Brownout - Response Time Control for Cloud Applications

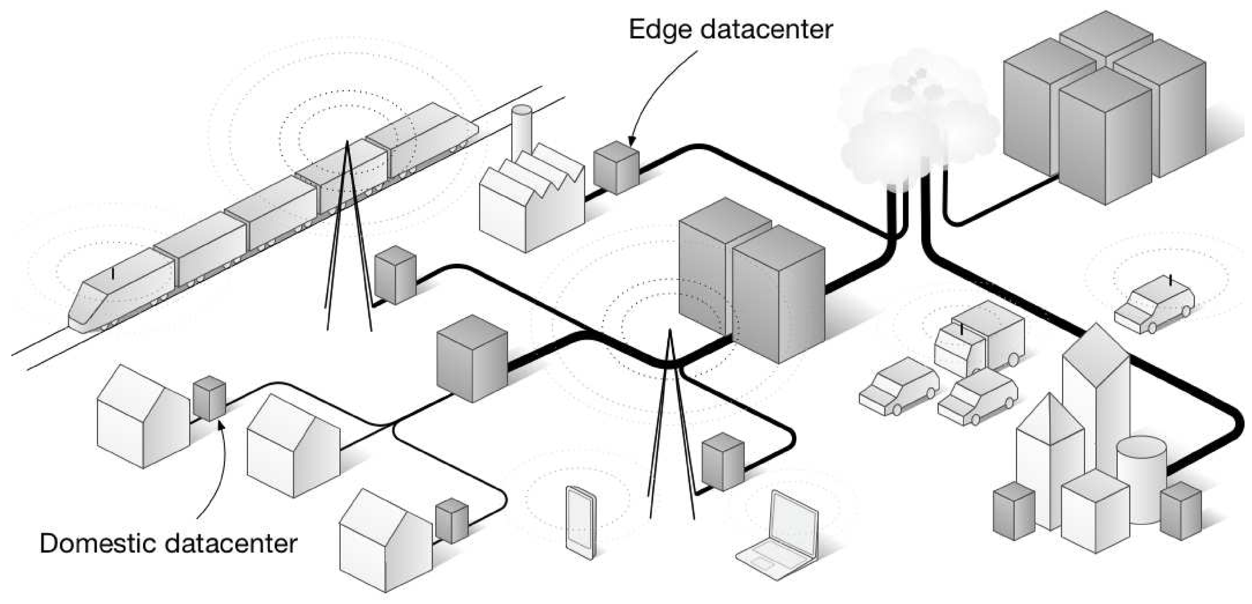
Tommi Nylander, Cristian Klein, Martina Maggio,
Maria Kihl, Karl-Erik Årzén

Lund University · Cloud Control · Umeå Universitet · WASP Wallenberg Autonomous Systems and Software Program

An Autonomous & Distributed Cloud

As part of the Autonomous Cloud project, my research will be focused on:

- Control-based resource management in the cloud
- Meeting the requirements of mission critical applications
- Modeling and control of the distributed cloud infrastructure

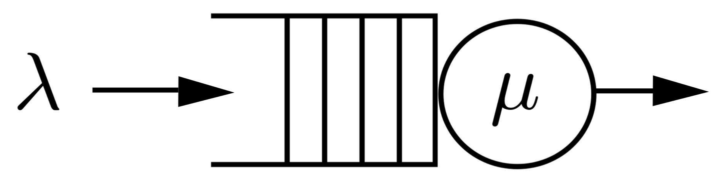


The Brownout Concept

- Introduced to the Cloud community by Klein et. al in 2014 [1]
- Assumption: Cloud applications contain optional content
- Goal: Control response times or maximize revenue?
- Control signal: Dimmer value θ:
  - Probability $0 \leq \theta \leq 1$ to answer with optional content

