	Relative amount of resources assigned to different jobs *(vector)*	Average rack outlet temperature
CRACs	Supplied air temperature *(scalar)*	Desired supplied air temperature *(scalar)*	Supplied air temperature

Thermal constraints

- **Bound the inlet temperature of each rack**

$$\mathsf{T}_{\mathrm{in},\mathcal{R}_i}(k) \le \overline{\mathsf{T}}_{\mathrm{in},\mathcal{R}_i}, \qquad i = 1, \ldots, R \longleftarrow \textit{Number of racks}$$

- **Inlet-outlet relationship**

$$T_{\mathrm{in},\mathcal{R}_i}(k) = \underbrace{\sum_{j\in\mathcal{R}_v} \gamma_{i,j} T_{\mathrm{out},j}(k)}_{\textit{Servers on the same rack}} + \underbrace{\underbrace{\sum_{j\in\mathcal{C}} \gamma_{i,j} T_{\mathrm{out},j}(k)}_{\textit{CRACs}} + \underbrace{\sum_{\substack{j=1\\ j\neq v}}^{R} \gamma_{i,\mathcal{R}_j} \mathsf{T}_{\mathrm{out},\mathcal{R}_j}(k)}_{\textit{Other racks}}}$$

$$\downarrow \qquad\qquad \swarrow$$

$$F_i \mathsf{x}_i(k) + \sum_{\substack{j=1\\ j\neq i}}^{S} F_{i,z_j} \mathsf{z}_j(k) \le \overline{\mathsf{z}}_i$$

S: number of subsystems (*S*=*R*+*C*)

[Tang et al. ICISIP 2006]

One step ahead coordinated control

- **The coordinated control problem can be written as**

$$\min_{\mathcal{U}_1} \sum_{i=1}^{S} \mathsf{c}_{i,u}^{T}(k)\hat{\mathsf{u}}_i(k|k)$$

s.t.

$$\underline{\mathsf{u}}_i \leq \hat{\mathsf{u}}_i(k|k) \leq \overline{\mathsf{u}}_i,$$

$$H_i\hat{\mathsf{u}}_i(k|k) \leq 1$$

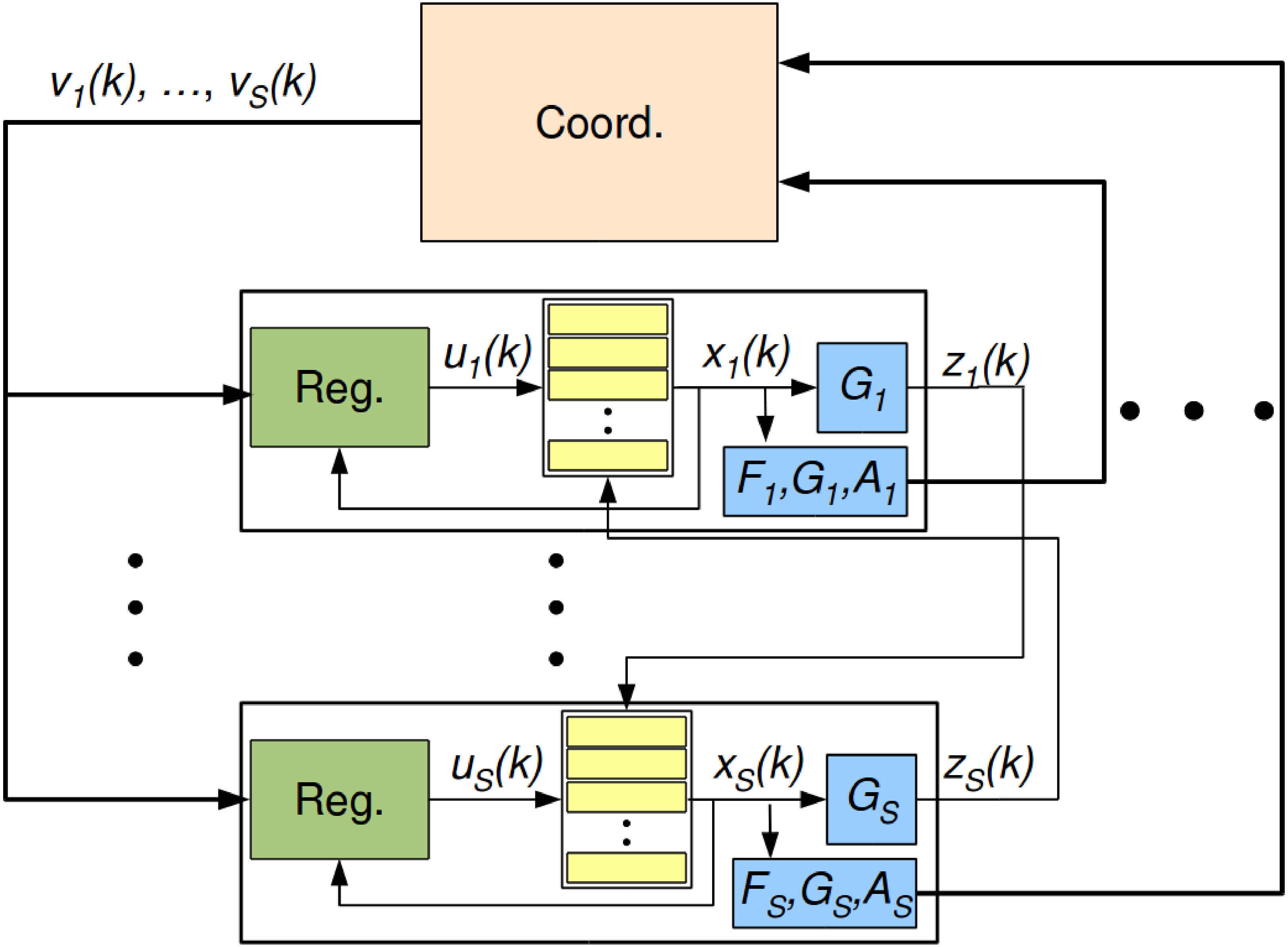
$$F_i B_i(k)\hat{\mathsf{u}}_i(k|k) + \overbrace{\sum_{\substack{j=1\\ j\neq i}}^{S} F_{i,z_j} G_j B_j(k)\hat{\mathsf{u}}_j(k|k)}^{\textit{Coupling term}} + \mathsf{k}_i(k) \leq \overline{\mathsf{z}}_i,$$

