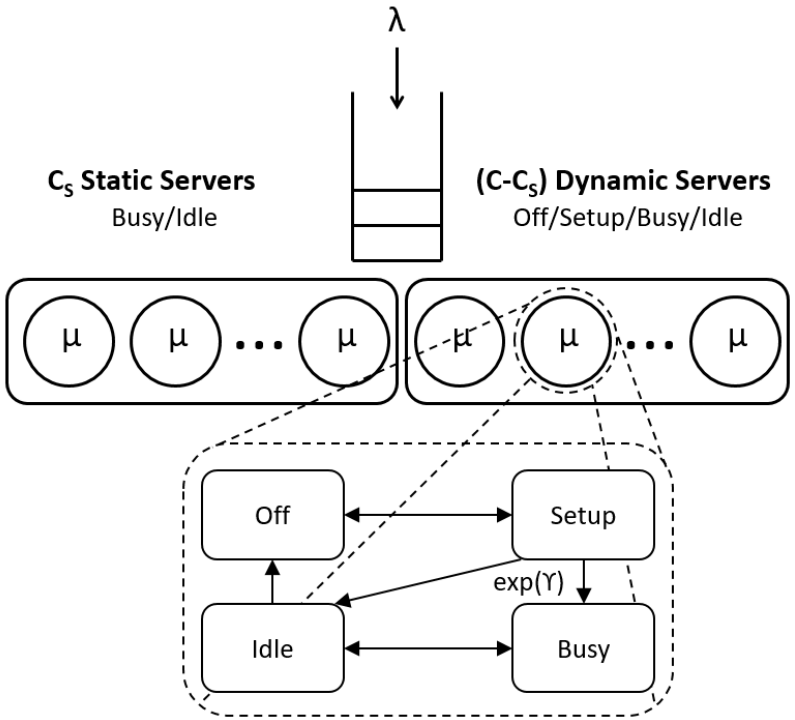
Asymptotic Bounds for Energy-Aware Queueing Systems with Setup Times

Vincent J. Maccio, Douglas G. Down

Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada

Model

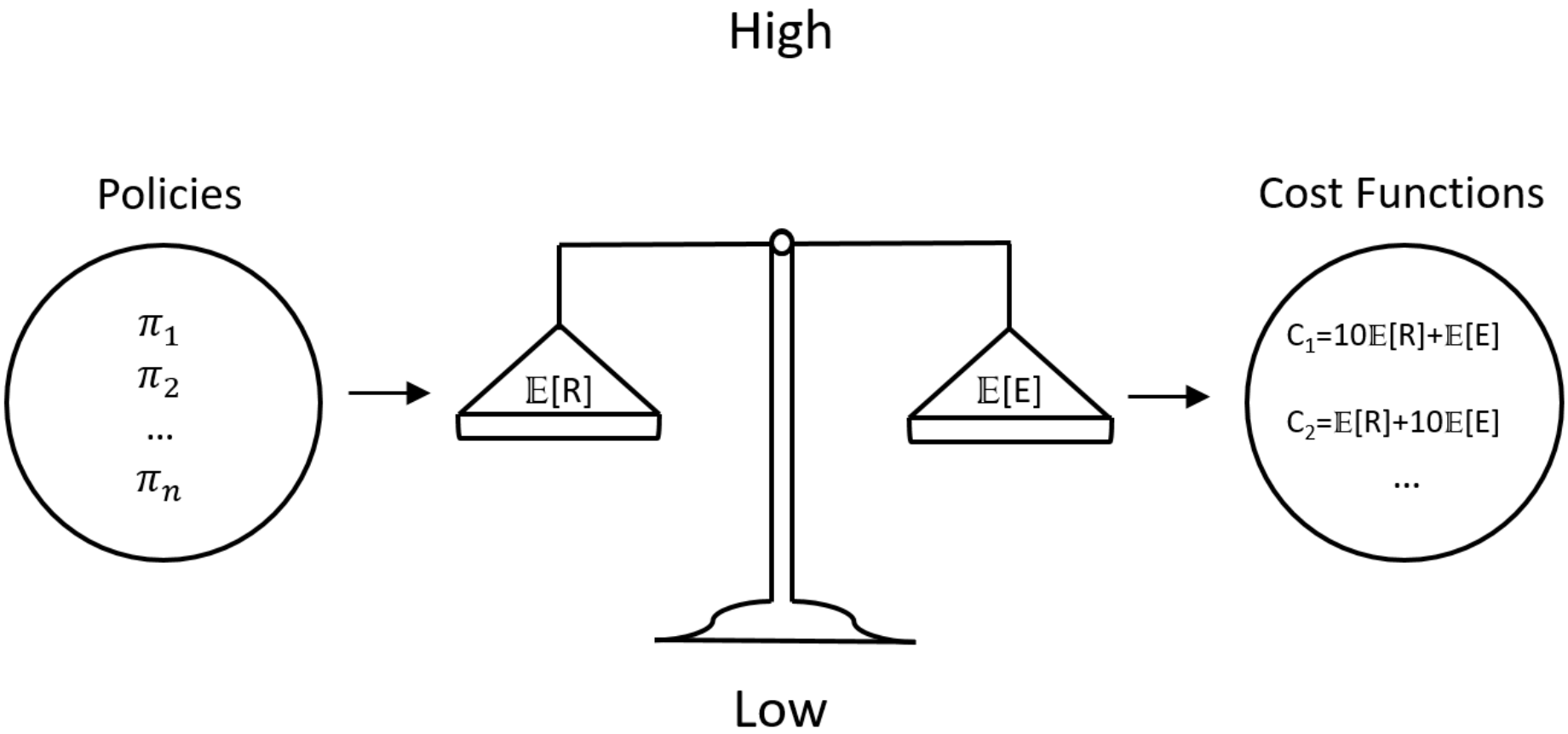
- Jobs arrive following a Possion process with rate λ
- Processing times are exponentially distributed with rate μ
- C homogeneous servers
- Setup times are exponentially distributed with rate γ
- C_S servers will always remain on



