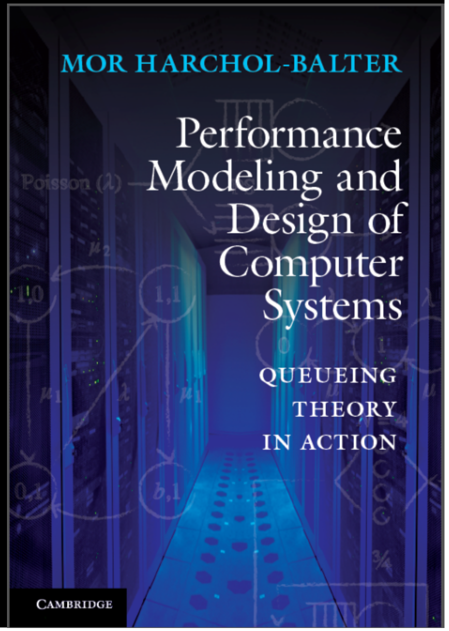
Power Management in Data Centers: Theory & Practice

Mor Harchol-Balter
Computer Science Dept
Carnegie Mellon University

Anshul Gandhi, Sherwin Doroudi,
Alan Scheller-Wolf, Mike Kozuch



Power is Expensive

Annual U.S. data center energy consumption

||

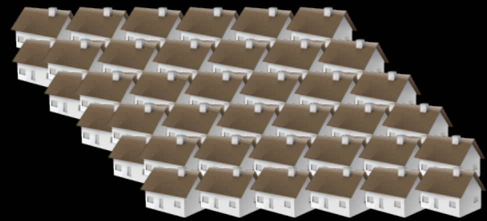
100 Billion kWh or 7.4 Billion dollars

||

Electricity consumed by 9 million homes

||

As much CO_2 as all of Argentina



Sadly, most of this energy is wasted

[energystar.gov], [McKinsey & Co.], [Gartner]

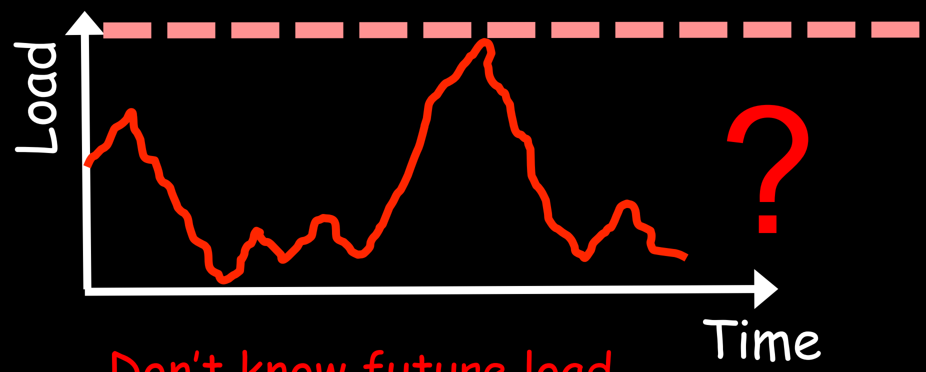
Most Power is Wasted

Servers only busy 5-30% time on average,
but they're left ON, wasting power. *[Gartner Report] [NYTimes]*

Setup time 260s 200W

- BUSY server: 200 Watts
- IDLE server: 140 Watts
- OFF server: 0 Watts

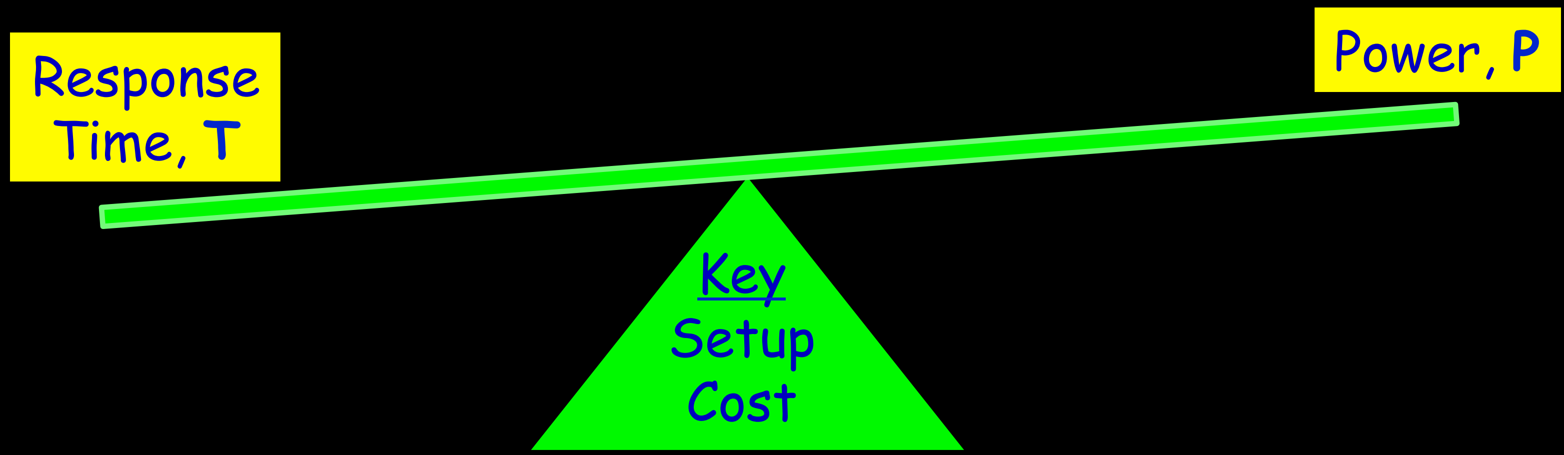
Intel Xeon E5520
2 quad-core 2.27 GHz
16 GB memory



ALWAYS ON:
Provision for Peak

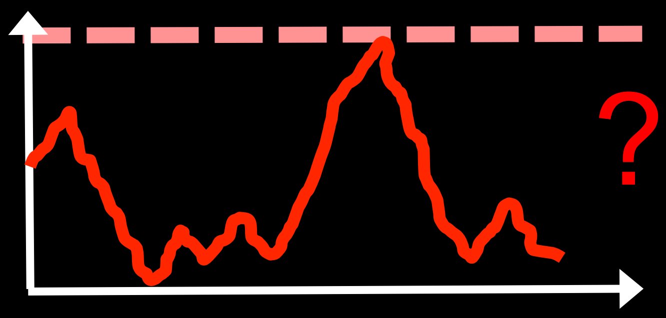
Don't know future load

Talk Thesis



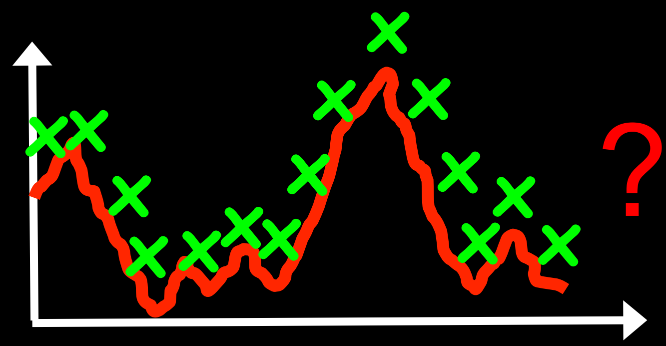
ALWAYS ON

+ Low response time
- Wastes power



ON/OFF

- High response time
+ Might save power



Outline

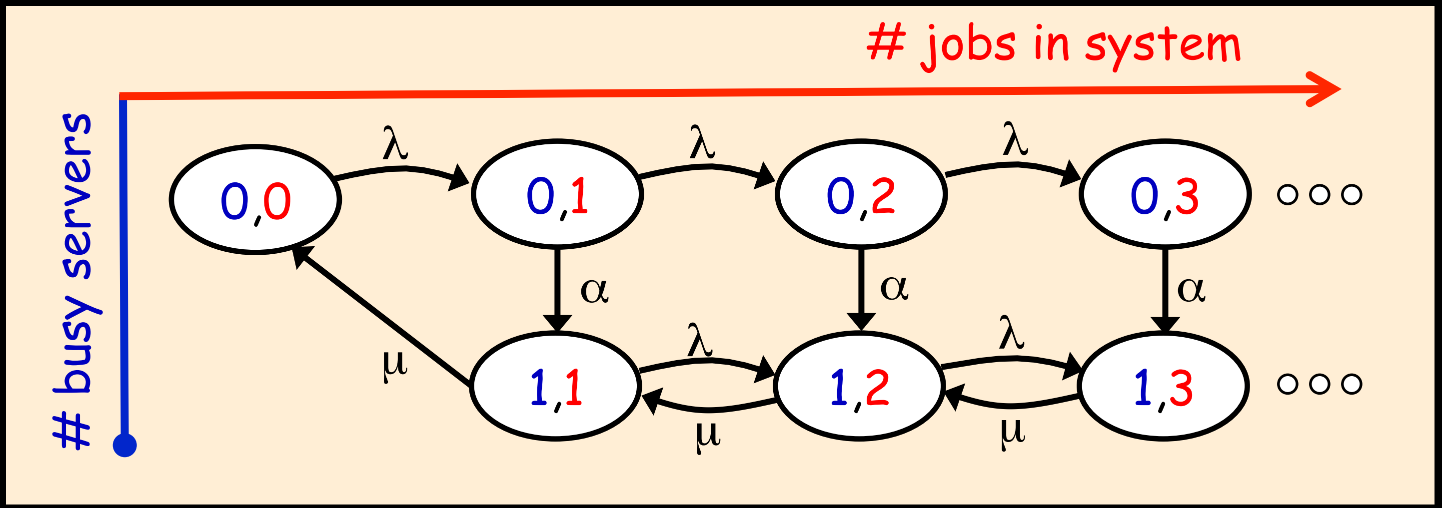
- Part I: Theory - M/M/k
  -- What is the effect of setup time?