$$\mathcal{U}_1 = \left\{\hat{\mathsf{u}}_1(k|k), \ldots, \hat{\mathsf{u}}_S(k|k)\right\},$$

Problem analysis

- **Data centers are large-scale systems**
  - It may be impossible to collect data from all of the sensors and compute the control actions with a single controller
- **Every subsystem affects other subsystems only through its output**
  - *For racks, the output has a much lower dimension than the state*
- ***Is it possible to exploit the particular coupling, in order to derive an effective hierarchical control strategy?***

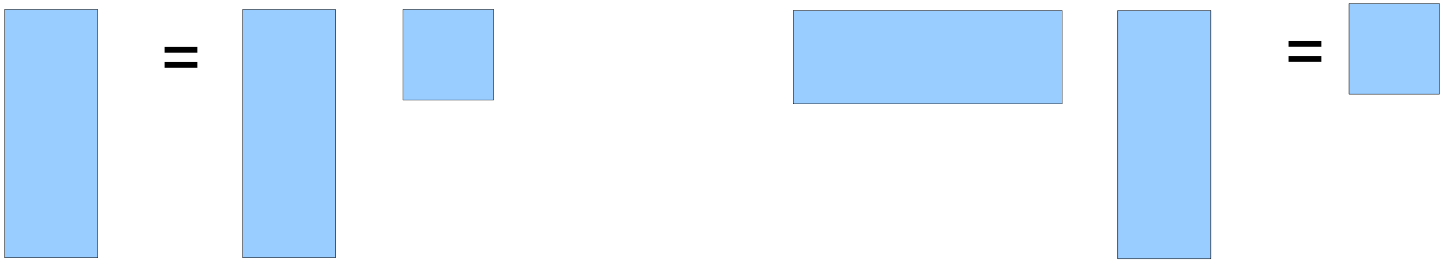
Proposed hierarchy strategy



Coordinator optimization problem

- **Stated in terms of the virtual control inputs {*v*$_i$*(k)* }**
- **A collection of matrices {*M*$_i$*(k)*} approximate the actions of local controllers**
    - The coordinator considers the following relationship

$$\mathrm{u}_i(k) = M_i(k)\mathrm{v}_i(k) \qquad G_iB_i(k)M_i(k) = I$$



- **Multiple collections of {*M*$_i$*(k)*} matrices can be defined, but not all of them yield an overall optimal control strategy**

Stage one – coordinator problem

- **Global optimization problem in the {*v*$_i$*(k)*} variables**

$$\min_{\mathcal{V}_1} \sum_{i=1}^{S} \mathsf{c}_{i,u}^T(k) M_i(k) \hat{\mathsf{v}}_i(k|k)$$

s.t.

$$\underline{\mathsf{u}}_i \leq M_i(k)\hat{\mathsf{v}}_i(k|k) \leq \overline{\mathsf{u}}_i,$$
$$H_i M_i(k)\hat{\mathsf{v}}_i(k|k) \leq 1$$
$$F_i B_i(k) M_i(k)\hat{\mathsf{v}}_i(k|k) + F_{i,z}\hat{\mathsf{v}}(k|k) + \mathsf{k}_i(k) \leq \overline{\mathsf{z}}_i.$$

$$\mathcal{V}_1 = \{\hat{\mathsf{v}}_1(k|k), \ldots, \hat{\mathsf{v}}_S(k|k)\}$$

Stage two – regulator problem

- **Every regulator solves a local optimization problem**

$$\min_{\hat{\mathrm{u}}_i(k|k)} \mathrm{c}_{i,u}^T(k)\hat{\mathrm{u}}_i(k|k)$$

s.t.

