Ming Hsieh Department of Electrical Engineering

USC Viterbi School of Engineering

Energy-Efficient Video-Sharing Servers

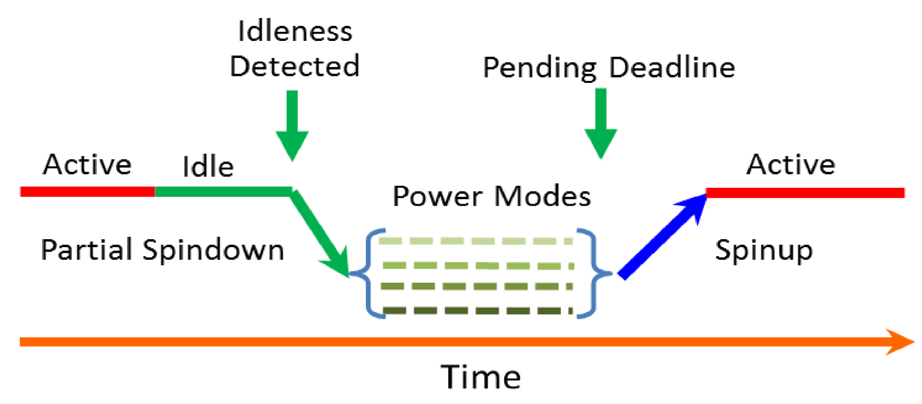
Hang Yuan, C.-C. Jay Kuo

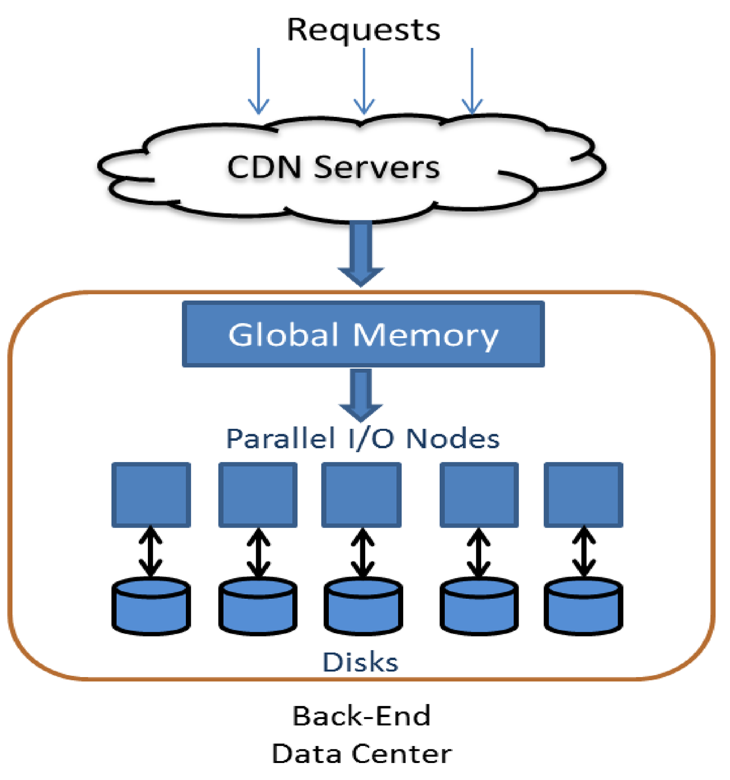
Motivation

- Data centers consume 30 billion watts of electricity
  - **90%** of the power is consumed by idle devices
- Ubiquity of large-scale video-sharing services
  - 72 hours of videos are being uploaded to YouTube every minute
  - Internet video will account for **57%** of the Internet traffic in 2014
- Storage systems with a large number of disks
  - Consume **25%~35%** of the total energy
  - Video-sharing services rely on such storage systems

Background

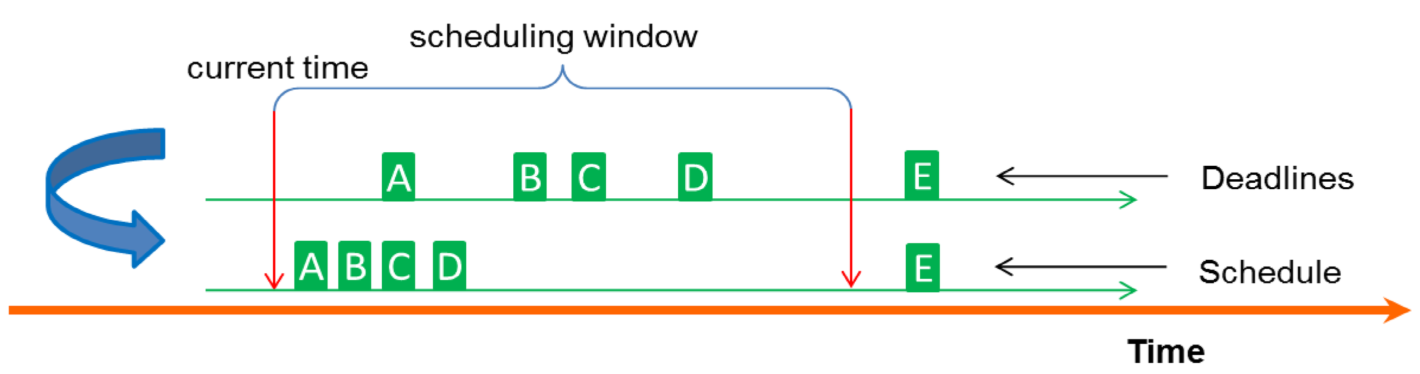
- Disk Energy Management
  - Use multi-speed disks
- Video-Sharing Services
  - Huge number of videos and users
  - Diverse video repository