- Part II: Systems Implementation
  Dynamic power management in practice

M/M/1/Setup

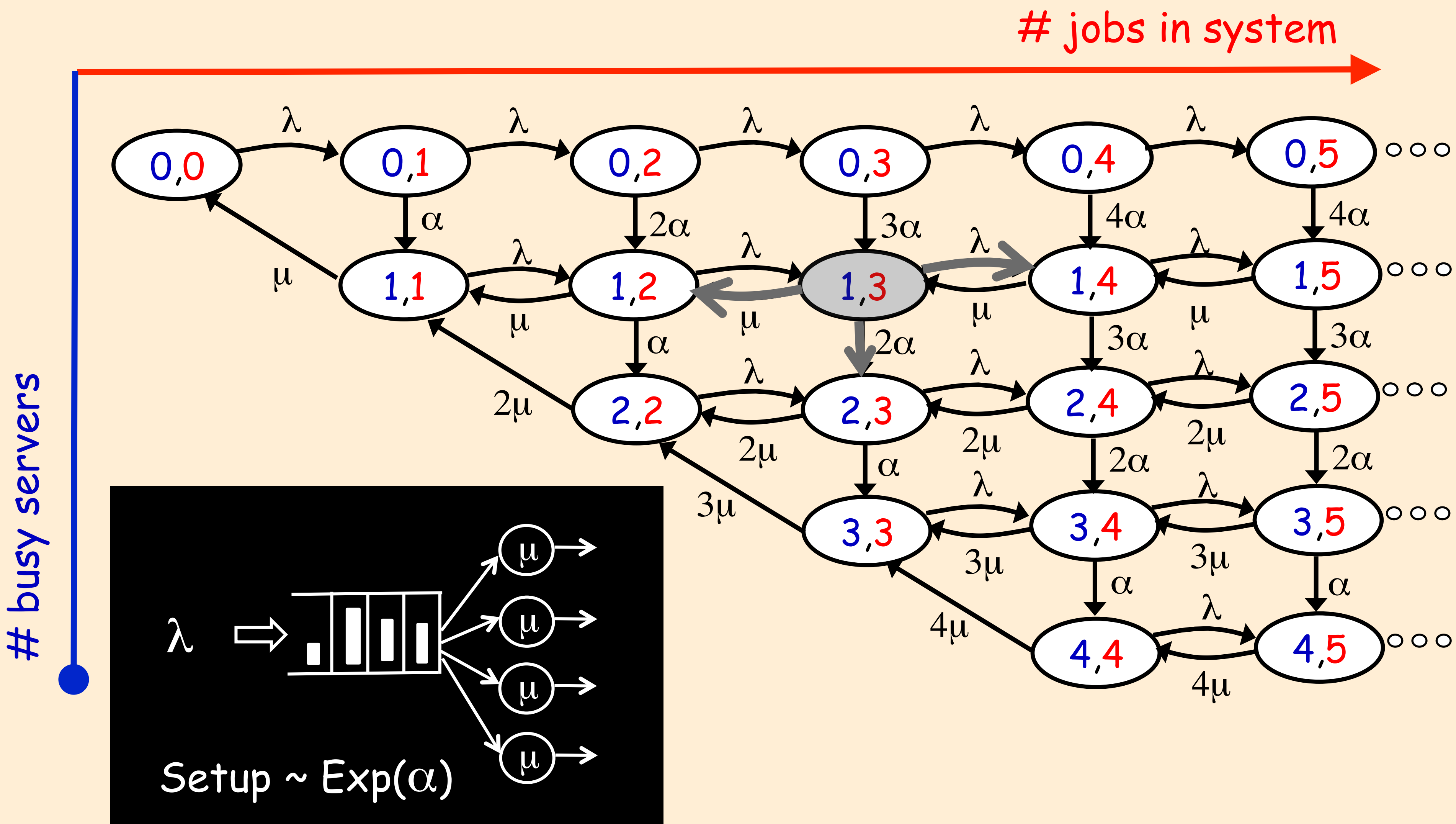
Server turns off when idle.

Setup ~ Exp(α)



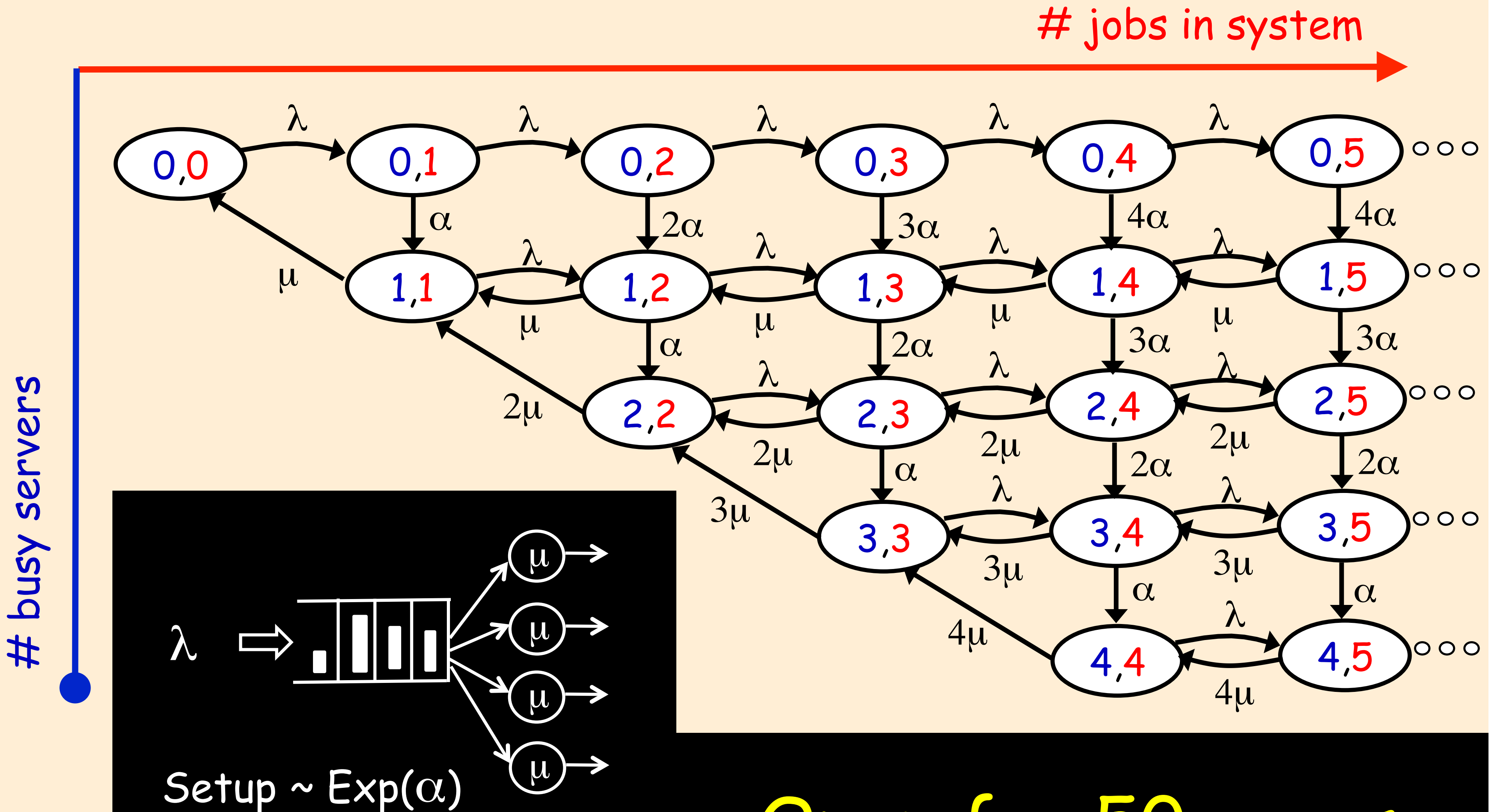
[Welch '64] $E[T^{M/M/1/Setup}] = E[T^{M/M/1}] + E[Setup]$