$$\underline{\mathrm{u}}_i \leq \hat{\mathrm{u}}_i(k|k) \leq \overline{\mathrm{u}}_i$$

$$H_i\hat{\mathrm{u}}_i(k|k) \leq 1$$

$$F_iB_i(k)\hat{\mathrm{u}}_i(k|k) + \sum_{\substack{l=1\\ l\neq i}}^{S} F_{i,z_l}\hat{\mathrm{v}}_l(k|k) + \mathrm{k}_i(k) \leq \overline{\mathrm{z}}_i$$

$$\boxed{\hat{\mathrm{v}}_i^*(k|k) = G_iB_i(k)\hat{\mathrm{u}}_i(k|k)}$$

Ensures coherence among optimization problems

Properties of the control strategy

- **If the coordinator problem has an admissible solution, then all of the local regulator problems have an admissible solution**

- **The condition** $\hat{v}_i^*(k|k) = G_i B_i(k) \hat{u}_i(k|k)$
  - can be weakened to

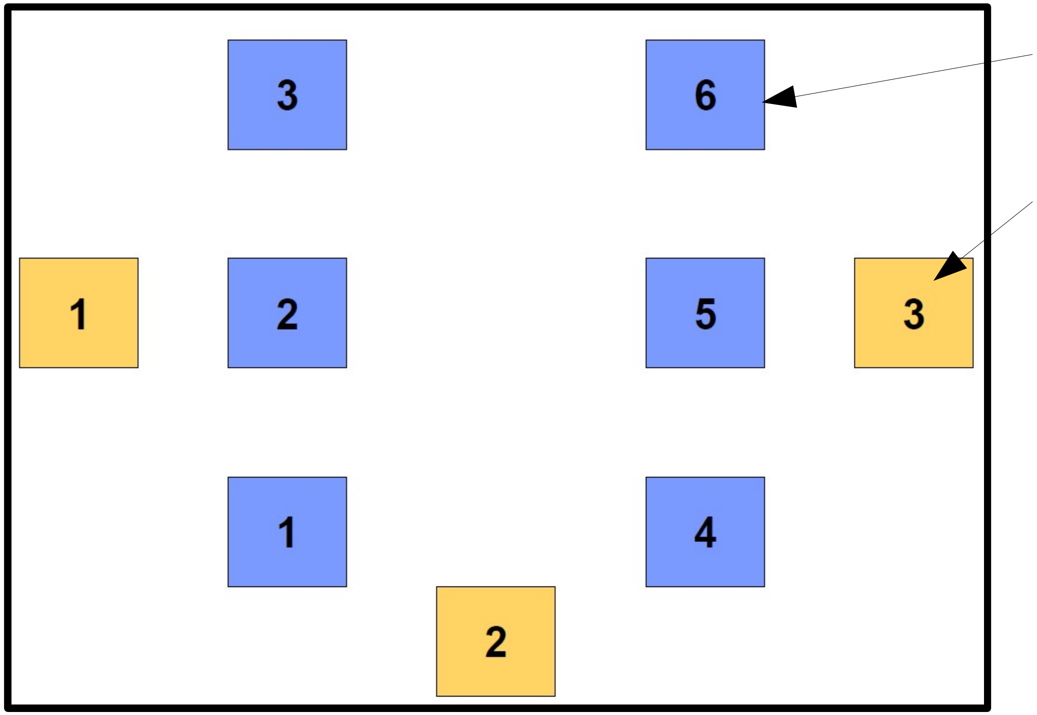
$$F_{j,z_i} \hat{v}_i^*(k|k) \geq F_{j,z_i} M_i(k) \hat{u}_i(k|k) \quad i,j = 1,\ldots,S,\ i \neq j$$

- **There always exists a collection of matrices {$M_i(k)$} such that the performance obtained by the hierarchy control strategy equals the performance of the centralized coordinated controller**

Choice of the {$M_i(k)$} matrices

- **Matrices {$M_i(k)$} are used by the coordinator to approximate the actions of local controllers**
    - Choice based on the a priori information about local subsystems
- **A partial characterization of the {$M_i(k)$} matrices is given in the paper**
- **How relevant is the is the effect of the {$M_i(k)$} matrices on the overall system performance?**

