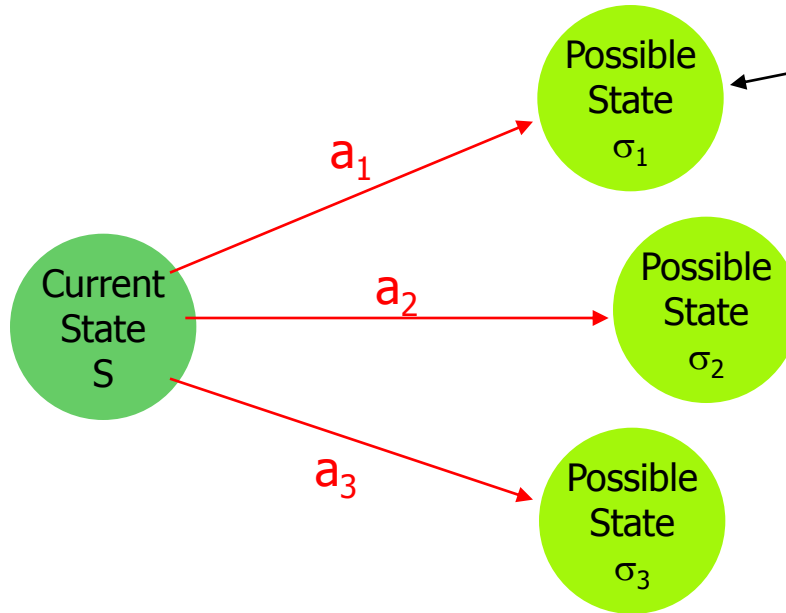


Autonomic Computing Grand Challenge

Jeff Kephart
IBM Research

- Build computing systems that manage themselves in accordance with high-level objectives from humans

Traditional policies have focused on the *current* state

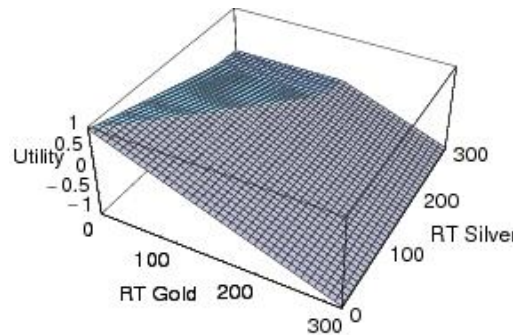


AC demands that we focus on the *desired* state

Utility functions have been advocated as a means for expressing high-level objectives, and models and optimization as a means for managing to those objectives.