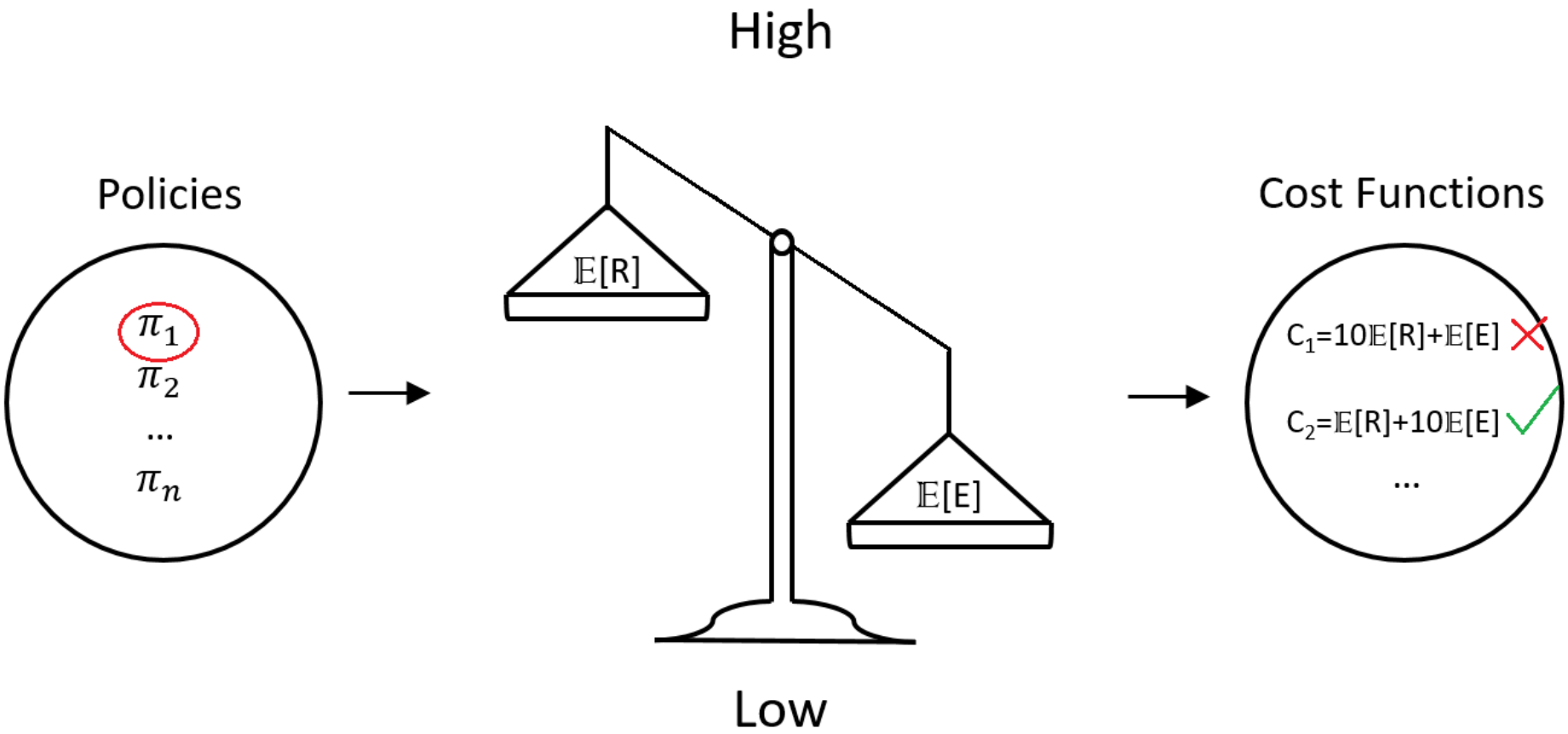
Metrics and Costs

- Why turn a server off? Why keep a server on?
- The expected response time $\mathbb{E}[R]$ and expected energy cost $\mathbb{E}[E]$
  - A weighted sum of the expected number of servers multiplied with a corresponding cost:
    $\mathbb{E}[E] = E_{Busy}\mathbb{E}[C_{Busy}] + E_{idle}\mathbb{E}[C_{idle}] + E_{setup}\mathbb{E}[C_{setup}]$
- We call a cost function *Well Formed*, if it is non-decreasing in and only dependent on $\mathbb{E}[R]$ and $\mathbb{E}[E]$
  - Example cost functions: $\mathbb{E}[E]\mathbb{E}[R]$ or $\mathbb{E}[R] + \beta\mathbb{E}[E]$
- Minimize the cost function via a system policy

