Privay-preserving Systems for Processing Personal Data

DMSN Seminar 2017

Yousef Amar

Queen Mary University of London

2017-03-14

Background
What is Personal Data?

Background
What is Personal Data?

- Online and social media data
  - Facebook, Twitter, Instagram...
  - Email
  - Online banking

Background
What is Personal Data?

- Online and social media data
  - Facebook, Twitter, Instagram...
  - Email
  - Online banking
- Smartphone sensors and wearable
  - Message history, GPS, accelerometers, gyroscopes, temperature, microphone, Bluetooth...
  - Smartwatches, GSR, heart rate, step counter...

Background
What is Personal Data?

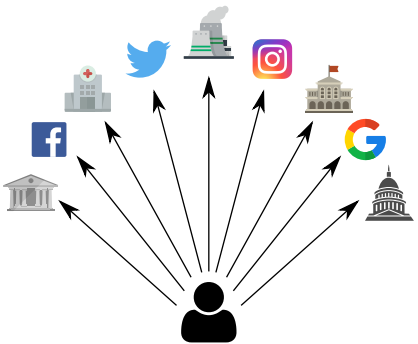
- Online and social media data
  - Facebook, Twitter, Instagram...
  - Email
  - Online banking
- Smartphone sensors and wearable
  - Message history, GPS, accelerometers, gyroscopes, temperature, microphone, Bluetooth...
  - Smartwatches, GSR, heart rate, step counter...
- IoT devices in the home
  - Smart devices (TVs, fridges, etc)
  - Lighting, heating, energy usage, proximity sensors...

Background
What is Personal Data?

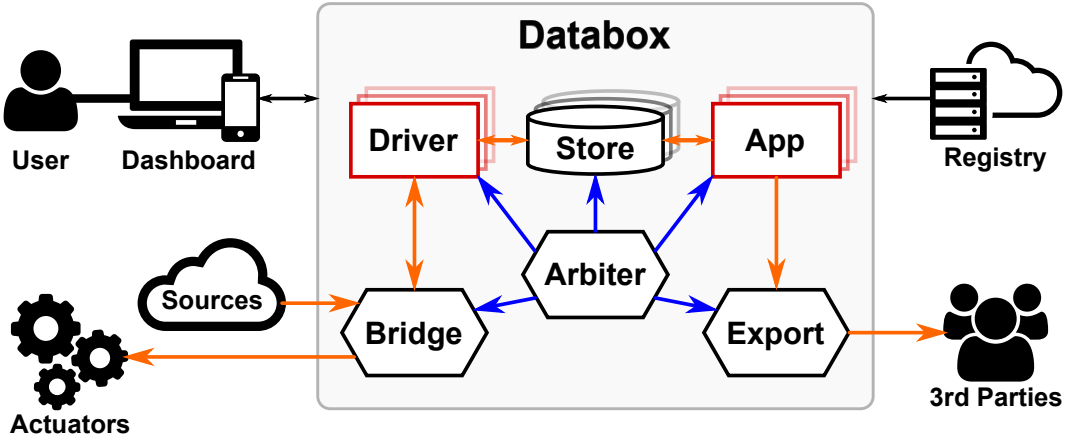
- Online and social media data
  - Facebook, Twitter, Instagram...
  - Email
  - Online banking
- Smartphone sensors and wearable
  - Message history, GPS, accelerometers, gyroscopes, temperature, microphone, Bluetooth...
  - Smartwatches, GSR, heart rate, step counter...
- IoT devices in the home
  - Smart devices (TVs, fridges, etc)
  - Lighting, heating, energy usage, proximity sensors...
- Discrete files/blobs/documents vs naturally time series