M/M/k/Setup (k=4)



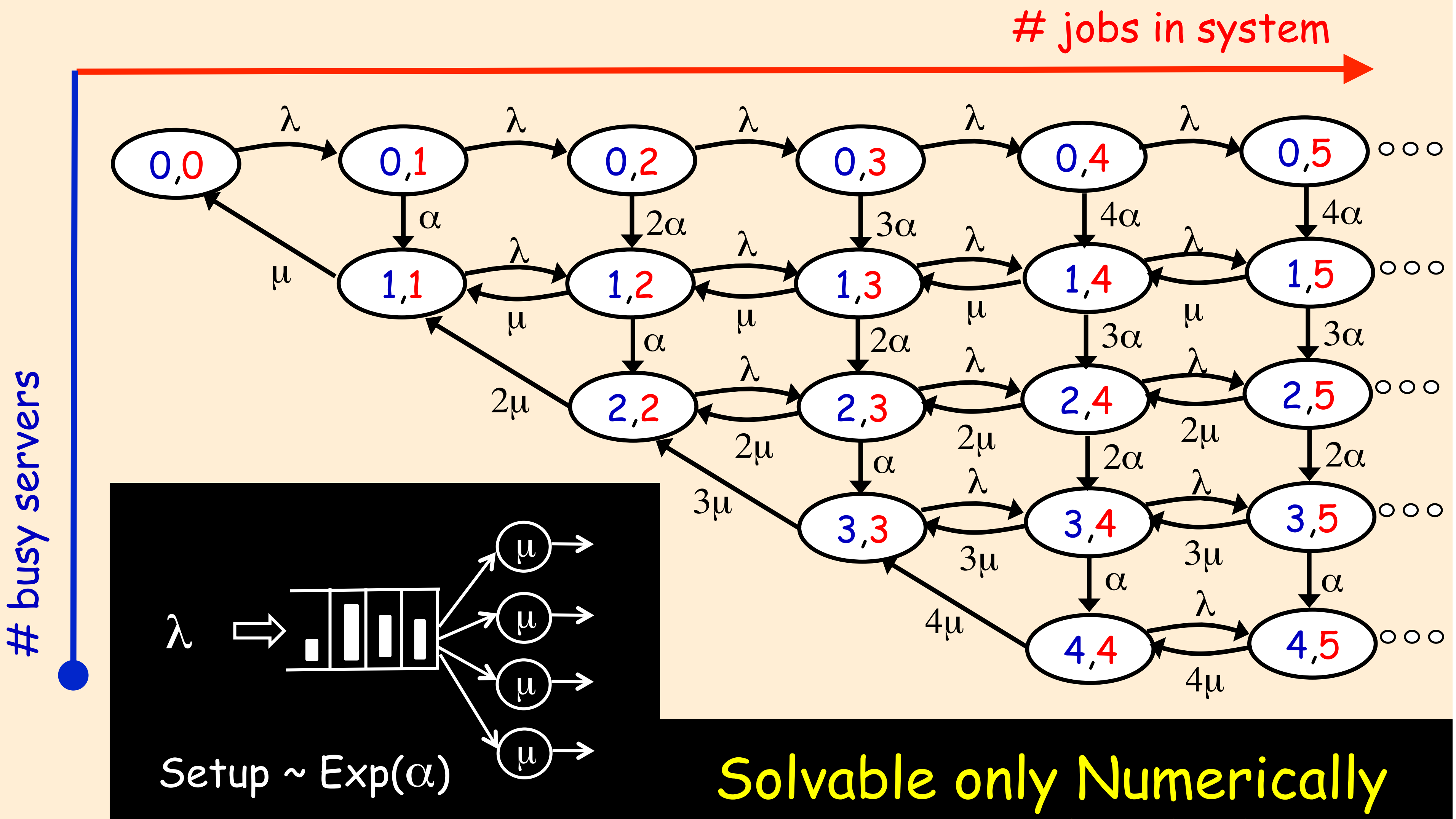
jobs in system
busy servers
0,0
0,1
0,2
0,3
0,4
0,5
1,1
1,2
1,3
1,4
1,5
2,2
2,3
2,4
2,5
3,3
3,4
3,5
4,4
4,5
λ
μ
2μ
3μ
4μ
α
2α
3α
4α
Setup ~ Exp(α)

M/M/k/Setup (k=4)



Open for 50 years

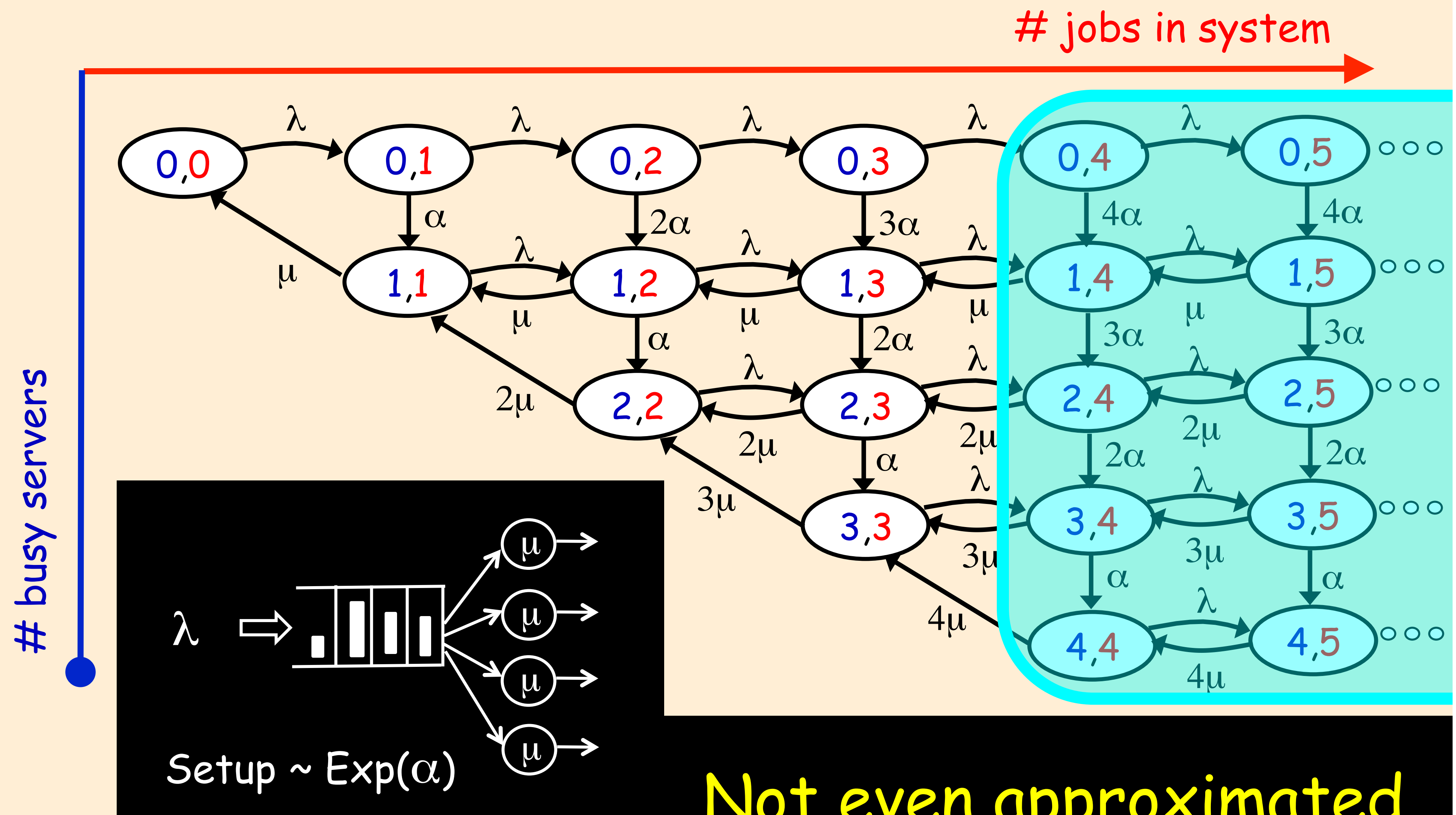
M/M/k/Setup (k=4)