Modeling of Server Replicas

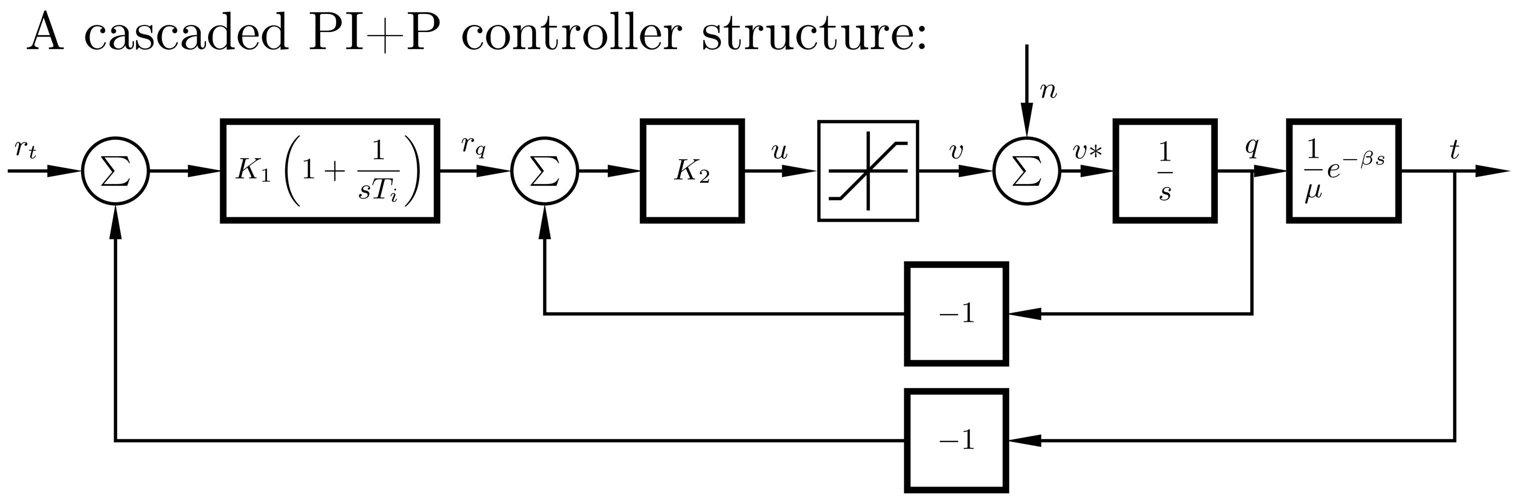
- Goal: Model the influence of dimmer values θ on response times t



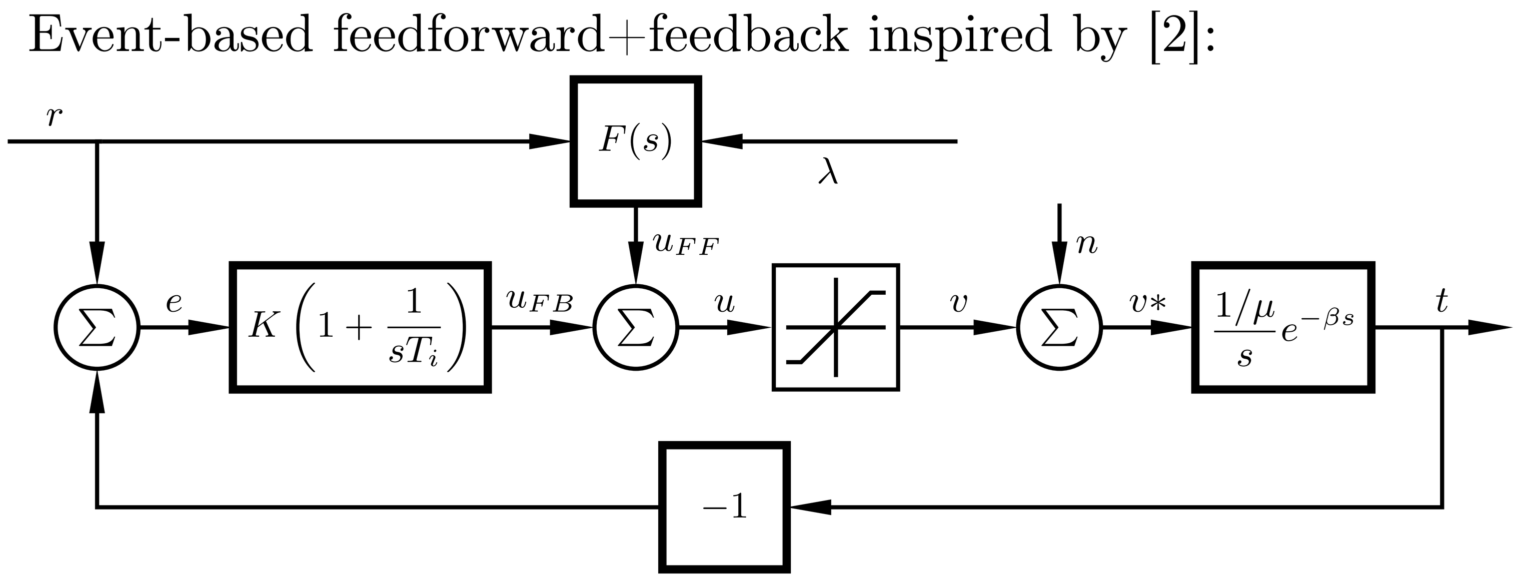
- Continuous time flow model approach: $\frac{dq}{dt} = \lambda - \mu = v$
- Linear model: $T(s) = \frac{1/\mu}{s} e^{-\beta s} V(s)$
- Dimmer value θ determines service time: $\frac{1}{\mu} = \theta t_f + (1-\theta)\, t_r$
- Transform: $\theta = \left(\frac{1}{\lambda - v} - t_r\right) \Big/ (t_f - t_r)$