Background
The Status Quo

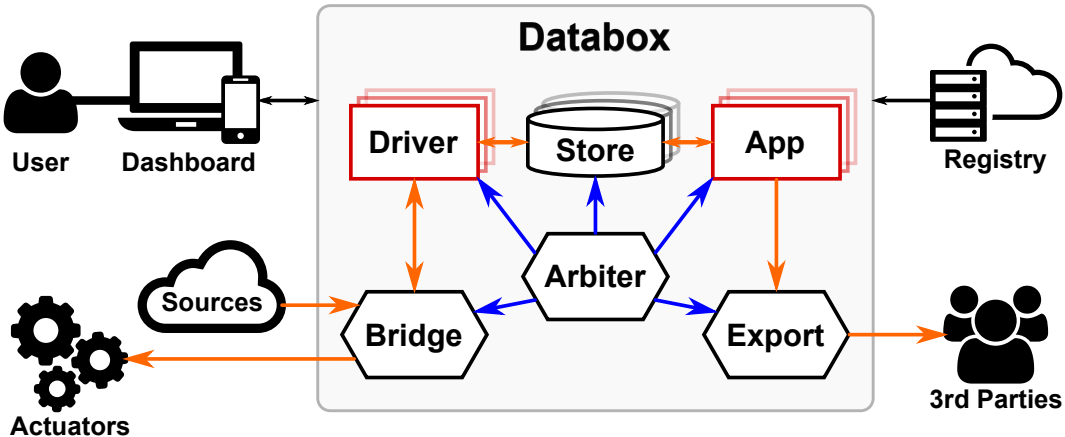
- Our digital footprint is explosively increasing
- Our data is scattered all over the cloud
  - Silos of data inaccessible
  - Limited analytics; tech and legal bounds
  - Data breaches on the rise
- Cloud solutions, e.g. homomorphic encryption face same issues
- Most data doesn't even need to leave your home/phone; costs power, bandwidth, latency, and money
- Need for different architectural paradigm



Background
Databox



Background
Databox



How can we design safe, scalable access control systems with arbitrary restrictions in this context?

Inter-container Communication

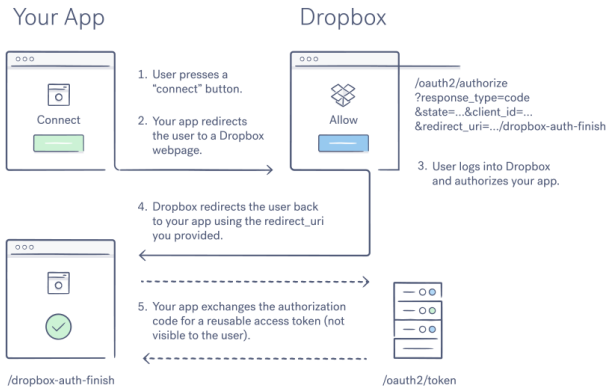
- All communication over HTTPS; certs generated run-time; system root CA
- RESTful APIs for all operations
- Direct mapping of HTTP methods to CRUD functions
- Per-route granular permissions
- Network-level isolation additionally
- Content Security Policy (CSP) to sandbox UIs

```
{
  "target": "smartphone-store",
  "path": "/accelerometer/ts/latest",
  "method": "POST"
}

{
  "target": "smartphone-store",
  "path": "/(sub|unsub)/gps/*",
  "method": "GET"
}
```

Delegated Authorisation

- Data can be protected through basic authentication; not granular enough
- Many APIs use token-based authorization, e.g. OAuth 2.0 (Twitter coursework)



Delegated Authorisation

Macaroons

- Google Research: Macaroons
  - A standard similar to signed cookies
  - Can be attenuated by “caveats”
  - Embedded permissions
  - Minting and verification can be separated through shared secret keys

```
target = smartphone-store
path = /(sub|unsub)/gps/*
method = GET
time < 1489405851417

target = smartphone-store
path = /light/ts/range
method = GET
startTimestamp >= 1489405234352
endTimestamp <= 1489405259525
```

Resource Discovery

- API for describing APIs
- Directory servers
- Many competing standards
  - Resource Description Framework (RDF)
  - Web Application Description Language (WADL)
  - Web Services Description Language (WSDL)
  - eXtensible Resource Descriptor (XRD)
- Subject-predicate-object style pervalent
- Different formats and applications — XML for REST, SOAP, OpenID