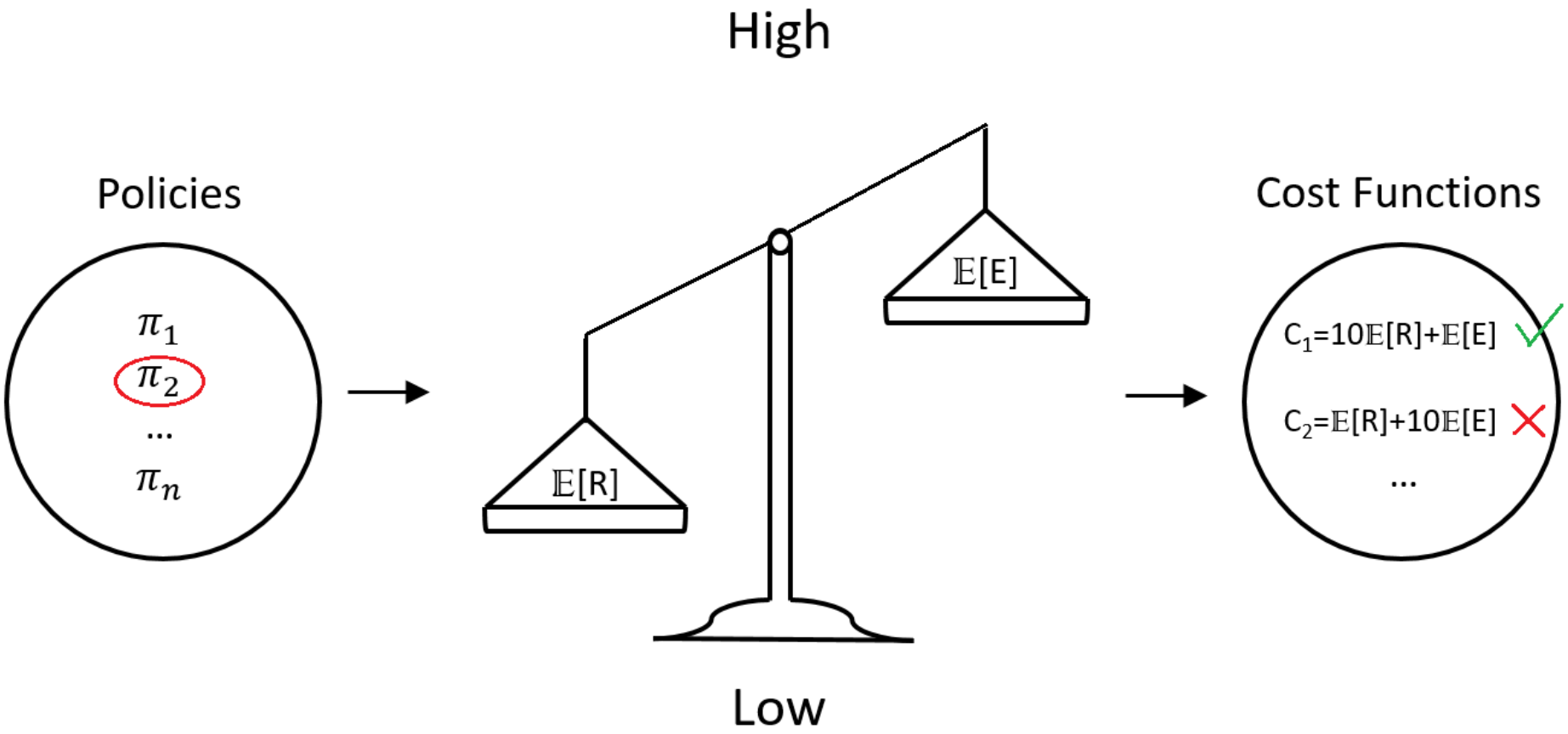
Potential Difficulties



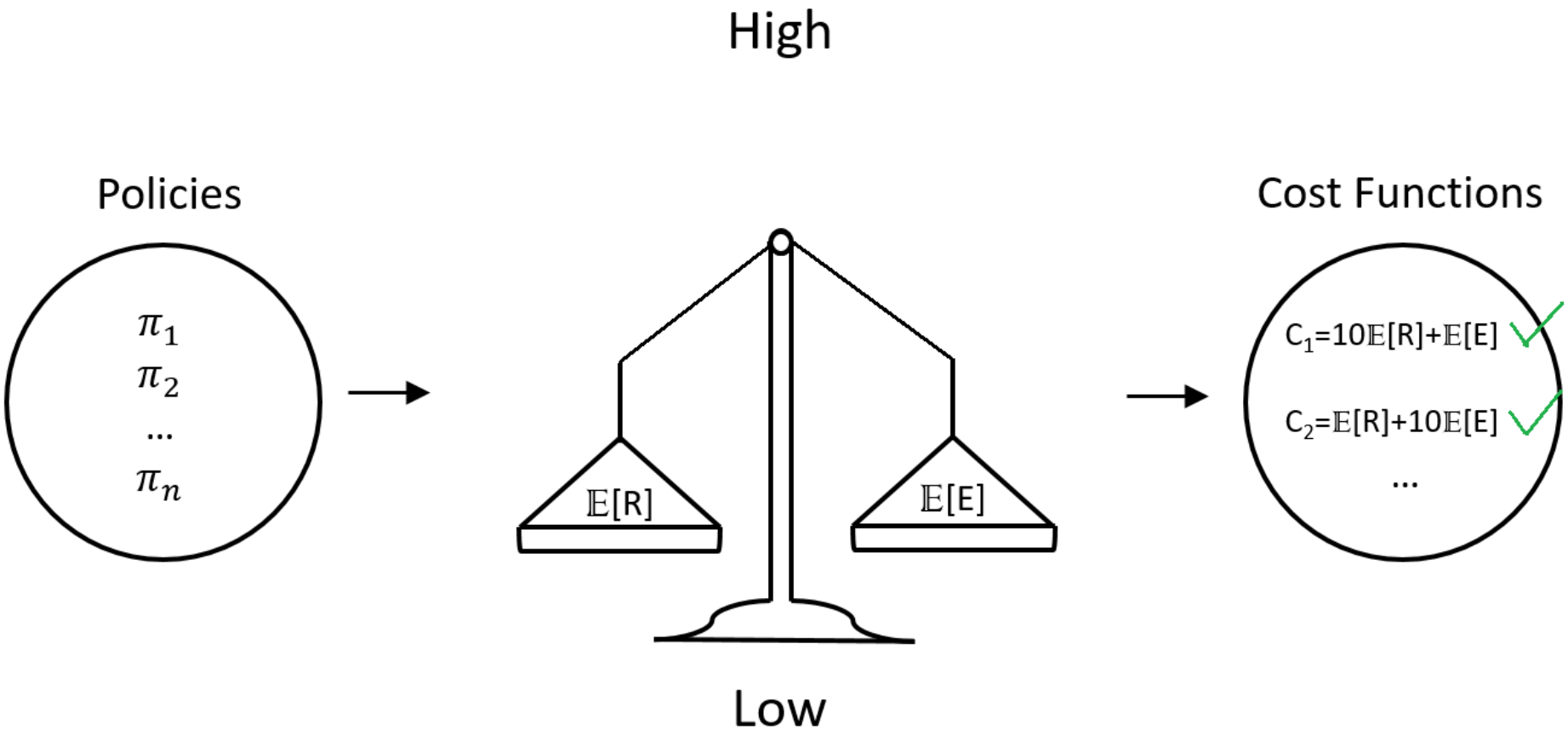
Potential Difficulties



Potential Difficulties



Potential Difficulties



Asymptotic Regime

- Many servers regime - the number of servers grows, but the load on the system remains fixed:
  - $C \to \infty$, $\lambda \to \infty$, $\rho = \lambda / C\mu$, where $0 < \rho < 1$

C	λ	μ	ρ
10	5	1	0.5
100	50	1	0.5
1000	500	1	0.5
∞	$\infty/2$	1	0.5

- Difficult to determine the degree to which servers are switching on and off; does the system thrash?
- Note: heavy traffic regimes are not of particular interest here, as this would essentially result in all servers being always on.

Class A Policy

- A **Class A** policy adheres to a threshold scheme for server setups
  - Shown to be optimal in Maccio and Down (2015)
  - $\Theta(C^2)$ thresholds to determine
- A **Class A** policy never turns a server off if there is a job it could be processing
  - Shown to be optimal in Maccio and Down (2015)
- The vast majority of policies studied in the literature are **Class A** policies

Result 1

Theorem

All **Class A** *policies asymptotically minimize the expected response time.*

$$\mathbb{E}[R] \to 1/\mu.$$

Class B Policy

- A **Class B** policy is a **Class A** policy.
- There exists an $\alpha < 1$ such that $C_S < C\rho + C^{\alpha}$.
  - Avoids too many servers being always on - this would result in guaranteed suboptimal energy costs
- If the ith server turns off when there are j jobs in the system, then the ith servers will not begin turning on until there's at least $j + 1$ jobs in the system
  - Shown to be optimal in Maccio and Down (2015)
- Again, the vast majority of policies studied in the literature are **Class B** policies

Class B policy

Theorem

All **Class B** *policies asymptotically minimize the expected energy cost, i.e. for any* **Class B** *policy* π_B,

$$\mathbb{E}[E^{\pi_B}]/\lambda \to E_{Busy}/\mu.$$

Corollary

All **Class B** *policies asymptotically minimize all well-formed cost function.*

Key implication: there is no tradeoff (at least asymptotically) between expected response time and expected energy costs.

What's Going On?

- Do no jobs have to wait? Are there no servers turning on? Are there no setup costs?
- No, in fact, an infinite number of jobs have to wait for a finite amount of time, and there are infinite costs associated with idling and setup server
- However, the proportion of jobs that have to wait to be served is negligible
- Likewise, the proportion of energy costs contributed by idling and setups if also negligible

What About Finite C?

