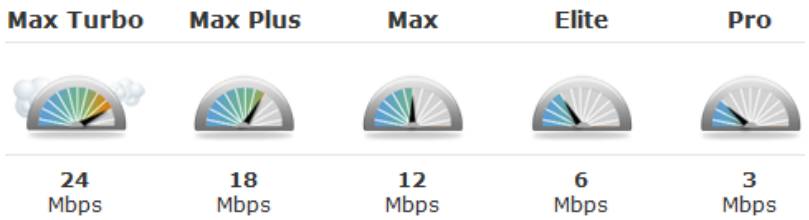
The Moving Cloud:
Predictive Placement in the Wild

Azarias Reda, Meritful
Brian Noble, University of Michigan

Latency

- Latency as a first class concern
  - With mobile devices, the “reach” is extending

- Latency is difficult:
  - Simply packing more bits doesn’t improve it
  - It has a hard upper bound
  - Additional services impact it the most
  - It’s doesn’t sell:



Humans and latency

- We are very sensitive to delay and jitter
  - Even in systems with adequate balance
- Humans notice even a modest increase in latency
  - Experiments with streaming, interactive services
- At a few hundred millisecs, many applications degrade
  - All too common in mobile connectivity

Latency in challenged networks

- Even more pronounced in many scenarios
  - Mobile connectivity
  - Shared access centers
  - Makes even simple network tasks unpleasant

- We have been looking at bandwidth primarily
  - Several approaches towards improving access
  - Some of these approaches help with latency too

Spoiler

- Provisioning data as close to demand as possible
  - Trading latency for storage and bandwidth
- The moving cloud:
  - Proposing a framework for proactive data delivery
  - Partial validation for components

Themes in the moving cloud

- People are creatures of habit
  - Move in patterns that can be probabilistically learned
  - Access data in patterns*

1. A proactive delivery infrastructure
   - Secure and extensible
2. Augmenting predictions with time bounds
   - Understanding temporal component of mobility
3. Leveraging the context of data access
   - Using context for data selection

Improved Delivery

An Example: Cyber foraging

- Augmenting computation with local surrogates
  - Cut on the latency to reach the cloud
  - Perform computationally intensive tasks on mobile devices
- Computation as a replaceable resource
  - Can be provisioned by nearby machines
  - Often a single hop away
- Try to define units of computation to offload – not easy

“Cloudlets”

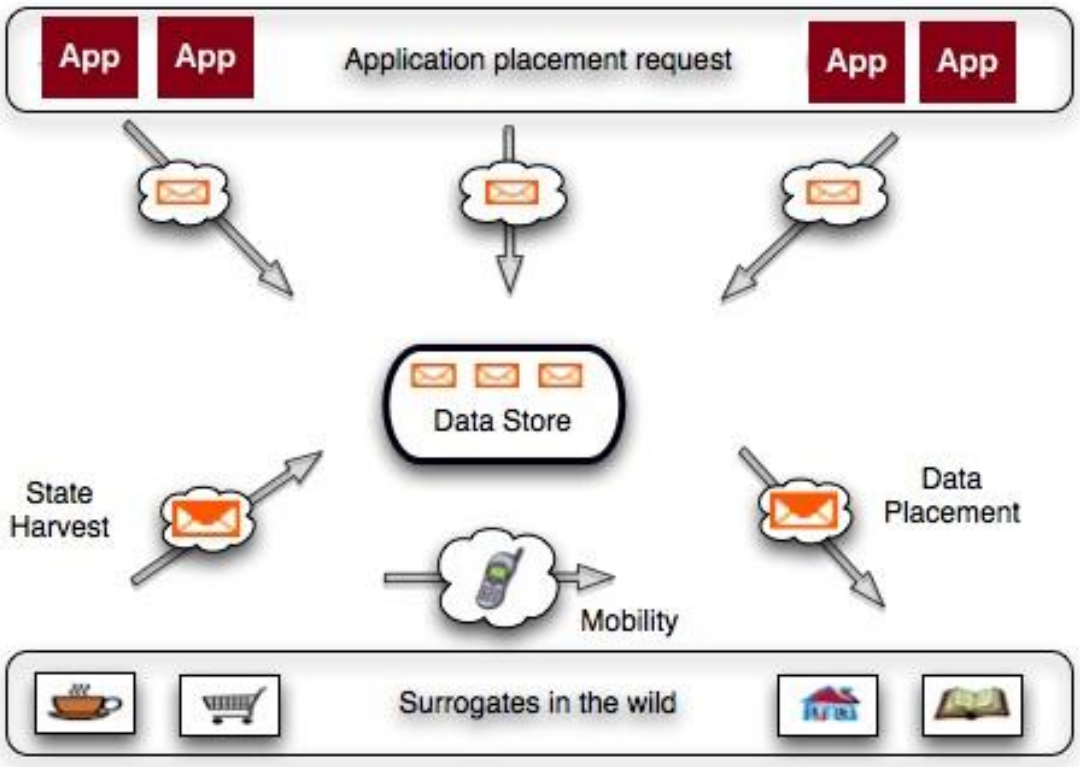
- The latest in cyber foraging
- “Data centers in a box”
  - Can be deployed alongside wireless access points
  - Provide on-demand augmentation
- Use virtual machines to encapsulate computation
  - “Base” VM at public nodes, and private “overlays”