Solvable only Numerically

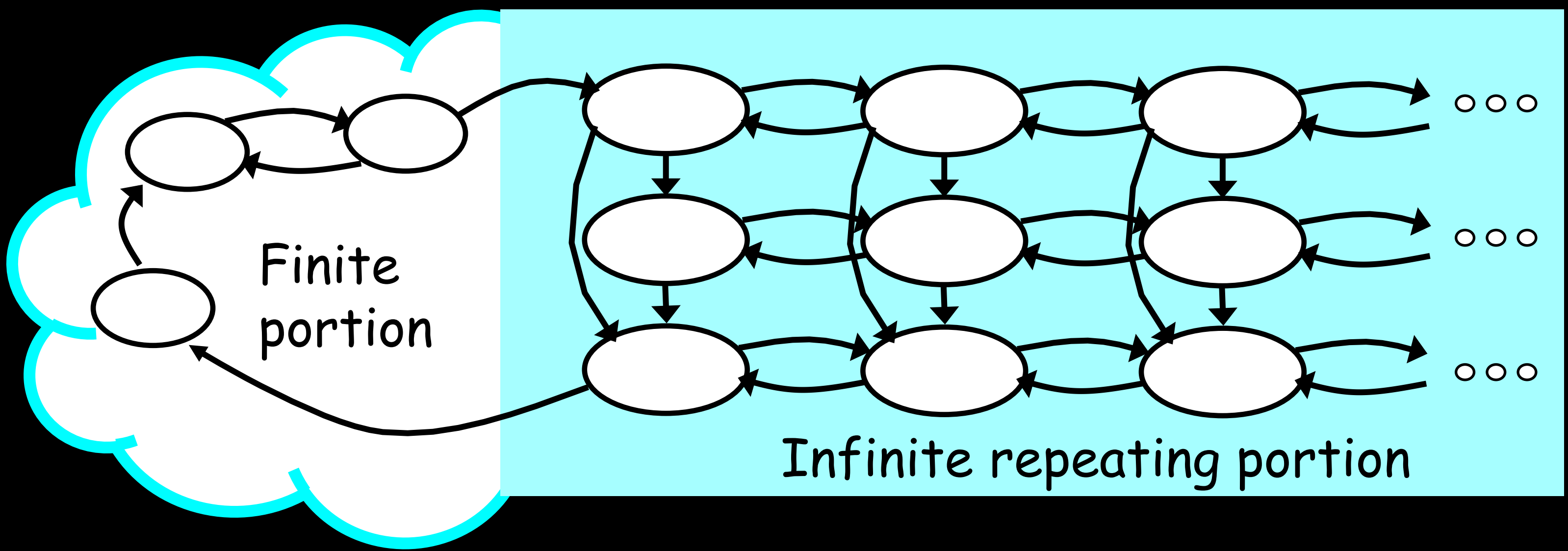
Matrix-Analytic (MA)

M/M/k/Setup (k=4)



Not even approximated

New Technique: RRR [Sigmetrics 13]

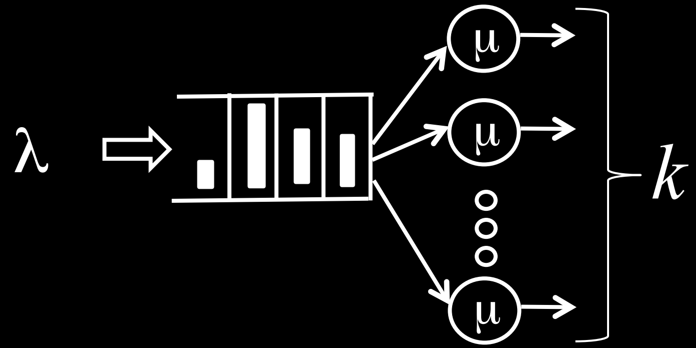


Recursive Renewal Reward (RRR)

- Exact. No iteration. No infinite sums.
- Yields transforms of response time & power.

Closed-form for all chains that are skip-free in horizontal direction and DAG in vertical direction.

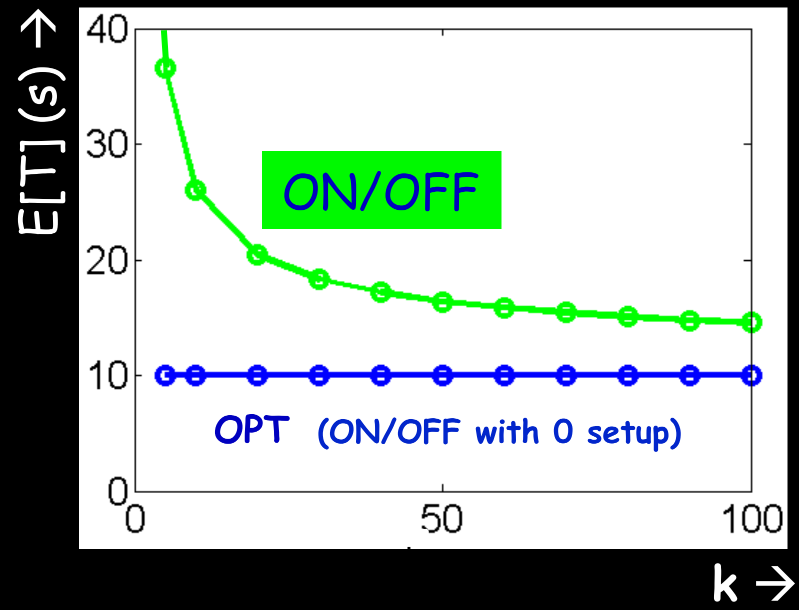
Results of Analysis

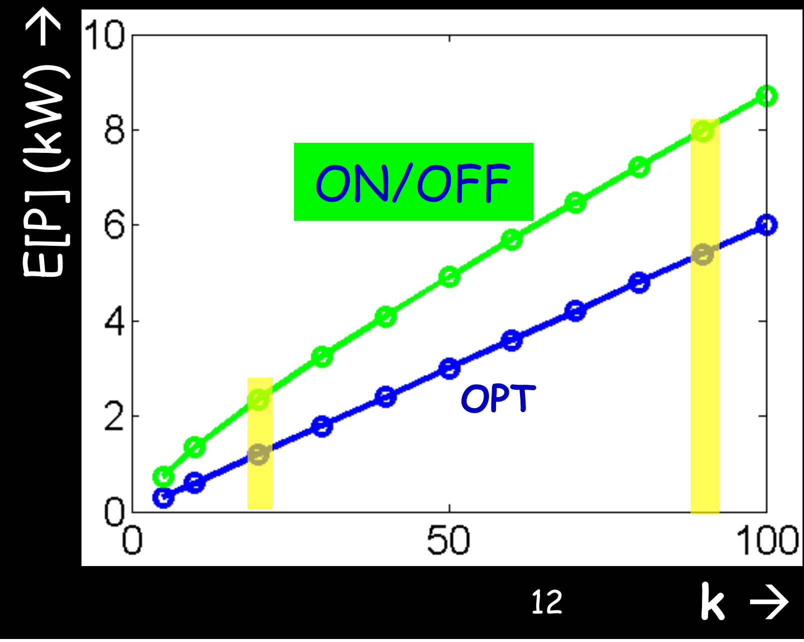


$$E[\text{Job size}] = 10s$$

$$E[\text{Setup}] = 100s$$

$$\text{fix utilization} = \lambda / k\mu = 30\%$$



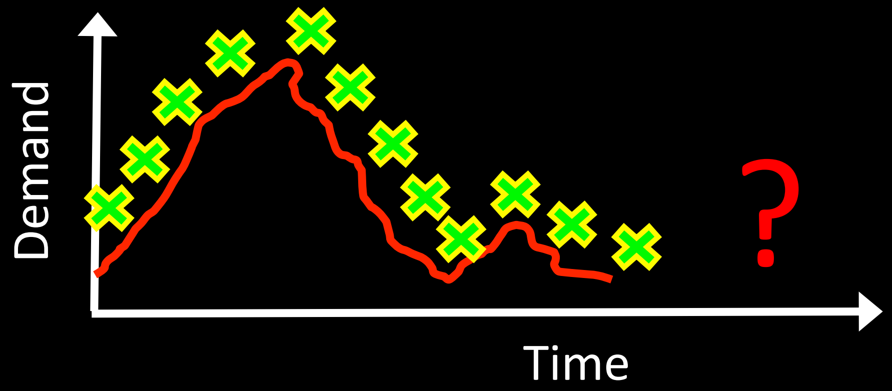


Outline

- Part I: Theory - M/M/k
  What is the effect of setup time?

