Big Control in the Datacenter

Matei Zaharia
Overview

- The Big Control Platform can enable new applications in the physical world, but where can we find meaningful large-scale testbeds?

- Key observation: one of the easiest-to-instrument IoT testbeds is the datacenter itself!

- Both providers and cloud users are already run sophisticated control loops for resource management, cost control, security and more!
Five Example Applications

1. Physical resource management
2. Powering granular computing
3. Network security
4. Automatic workload management
5. Cloud pricing & bidding
1. Physical Resource Management

- VM, container and function placement
- Network management (e.g. pFabric, Self-Driving Networks)
- Dynamic power management
- Fault detection & recovery
2. Granular Computing

- Decide where to place functions in real time
- Learn properties of each application
- Proactively replicate code images
- Control networking between functions
- Collect + monitor large amounts of logs

Directly impacts performance and cost!
3. Network Security

- A major concern for both private and public clouds
- Benefits from scalability, low latency, data fusion and machine learning
- Interesting opportunity to monitor many more software layers (e.g. application logs)
4. Automatic Workload Management

- Autoscaling is a major draw of the cloud, but current implementations are limited to stateless web applications.
- How should we scale storage? Or complex applications composed of interacting microservices?
- Examples: choosing best storage & compute primitives in GG (distributed compiler); adaptive replication in ReFlex; elastic pub-sub service.

Can be evaluated by end-users on existing public clouds.
5. Cloud Pricing and Bidding

- Already very valuable to manage costs in current clouds
- Can become a real-time market similar to financial or ad bidding
- Nontrivial scheduling problem with diverse resources (DRF)
- Interesting design question for Granular Compute APIs (e.g. unreliable compute or storage)
How Can We Get Started?

1. Application-level testing on current clouds

2. Evaluating new techniques in testbeds
App-Level Work in Current Clouds

- Deploy and try to autoscale a multi-tier application with complex interactions between the components (e.g. social network)

- Minimize cost to compile a program in GG, our distributed compiler built on AWS Lambda (involves computation + short-term & long-term storage)

- Run complex intrusion detection algorithms at large scale

- Design a pay-as-you go elastic key-value store on AWS
Physical Testbeds

- Minimize power cost of a datacenter running a standard workload
- Use BCP to improve flow schedulers or enforce network fairness
- Design a pricing mechanism for lambda functions & implement in BCP
- Evaluate impact of broader data collection on scheduling decisions
Conclusion

- The datacenter presents an interesting testbed for BCP – either from the provider’s point of view or the datacenter user’s
- Several ideas can be tested already on public clouds
- BCP can play a big role in building granular compute platforms