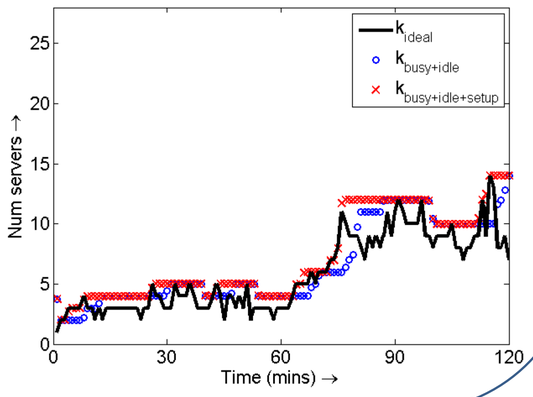
Dynamic Power Management for Multi-tier Data Centers

Anshul Gandhi, Timothy Zhu, Mor Harchol-Balter, Mike Kozuch* (CMU, *Intel Labs)

AutoScale [TOCS'12]

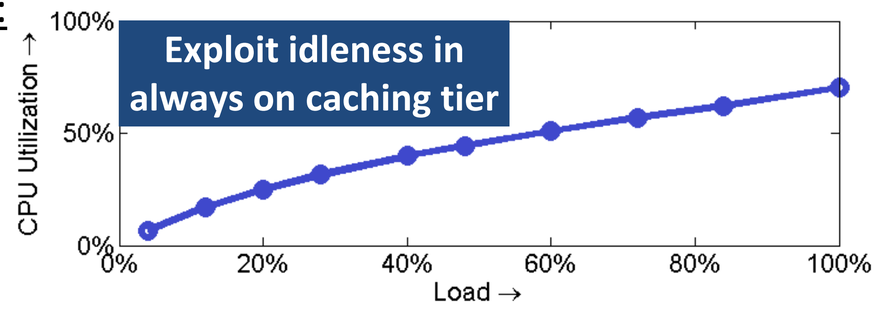
- Given unpredictable load and setup times, minimize power without violating 95%ile response time SLAs?
- Key ideas: 1. Delay turning off idle servers by t_{wait} secs
  2. Minimize idle servers by load skewing
  3. Monitor jobs in system, not request rate
- Theory: AutoScale achieves near-optimal power savings [Performance'10a]

Policy	T_{95}	P_{avg}
AlwaysOn	291ms	2,323W
Reactive	11,003ms	1,281W
Predictive MWA	7,740ms	1,276W
Predictive LR	2,544ms	2,161W
AutoScale	491ms	1,297W

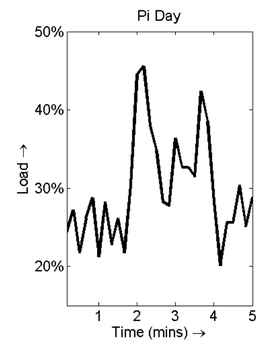


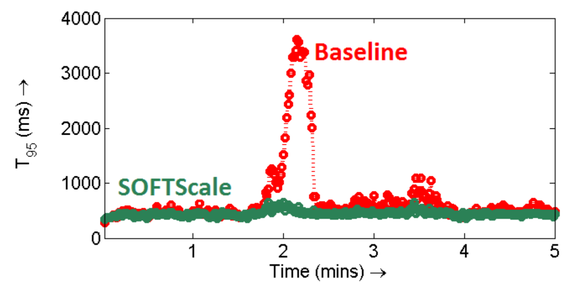
SOFTScale [Middleware'12]

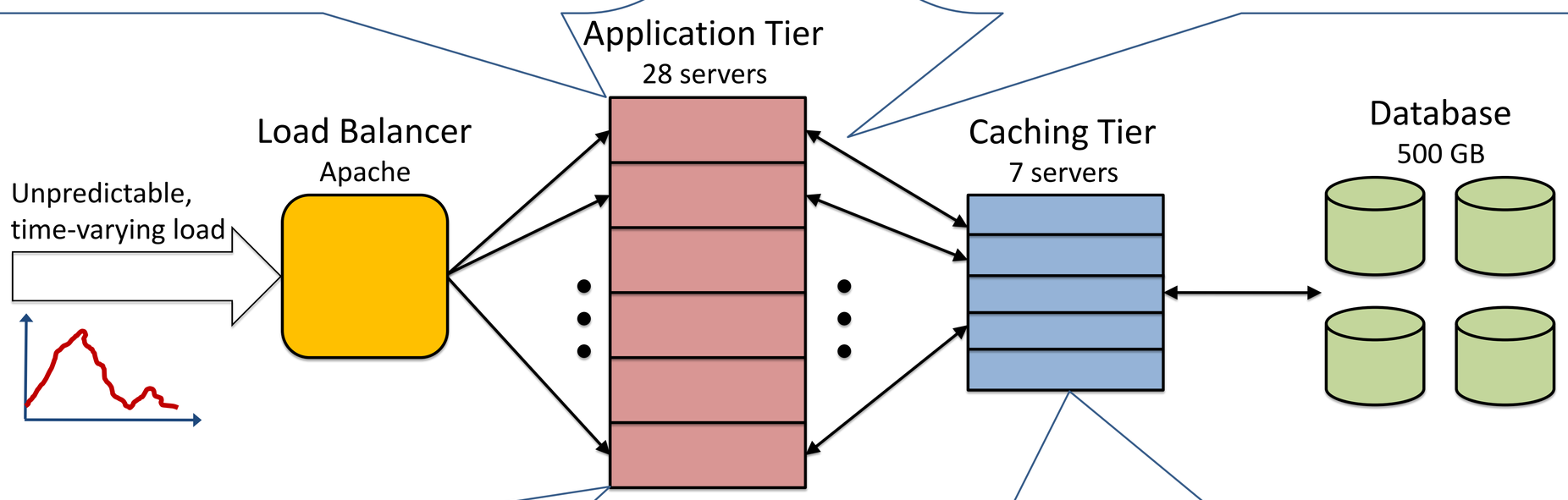
- Scalable application tier is vulnerable to load spikes
- Key idea:



- Spawn app work on caching tier and provide isolation

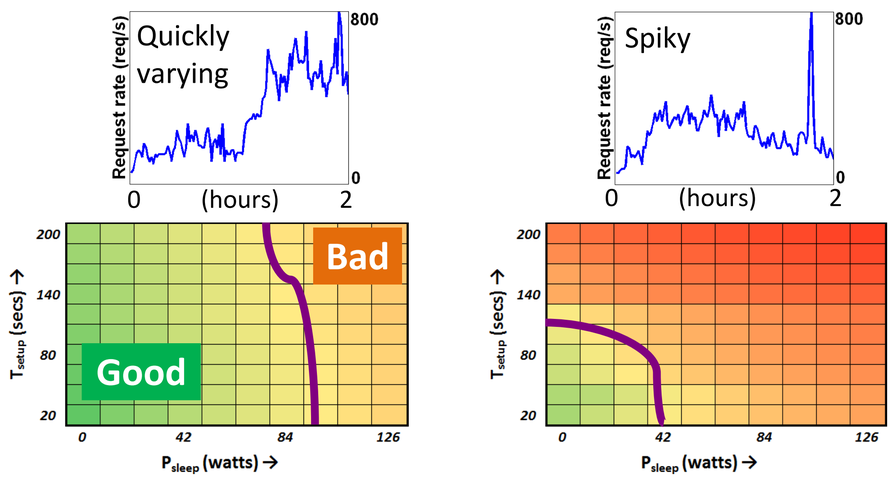






Sleep States [IGCC'12; HotPower'11]

- Sleep states are defined by (T_{setup} , P_{sleep})
- Which sleep states are useful for PPW = $(T_{95} \cdot P_{avg})^{-1}$

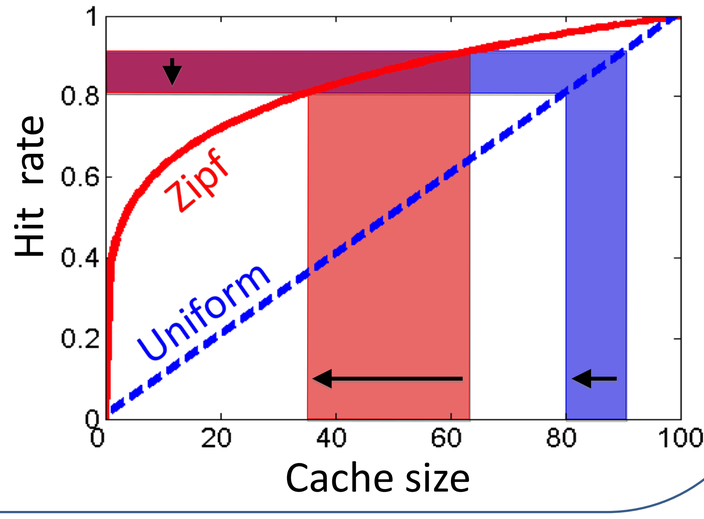


- Theory: Effect of T_{setup} goes down as data center size goes up [Performance'10b; Allerton'10]

CacheScale [HotCloud'12]

- Can we scale the caching tier?
- Caching tier only has 20% of the servers, but accounts for 33% of the total cost due to lots of expensive DRAM
- Key idea: When load drops, we can have lower hit rate
- More details in other poster

Lowering hit rate by even 10% can lead to 40% savings in cache size



Intel Science & Technology Center for Cloud Computing
Carnegie Mellon University
Georgia Tech
intel
PRINCETON UNIVERSITY
UC Berkeley