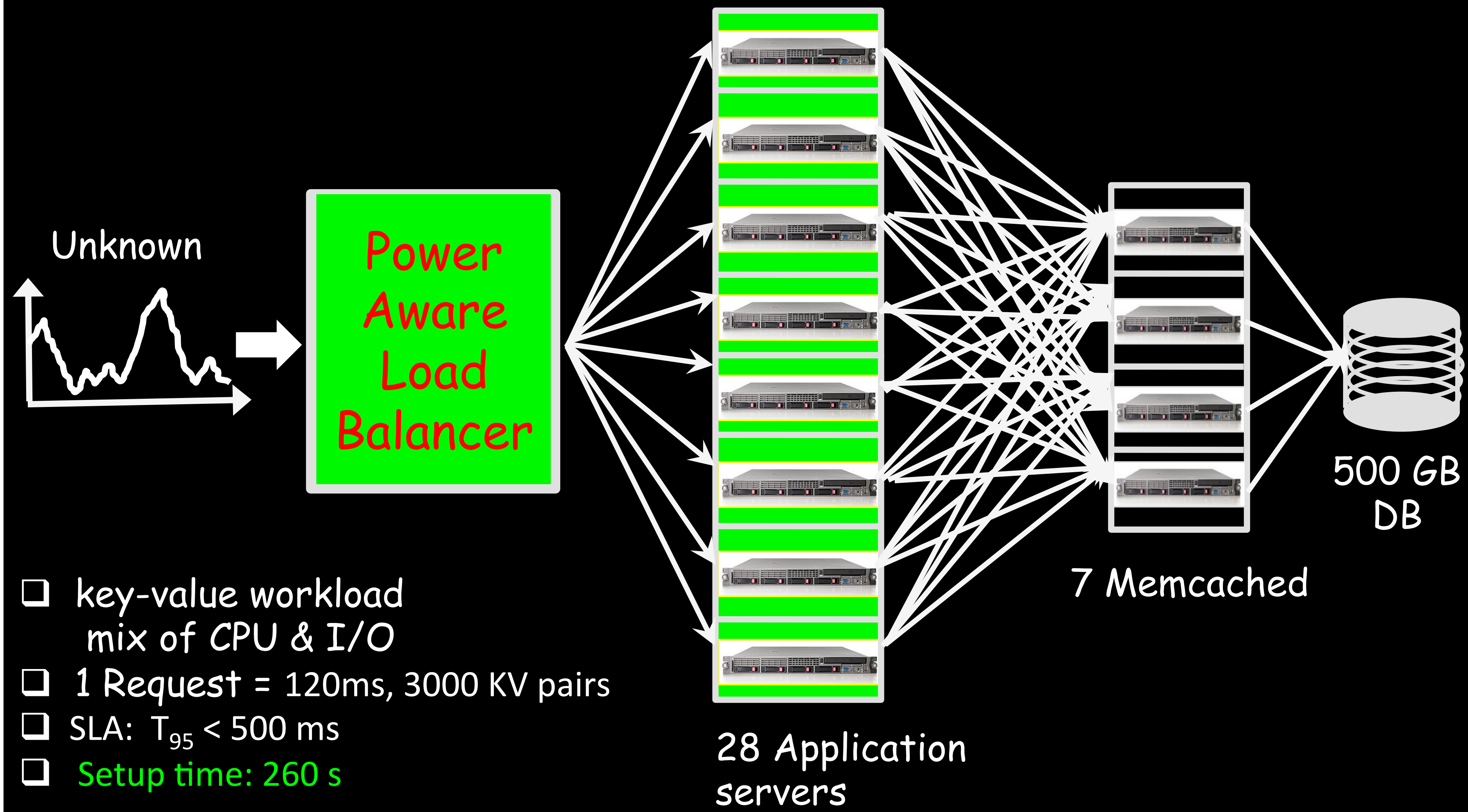
  -- Setup hurts a lot when k: small
  -- But setup much less painful when k: large
  -- ON/OFF allows us to achieve near optimal power

- Part II: Systems Implementation
  Dynamic power management in practice



-- Arrivals: NOT Poisson
   Very unpredictable!
-- Servers are time-sharing
-- Job sizes highly variable
-- Metric: $T_{95} < 500$ ms
-- Setup time = 260 s

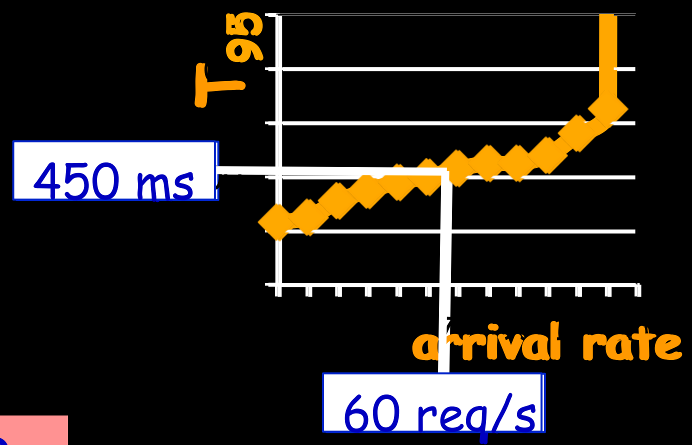
Our Data Center



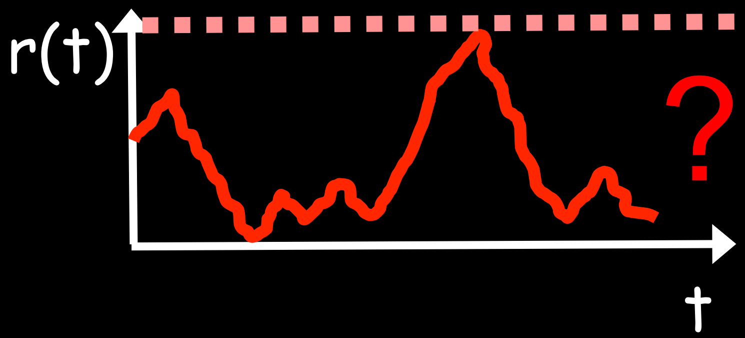
- key-value workload mix of CPU & I/O
- 1 Request = 120ms, 3000 KV pairs
- SLA: T_{95} < 500 ms
- Setup time: 260 s

Provisioning

At Single Server

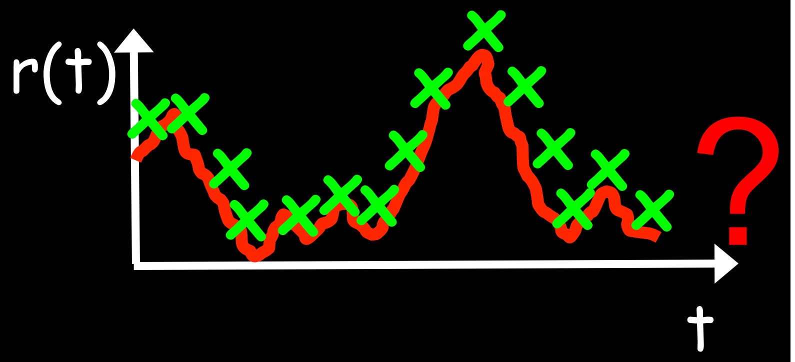


AlwaysOn



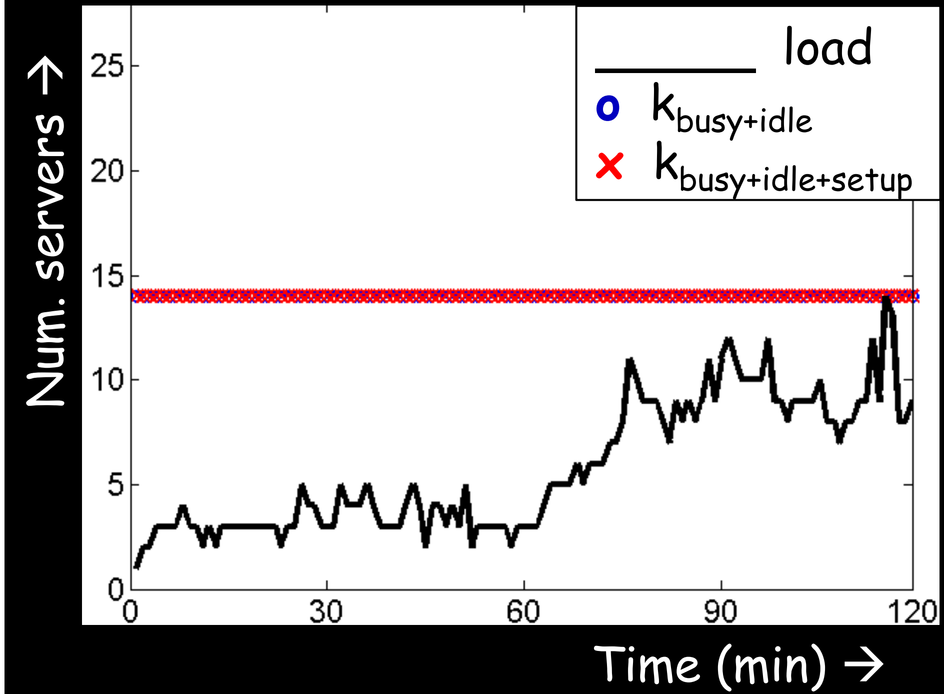
$$k = \left\lceil \frac{r_{\max}}{60} \right\rceil$$