The moving cloud and cloudlets

- Cloudlets augment computation
- However, VM matching and updating is hard
  - Many overlays to work with various base VMs
  - Updates are difficult to propagate

- The moving cloud simplifies targeted, proactive delivery
- Can be used to implement the cloudlet idea
  - Deliver targeted VMs ahead of time
  - Employ techniques for smarter and efficient delivery

High level design



Key principles:

- Proactive placement
- Extensibility
- Access models

Application model

- An application will have two components
- Server based component:
  - Deals with high fidelity, first level replica of its data
- A “mobile” component:
  - Provides the user facing interactions
  - Interacts with augmentation nodes
- A node is a publicly available storage and compute node

Application interaction model

- As updates to data happen in the wild:
  - Changes propagate back to the server component
  - Server component deals with data semantics

- The moving cloud should support a simple set of APIs
  - Pushing data into the service
  - Accessing data from nodes
  - Harvesting state from nodes
  - Versioning data pieces
  - Providing contextual access information

Mobility

- Several models for capturing human mobility
- Algorithms to predict next location of individuals
  - Some with very good success, close to 90%
- We worked on adding time bounds to location prediction
  - Important for proactive data delivery
  - Provides actionable information
  - Needed to deliver fresh data

Mobility

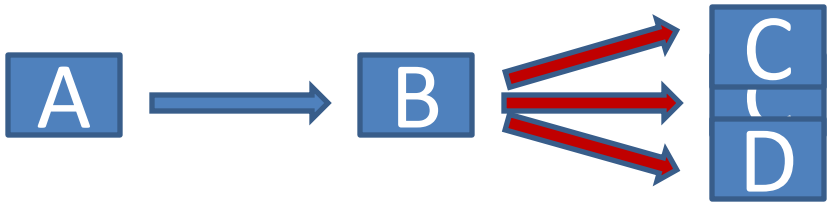
- Approaches:
  - Fingerprints to identify distinct routes
  - Probabilistically analyzing associated route times

- A route is a recorded mobility edge between two nodes
  - Has an identifying precedent (its fingerprint)
  - A measured time of travel

Time augmentation

- *M* captures the statistical distribution of route times
  - Used to estimate expected route time

- For the fingerprint, we use the *previous* two locations
  - For the following route, (A,B) is used as the fingerprint:



  - A fingerprint is not unique, can lead to multiple routes

Time augmentation

- A second order Markov chain to represent fingerprints
  - A fingerprint matrix
  - A sparse matrix used to decide on route choices
- The Markov chain is used to select next possible route
  - Statistical distribution is used to estimate route time
- Capture deviation from ground truth to estimate error
  - Used to establish time bound confidence
  - Good first results with the CROWDAD mobility data

Contextual access behavior

- Data access patterns used for system design
  - Temporal and spatial locality in access
  - Clusters of files accessed together
- We posit access can be correlated with context of access
  - For example: location and time
  - “office data” vs. “home data”
- Patterns emerge over time, and can be used for delivery

Applications in challenged networks

- Predictors work well in the absence of detailed location
  - Location at the granularity of internet kiosks, for example
- Pushing data proactively reduces access latency
- However, special sensitivity to prediction confidence
  - Since resources are limited to begin with
- We have modified Sulula to support this style of delivery

A few challenges/directions

- **Consistency models**
- Data that was consistent at delivery might not remain so
- Harvesting residual state from nodes is important
- Vertical and horizontal consistency
- Need robust versioning and consistency

Challenges

- Improving prediction certainty
- Especially important in constrained environments
- Utility models for deciding data placement
- Dealing with phase changes in human mobility
  - Quick learning when errors increase
  - Enough memory to revert back to old routines

Challenges

- System level support for modeling access context
- A lot of work in the space of application hints
  - For optimizing network and energy use
- Capturing and communicating patterns in access context

Thank you!

Azarias Reda

azarias@umich.edu