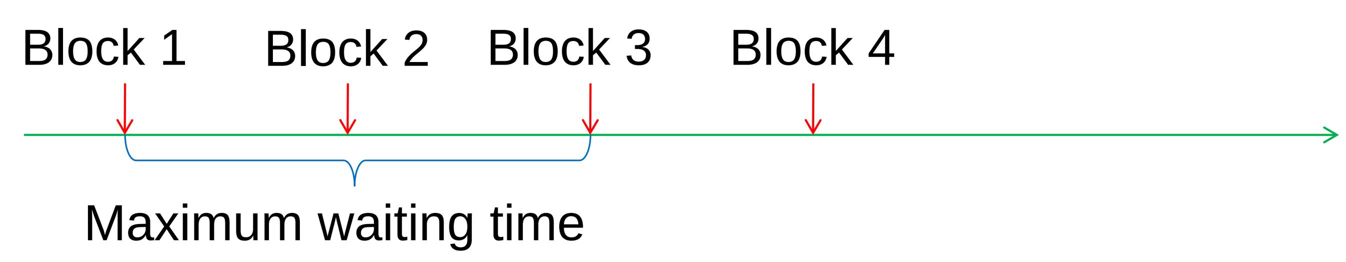




Modeling

Disk 1	Video 1 block 1	Video 2 block 1	Video 4 block 1
Disk 2	Video 1 block 2	Video 2 block 2	Video 4 block 2
Disk 3	Video 1 block 3	Video 3 block 1	Video 4 block 3
Disk 4	Video 1 block 4	Video 3 block 2	Video 5 block 1
Disk 5	Video 6 block 1	Video 3 block 3	Video 5 block 2



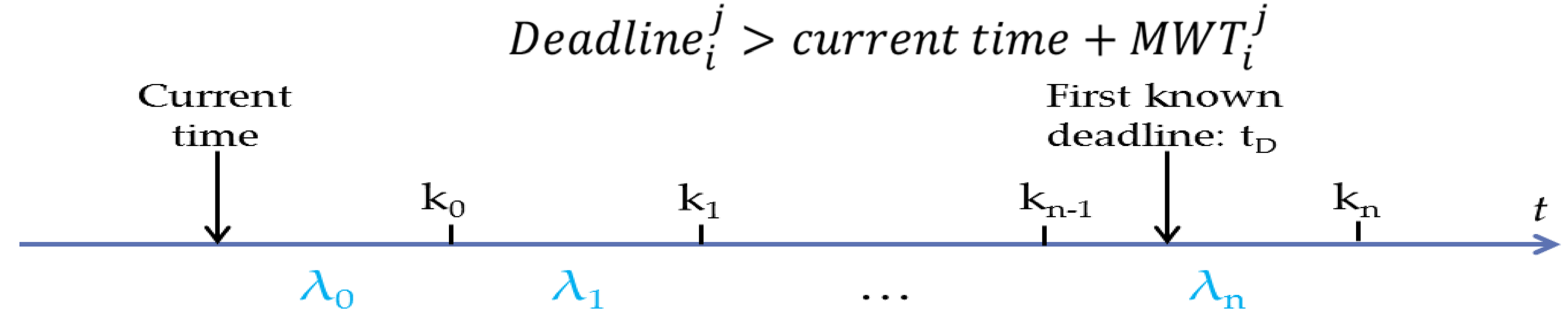


- Known deadlines (already requested)

$$Deadline_i^j = t_{req} + MWT_i^j + AD$$

- Unknown deadlines (future requests)

$$Deadline_i^j > current\ time + MWT_i^j$$



Approach

$$T_s = \min(T_r, t_D);\ \ t_D \in (k_{n-1}, k_n]$$

T_s: sleep time;

T_r: time until the first unknown deadline;