ON/OFF

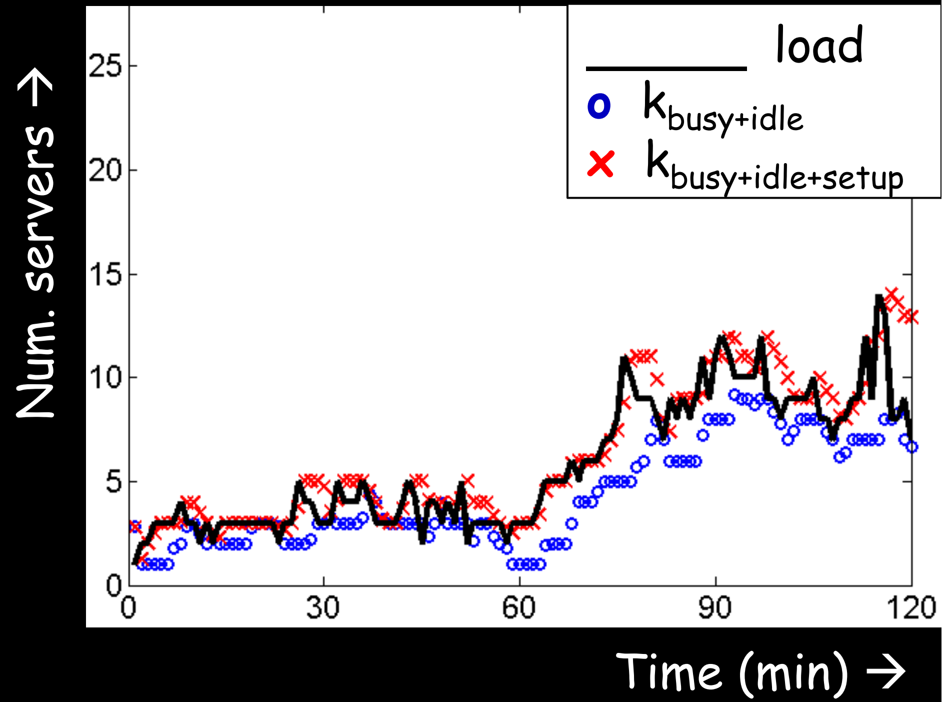


$$k(t) = \left\lceil \frac{r(t)}{60} \right\rceil$$

AlwaysOn	ON/OFF



T_{95}=291ms, P_{avg}=2,323W



T_{95}=11,003ms, P_{avg}=1,281W

I'm late,
I'm late!

ON/OFF Variants

Reactive
Control-Theoretic

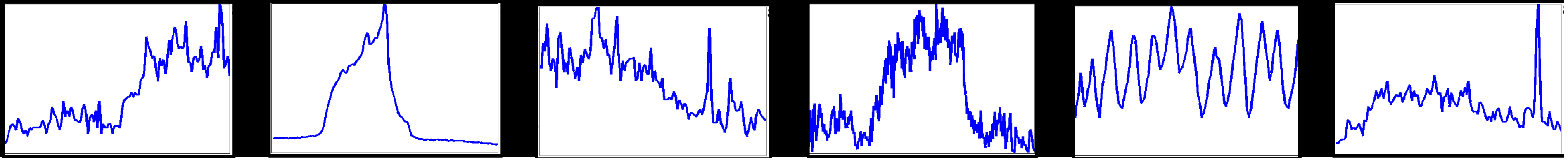
[Leite, Kusic, Mosse '10]
[Nathuji, Kansal, Ghaffark
[Fan, Weber, Barroso '07]
[Wang, Chen '08]
[Wood, Shenoy, … '07]

[Horvath, Skadron '08]
[Urgaonkar, Chandra '05]
[Bennani, Menasce '05]
[Gmach et al. '08]

Predictive

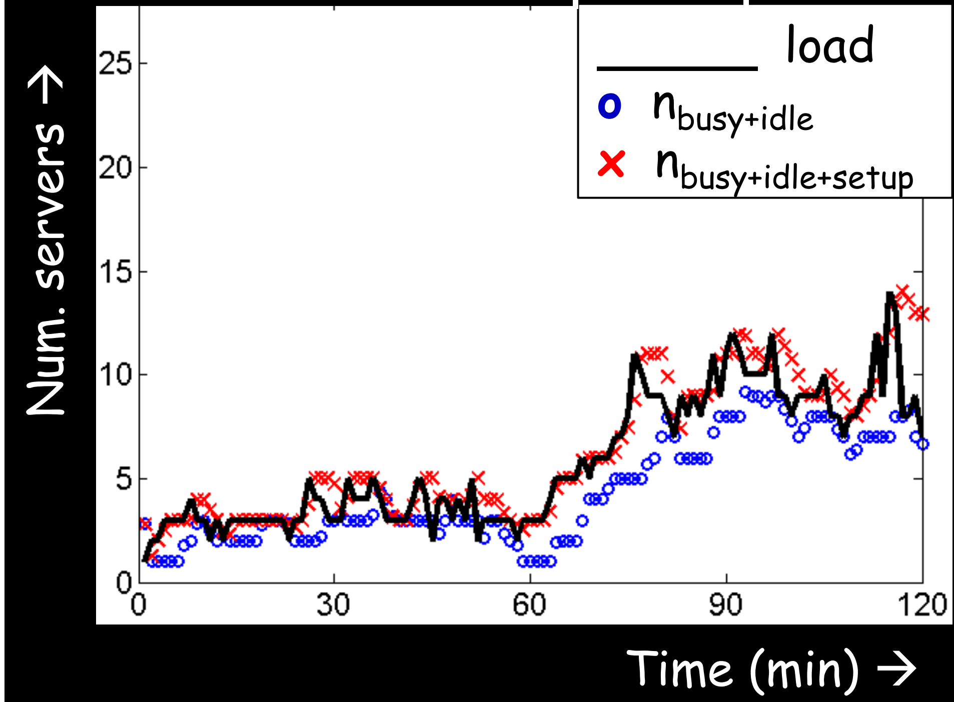
[Krioukov, …, Culler, Katz '10]
[Castellanos et al. '05]

[Chen, He, …, Zhao '08]
[Chen, Das, …, Gautam '05]
[Bobroff, Kuchut, Beaty '07]



ON/OFF

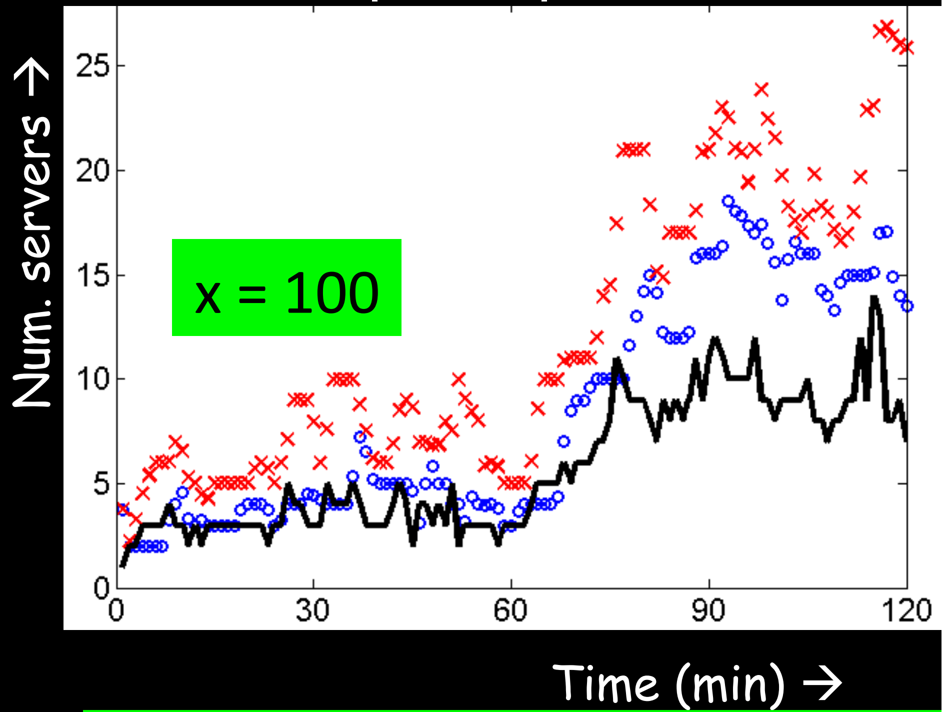
$$k(t) = \left\lceil \frac{r(t)}{60} \right\rceil$$



T_{95}=11,003ms, P_{avg}=1,281W

ON/OFF+padding

$$k(t) = \left\lceil \frac{r(t)}{60} \right\rceil \cdot (1 + x\%)$$



T_{95}=487ms, P_{avg}=2,218W

A Better Idea: AutoScale

Existing ON/OFF policies are too quick to turn servers off …
then suffer huge setup lag.