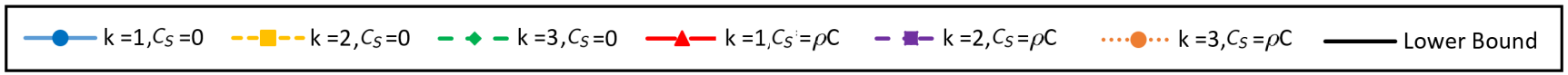
- How large does C need to be before this optimal behaviour is seen
- Examine via numerical experiments
  - Exact analysis of the underlying CTMC

The Staggered Threshold Policy

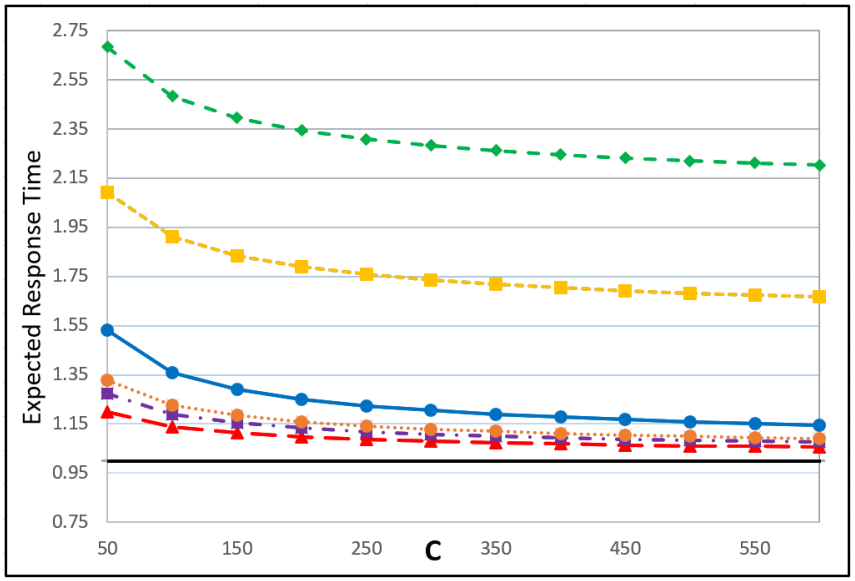
Choose a C_S and k:

- C_S servers always remain on
- Dynamic servers turn off when they idle
- The ith dynamic server will begin its setup when there are $ik + C_S$ or more jobs in the system