$p(t)$: pdf of T_s

$$p(t) = \begin{cases} \lambda_0 e^{-\lambda_0 t} & \text{if } t \le \min(k_0, t_D) \\ \alpha_i \lambda_i e^{-\lambda_i t} & \text{if } t \in (k_{i-1}, \min(k_i, t_D)] \\ \alpha_n \delta(t - t_D) e^{-\lambda_n t_D} & \text{if } t \ge t_D \end{cases}$$

where $i = 1,2,\ldots,n$ and $\alpha_x = \prod_{j=0}^{x-1} e^{(\lambda_{j+1}-\lambda_j)k_j}$

Energy Cost: $\frac{P_i \cdot T_s + O_i}{T_s + R_i}$

Delay Cost: $D_i = \frac{\lambda_k R_i^2}{2} + \frac{(\lambda_k R_i + n_l)\lambda_k R_i T}{2n_l}$

We want to minimize energy with a constraint on service delay

$$\min\{E\}, \text{subject to } D < D_c$$

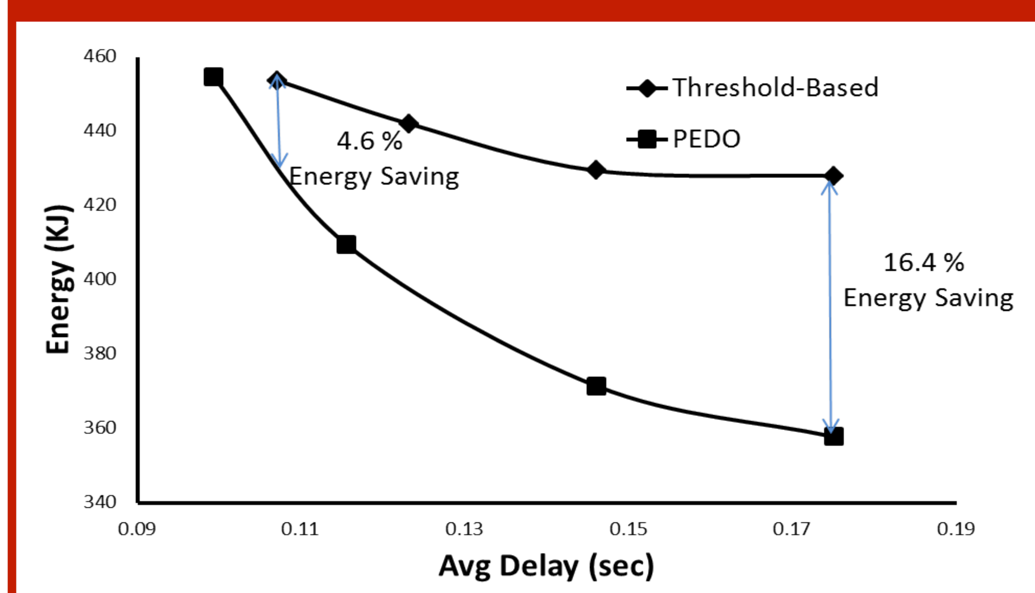
It can be solved using Lagrangian relaxation

$$\min\{C\}, \text{where } C = E + \mu D$$

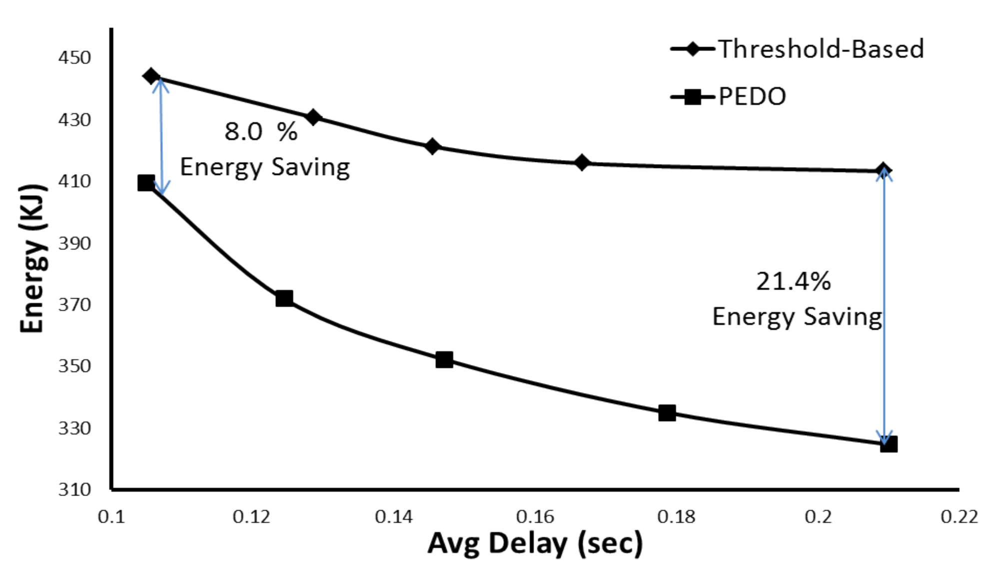
Each μ will solve the optimization for some particular D_c