Two new ideas

- Wait some time (t_{wait}) before turning idle server off
- "Un-balance" load: Pack jobs onto as few servers as possible w/o violating SLAs

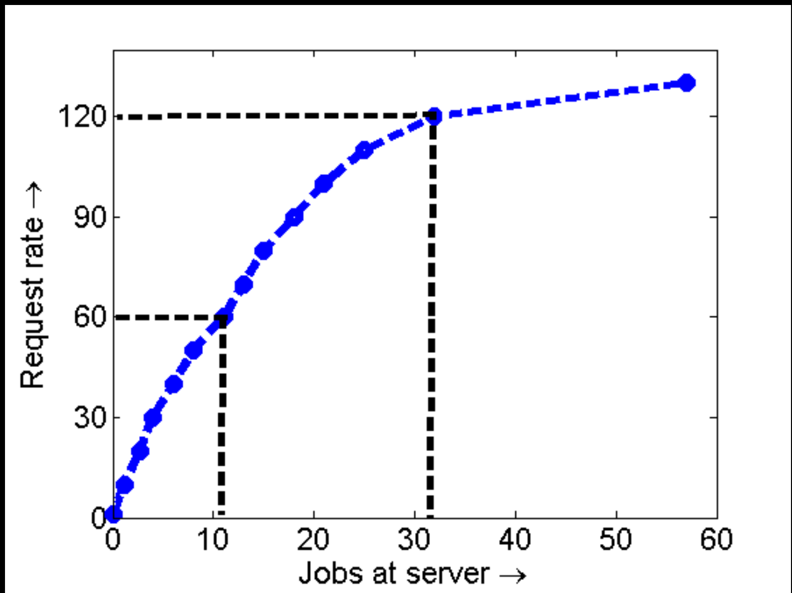
[Transactions on Computer Systems 2012]

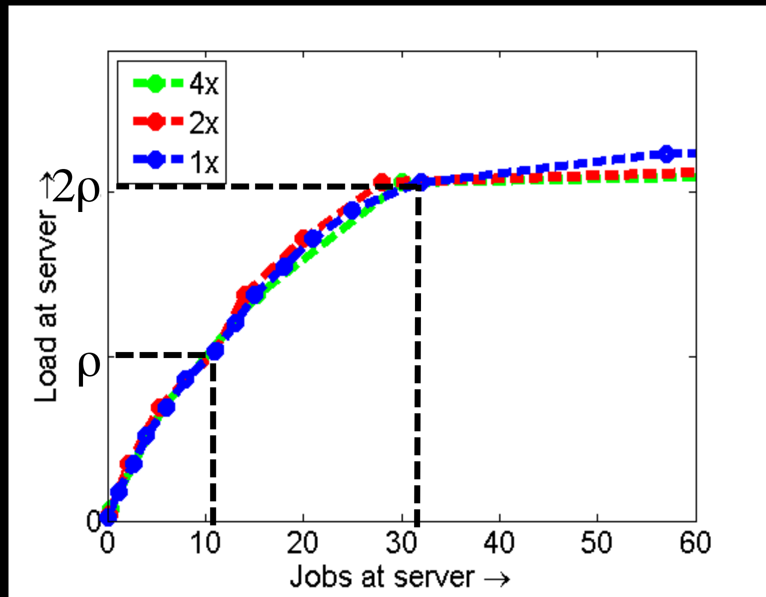
Scaling Up via AutoScale

Request rate is insufficient indicator of load.
jobs/server more robust indicator.

But not so obvious how to use # jobs/server …

10 jobs/server ⇔ load ρ
30 jobs/server ⇔ load 2ρ





[Transactions on Computer Systems 2012]

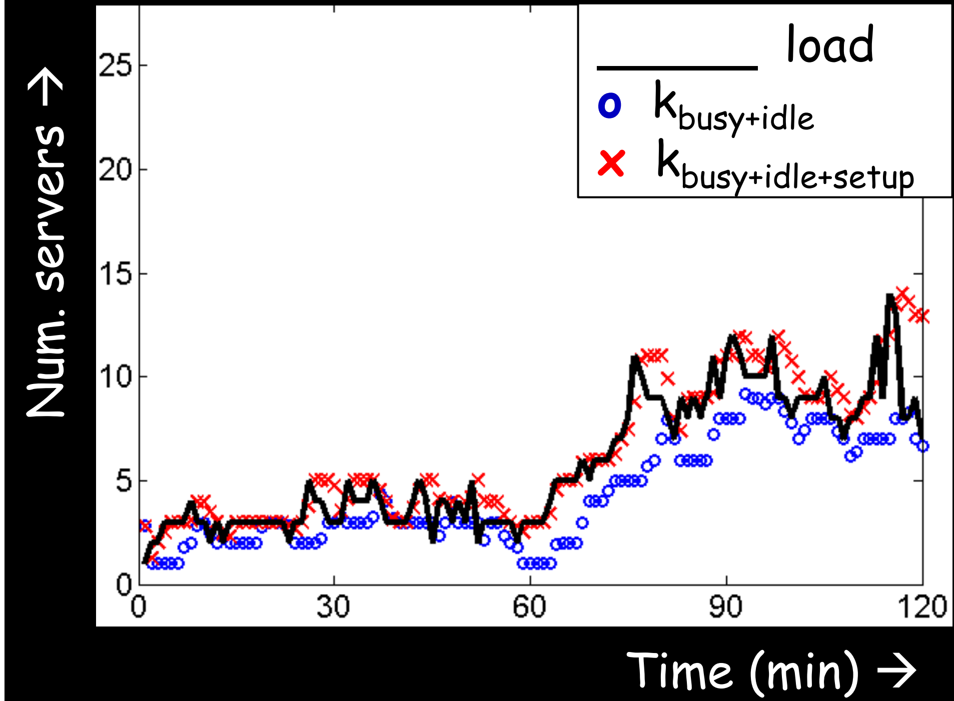
Why AutoScale works

Theorem : *As k → ∞, M/M/k with DelayedOff + Packing approaches square-root staffing.*

$$k_{avg}^{AutoScale} \rightarrow k_{avg}^{OPT} + \sqrt{k_{avg}^{OPT} \log\left(k_{avg}^{OPT}\right)}$$

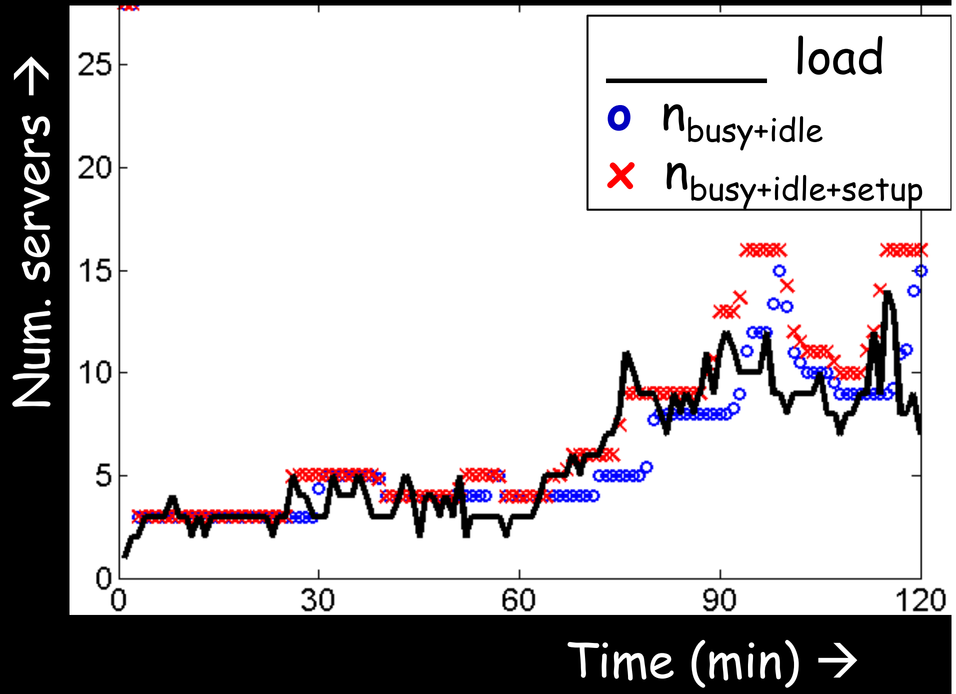
[Performance Evaluation, 2010 (b)]