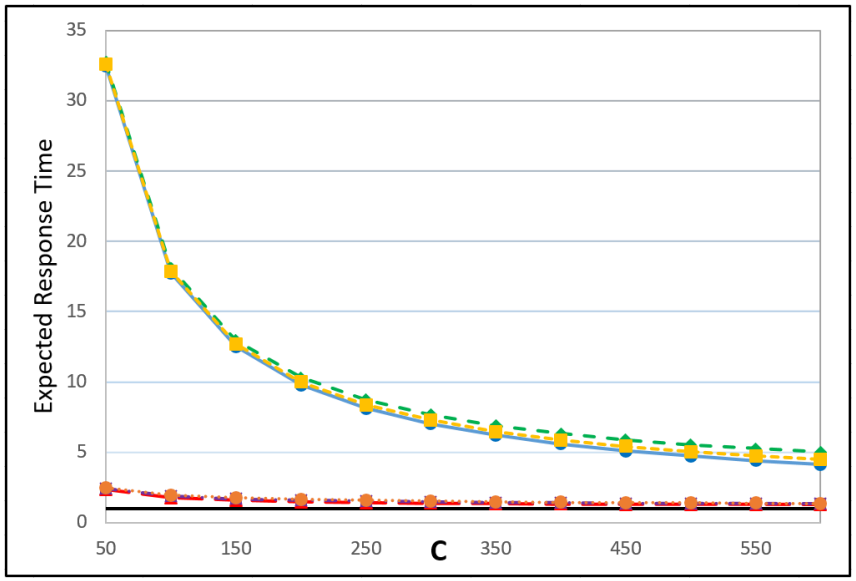
Numerical Results - Response Times



$\lambda = C/2,\ \mu = 1$

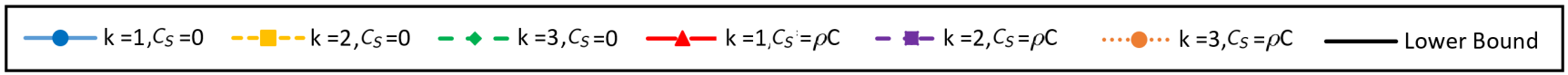


$\gamma = 0.01$

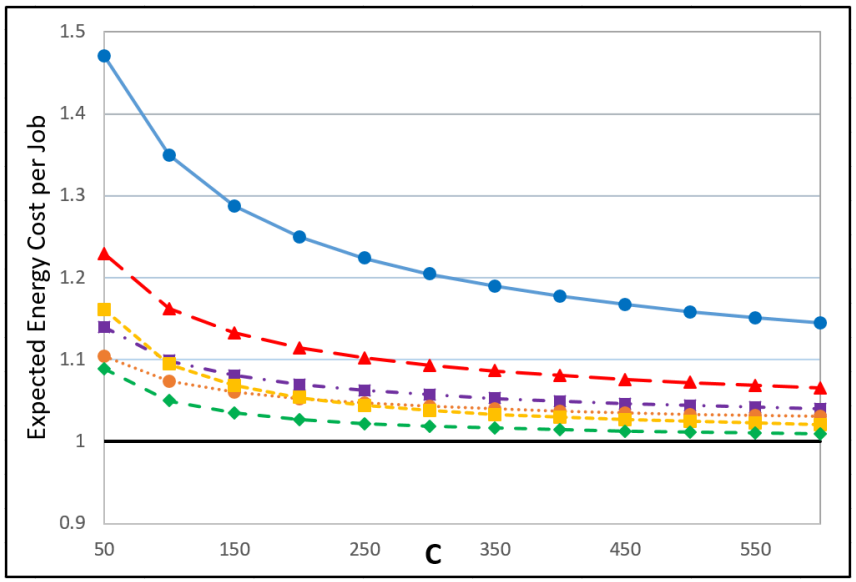


$\gamma = 0.0001$

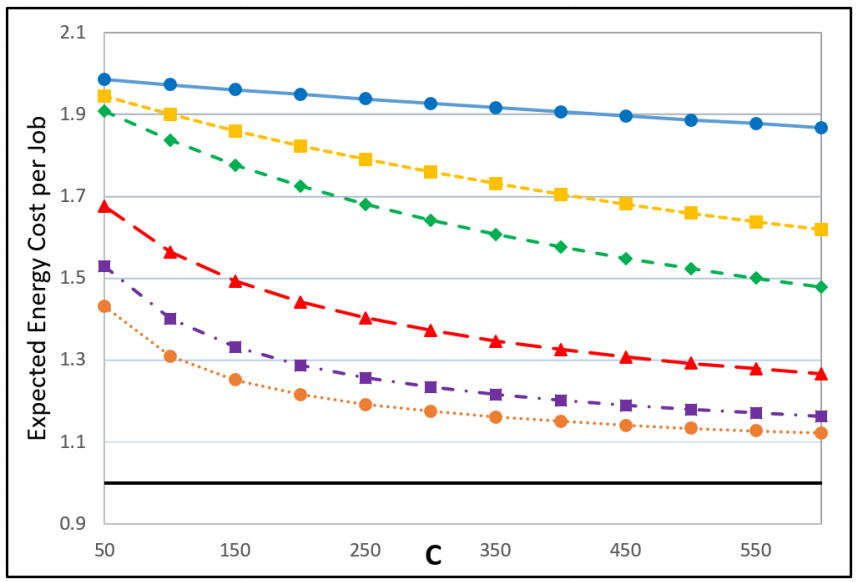
Numerical Results - Energy Costs



$$\lambda = C/2,\ \mu = 1,\ E_{Busy} = E_{setup} = 1,\ E_{idle} = 0.7$$



$\gamma = 0.01$



$\gamma = 0.0001$

Conclusion

- Under a fixed-load-many-server regime, a large (infinite) set of policies become equivalent and optimal for any well-formed cost function
- This asymptotic behaviour can be induced earlier by appropriately keeping some servers always on

Future work:

- Relaxing assumptions - distributional and also fixed arrival rate
- Convergence rates?