Control Strategies

- Simple PI controller: $C(s) = K\left(1 + \frac{1}{sT_i}\right)$
- A cascaded PI+P controller structure:



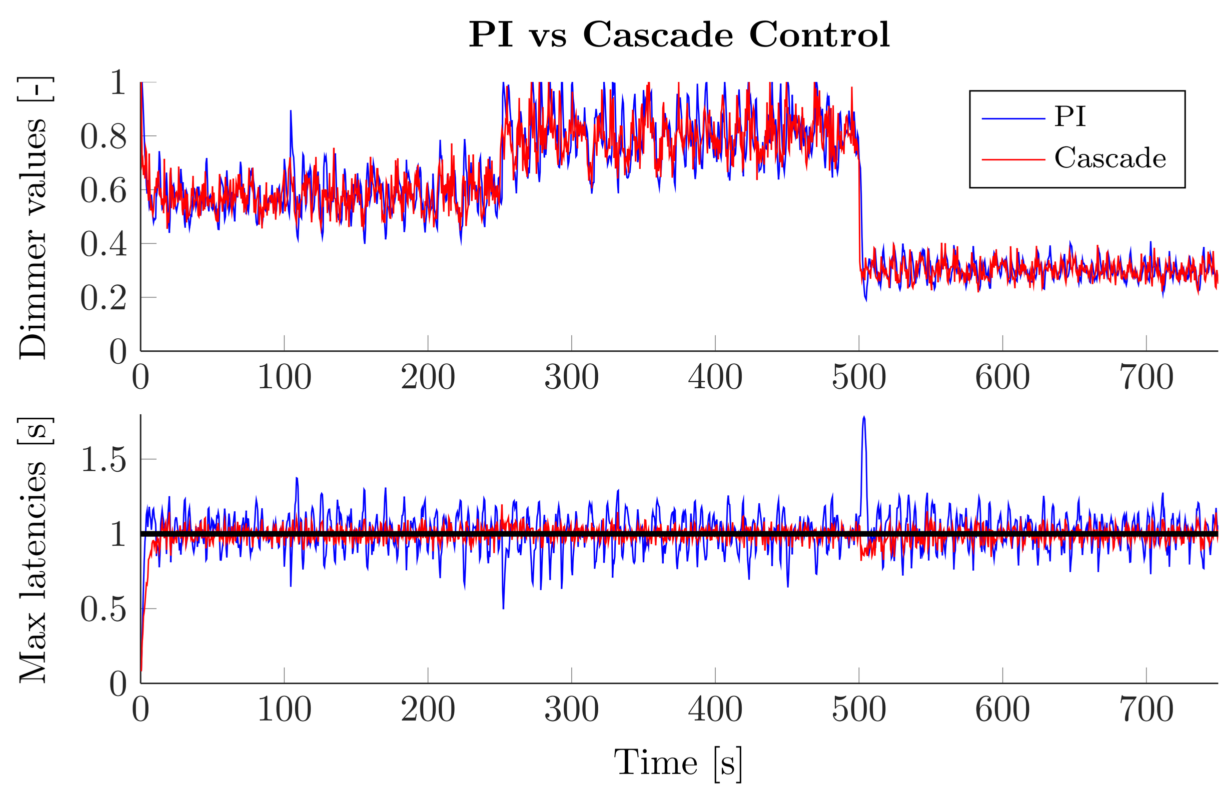
- Event-based feedforward+feedback inspired by [2]:

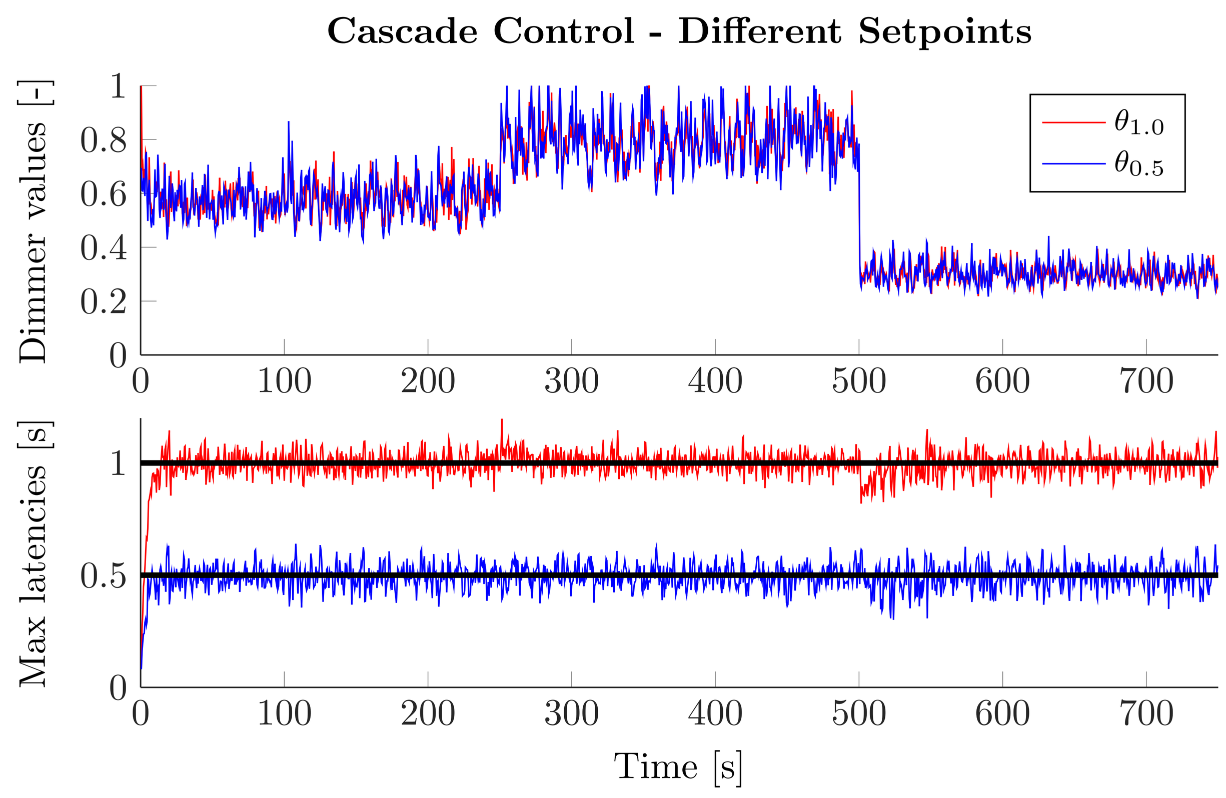


- Model Predictive Control: Minimize cost $J = \int Q_1\theta^2 + Q_2 t^2 dt$ subject to the dynamics and constraints

Preliminary Simulation Results

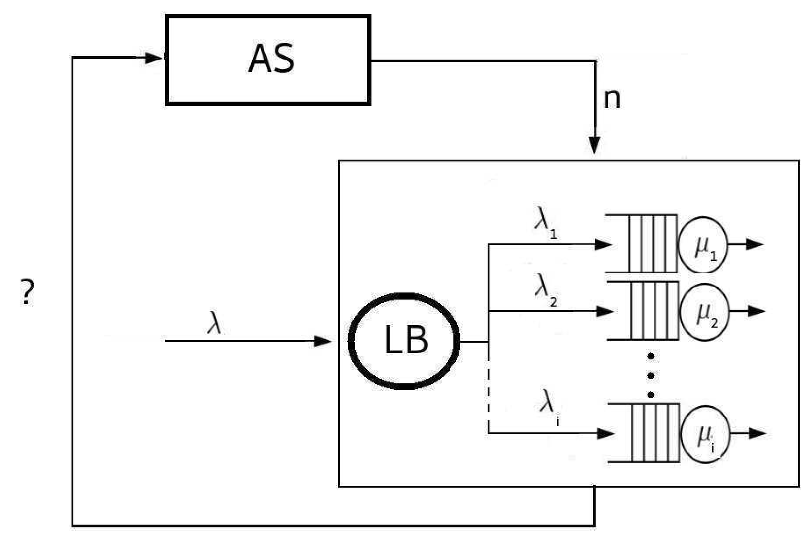
- Simulations performed in a discrete-event simulator
  - Written in Python by Cristian Klein and Martina Maggio
  - Open loop clients: $\bar{\lambda}_1 = 250/s$, $\bar{\lambda}_2 = 180/s$, $\bar{\lambda}_3 = 460/s$





Future Work

- Simulate event-based FF+FB and MPC controllers
- Test all controllers in a real hardware setup:
  - Extensions of cloud benchmarks RUBiS and RUBBoS
  - Open or closed loop clients?



- Autoscaling of multiple replicas:
  - Control signal: Desired number of replicas n
  - Measurement signal: Average dimmer values $\bar{\theta}$?
  - Distributed replicas: Where to deploy the resources?

References

[1] C. Klein, M. Maggio, K. E. Årzén, and F. Hernández-Rodriguez. "Brownout: Building More Robust Cloud Applications". In: *ICSE* 2014.

[2] D. Henriksson, Y. Lu, and T. Abdelzaher. "Improved Prediction for Web Server Delay Control". In: *ECRTS* 2004.