ON/OFF



T_{95}=11,003ms, P_{avg}=1,281W

AutoScale



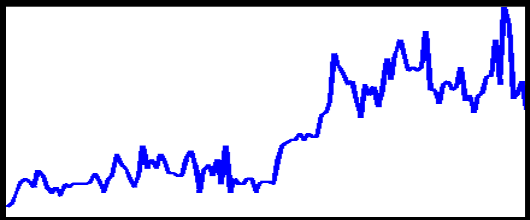
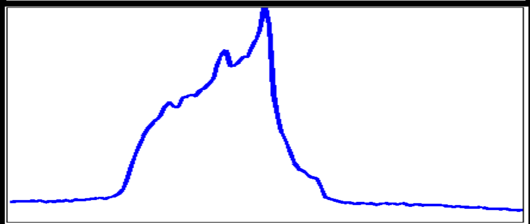
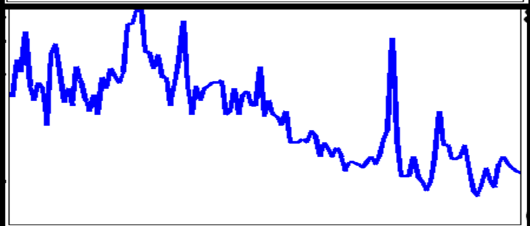
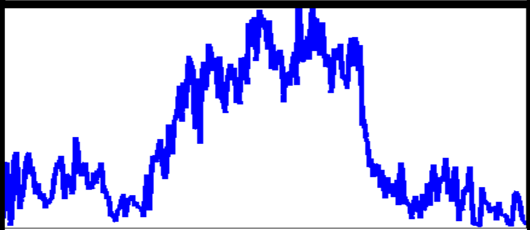
T_{95}=491ms, P_{avg}=1,297W

I'm late, I'm late!

Within 30% of OPT power on all our traces!

Facebook cluster-testing AS

Results

	AlwaysOn		ON/OFF		AutoScale	
	T_{95}	P_{avg}	T_{95}	P_{avg}	T_{95}	P_{avg}
	291 ms	2323 W	11,003 ms	1281 W	491 ms	1297 W
	271 ms	2205 W	3,802 ms	759 W	466 ms	1016 W
	289 ms	2363 W	4,227 ms	1,391 W	470 ms	1679 W
	377 ms	2263 W	> 1 min	849 W	556 ms	1412 W

[Transactions on Computer Systems 2012]

Conclusion

Dynamic power management → Managing the setup cost

Part I: Effect of setup in M/M/k

- First analysis of M/M/k/setup and M/M/∞/setup
- Introduced RRR technique for analyzing repeating Markov chains
- Effect of setup cost is very high for small k,
  but diminishes as k increases

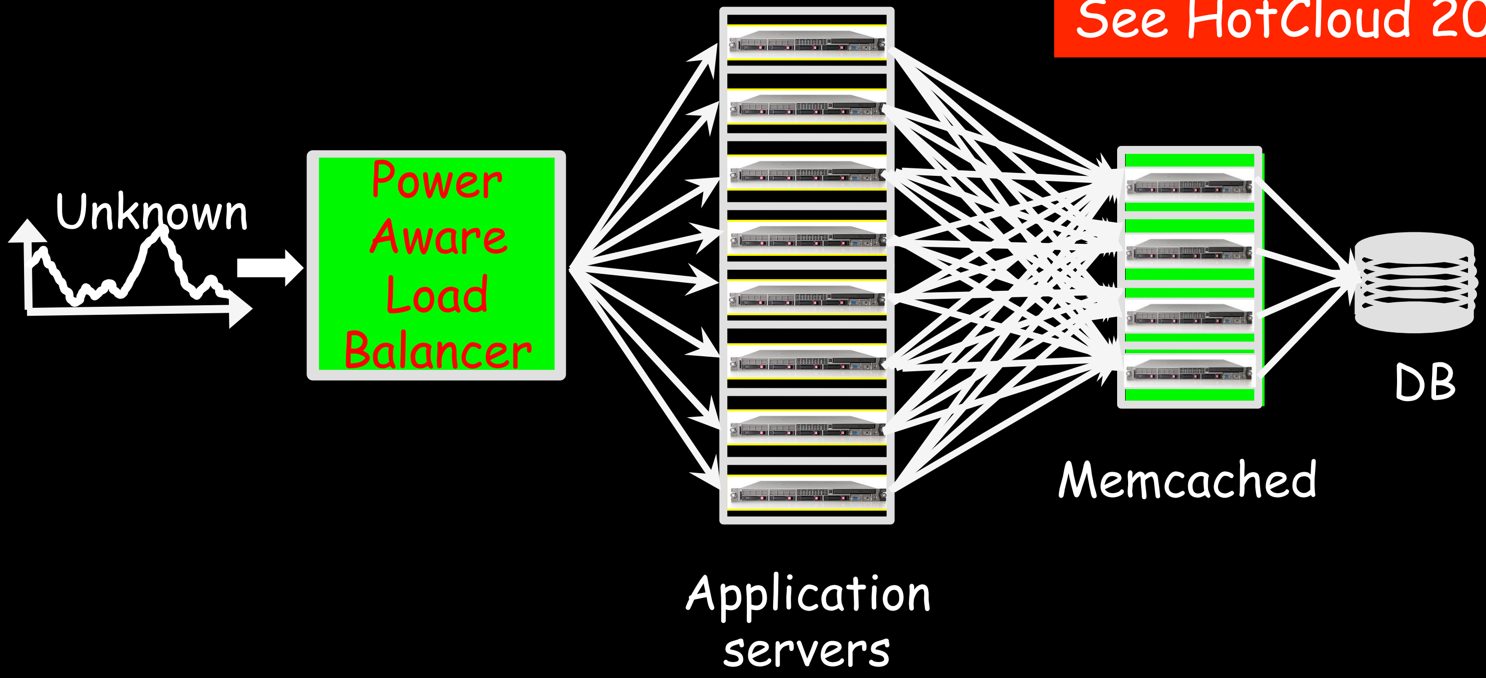
Part II: Managing the setup cost in data centers

- Non-Poisson arrival process; load unknown; unpredictable spikes
- Leaving servers AlwaysOn wastes power, but setup can be deadly.
- Lesson: Don't want to rush to turn servers off.
- Proposed AutoScale with Delayedoff, Packing routing & Non-linear Scaling.
- Demonstrated effectiveness of AutoScale in practice and theory.

Comments related to LCCC

- Scaling stateful servers?

See HotCloud 2012



- Tradeoffs between architectures:
  "Should we separate stateful from stateless?"

See Middleware 2012 – best of both

References

Anshul Gandhi, Sherwin Doroudi, Mor Harchol-Balter, Alan Scheller-Wolf. "Exact Analysis of the M/M/k/setup Class of Markov Chains via Recursive Renewal Reward." *ACM SIGMETRICS 2013 Conference,* June 2013.

Anshul Gandhi, Mor Harchol-Balter, R. Raghunathan, Mike Kozuch. "AutoScale: Dynamic, Robust Capacity Management for Multi-Tier Data Centers." *ACM Transactions on Computer Systems,* vol. 30, No. 4, Article 14, 2012, pp. 1-26.

Anshul Gandhi, Timothy Zhu, Mor Harchol-Balter, Mike Kozuch, "SOFTScale: Stealing Opportunistically for Transient Scaling." *Middleware* 2012.

Timothy Zhu, Anshul Gandhi, Mor Harchol-Balter, Mike Kozuch. "Saving Cash by Using Less Cache." HotCloud 2012.

Anshul Gandhi, Mor Harchol-Balter, and Ivo Adan. "Server farms with setup costs." *Performance Evaluation,* vol. 67, no. 11, 2010, pp. 1123-1138.

Anshul Gandhi, Varun Gupta, Mor Harchol-Balter, and Michael Kozuch. "Optimality Analysis of Energy-Performance Trade-off for Server Farm Management." *Performance Evaluation* vol. 67, no. 11, 2010, pp. 1155-1171.