Kephart and Walsh, Policy04

J. Strunk, *Using Utility Functions to Control a Distributed Storage System*, CMU PhD thesis, 2008

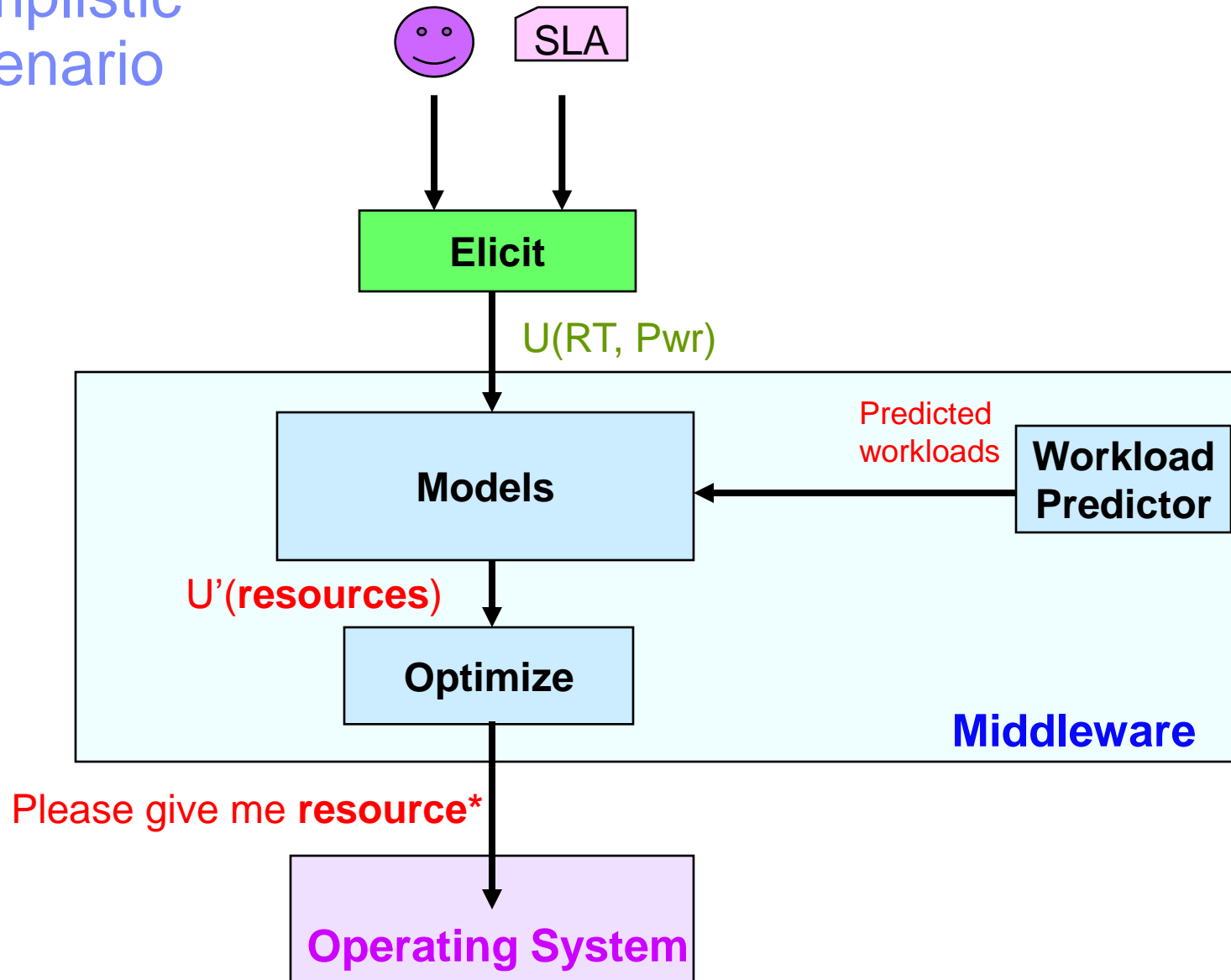


AC Baby Grand Challenge: Power-aware Middleware

- Huge and growing interest in energy-efficient data centers
 - Data centers are running out of power; costly to build new ones
 - Energy costs are rising quickly
 - Significant CO₂ emissions attributed to data center operation
- Myriad approaches are being developed and marketed
 - From chip to chiller
 - 10-12 orders of magnitude in temporal scale
 - ~5 orders of magnitude in spatial scale
- Two generic strategies
 - Turn things off
 - Slow things down

} But these methods affect performance and availability adversely!
- Energy management is intrinsically about trading power vs. performance, availability, etc. and it is the *middleware* level at which SLAs are expressed and managed
- Challenge: How can middleware work effectively with other levels of the stack to manage systems to specified performance, availability and energy objectives ?
 - Middleware operates on time scale of several seconds+
 - We want middleware to affect what happens at microsecond time scale
 - We want middleware to affect what happens at hour time scale (coordinating with facilities management)

Simplistic Scenario



Assorted Sub-challenges

- Elicitation
 - How to elicit utility; tradeoffs?
- Models
 - How to learn good models in complex environments from sparse data?
- Optimization: Architecture and algorithms
 - Decentralization seems essential
 - Multi-vendor environments
 - Multiple levels of stack must work in concert
 - Wide range of time scales, from microseconds to hours
 - How to cope with latencies in turning things on/off?
- Putting it all together at a data center scale (and beyond!)