Simulation

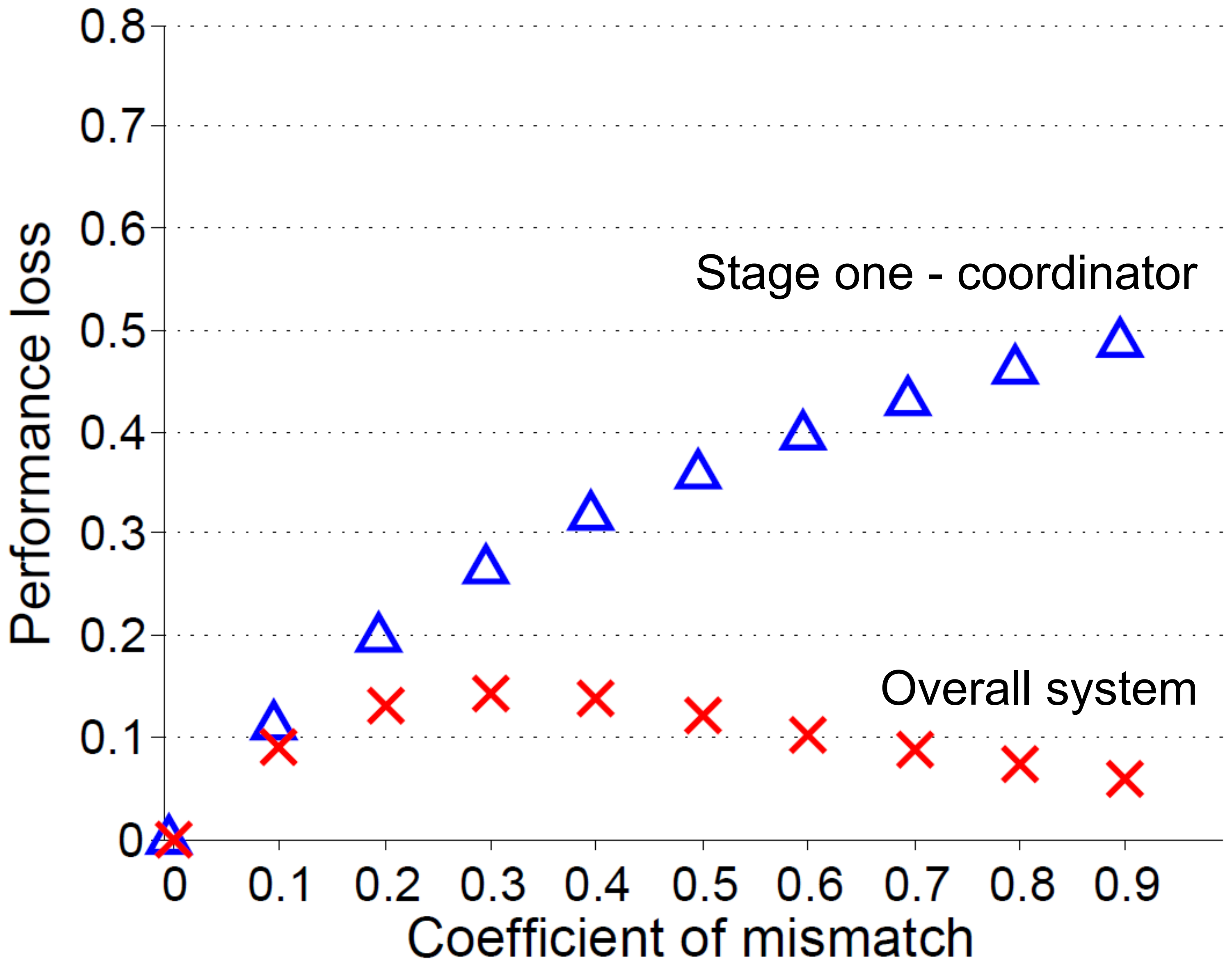


- **6 Racks**
- **3 CRACs**
- **6 job classes**
- **42 Server per rack**
  - Total of 1515 control variables

Simulation study

- **Objective: analyze the effects of a bad choice of the {$M_i(k)$} matrices**
- **Analysis based on**
  - the performance loss estimated at the first stage of the hierarchical strategy
  - the performance loss obtained at the second stage of the hierarchical strategy (overall system performance)
- **Coefficient of mismatch**
  - When the coefficient is 0, the {$M_i(k)$} matrices are optimally chosen
  - Higher values of the coefficient imply worse choices of the {$M_i(k)$} matrices
- **Analysis obtained via a Monte Carlo approach**
  - Each point is obtained averaging over 500 simulations

Performance loss



Conclusion and future work

- **The particular coupling among the subsystems and the large dimension of the control inputs makes hierarchical control an interesting approach**
    - In the particular case we discussed the hierarchal approach is effective

- **We are currently working in order to extend this approach for a multi-step MPC and for nonlinear cost functions**

- **Techniques to take advantage of the particular form of the input constraints are also being developed**