Optimal mode: $M = \underset{i}{argmin}\ C_i$

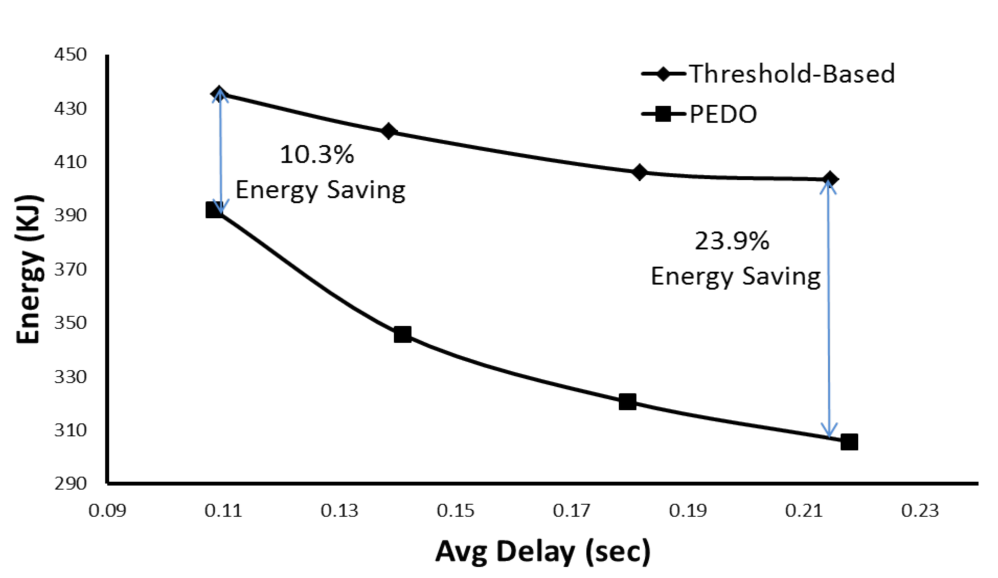
Results



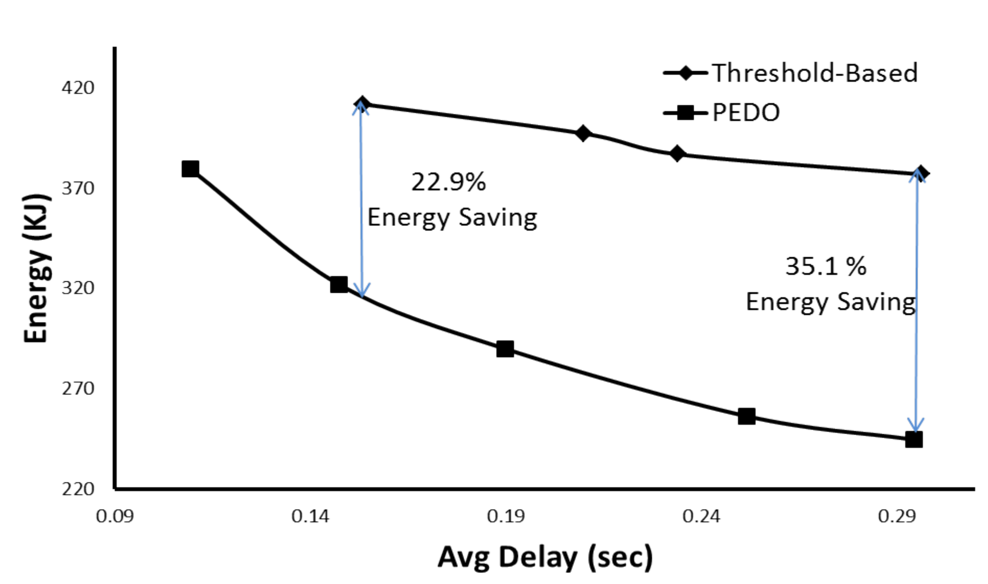
Inter-arrival time: 30 ms



Inter-arrival time: 40 ms



Inter-arrival time: 50 ms



Inter-arrival time: 100 ms

Conclusions

- Conducted the first study of the energy issues for large-scale VSS
- Jointly optimized energy and service delay
- Proposed a model for disk idle time in VSS
- Proposed a prediction-based approach to make optimized power mode decisions
- Future Work
  - Energy-aware caching
  - Energy-aware placement

Contact: hangyuan@usc.edu

Ming Hsieh Institute
Ming Hsieh Department of Electrical Engineering