Resource Discovery

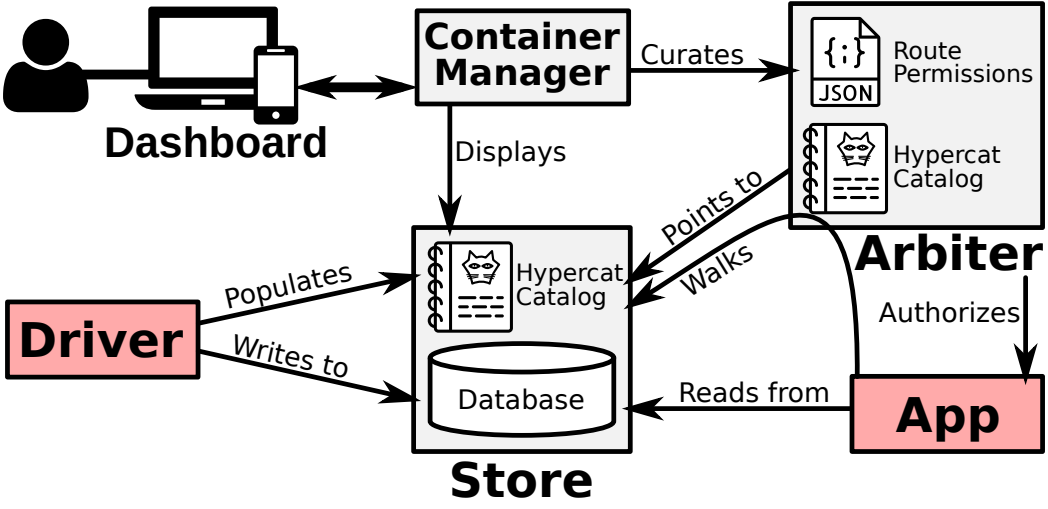
Hypercat

- Recently joined BSI Group (British Standards Institution)
- IoT-first specification design
- JSON/REST over XML/SOAP
- Only cataloguing; ontologies and authorisation extensible
- Discoverability vs accessibility
- Catalogues can be nested, allowing decentralisation and distribution

```
{
  "catalogue-metadata": [{
    "rel": "urn:X-hypercat:rels:isContentType",
    "val": "application/vnd.hypercat.catalogue+json"
  }, {
    "rel": "urn:X-hypercat:rels:hasDescription:en",
    "val": "A Databox Store"
  }],
  "items": [{
    "href": "http://some-store/light",
    "item-metadata": [{
      "rel": "urn:X-hypercat:rels:hasDescription:en",
      "val": "Light Datasource"
    }, {
      "rel": "urn:X-databox:rels:hasVendor",
      "val": "Databox Inc."
    }, {
      "rel": "urn:X-databox:rels:isActuator",
      "val": false
    }]
  }]
}
```

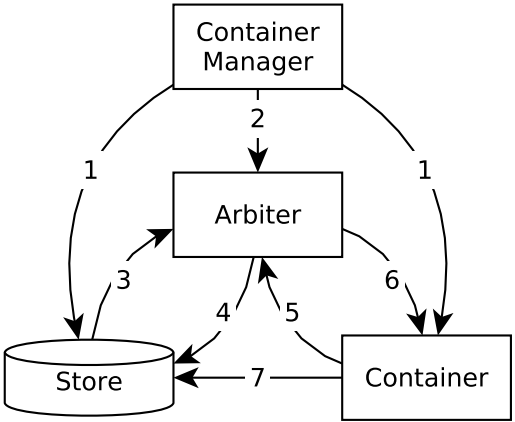
Implementation

Container Relationships

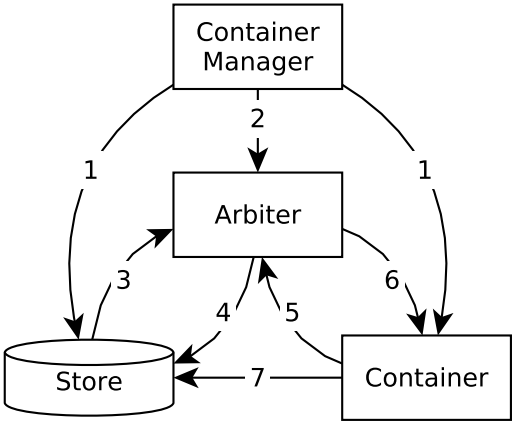


Implementation

Authorisation Flow

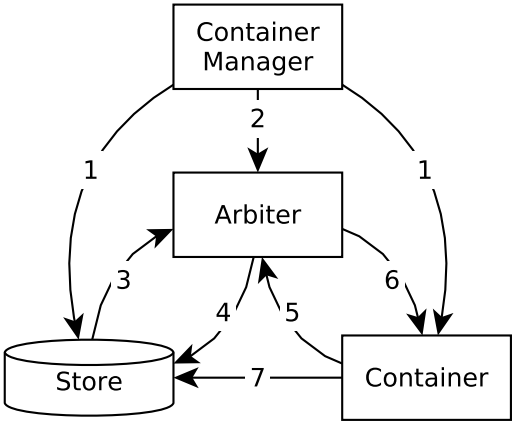


Implementation
Authorisation Flow



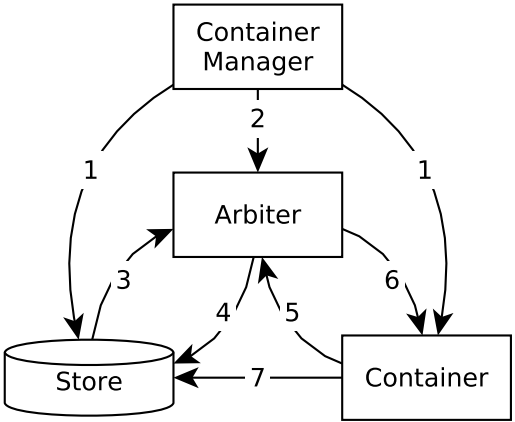
1. CM passes unique tokens

Implementation
Authorisation Flow



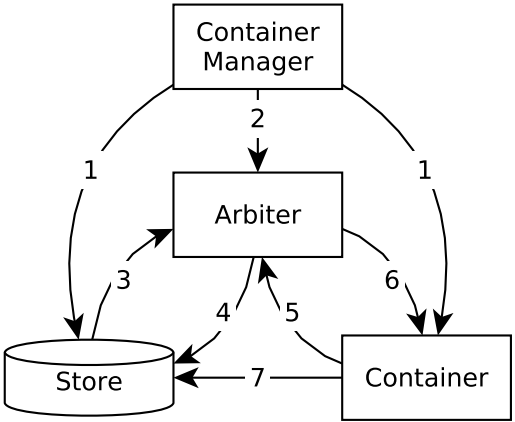
1. CM passes unique tokens
2. CM updates permissions

Implementation
Authorisation Flow



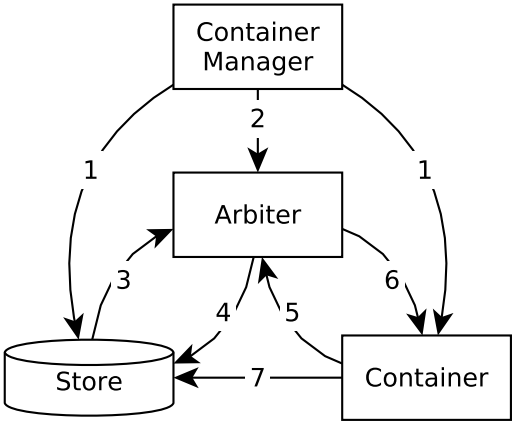
1. CM passes unique tokens
2. CM updates permissions
3. Store registers itself

Implementation
Authorisation Flow



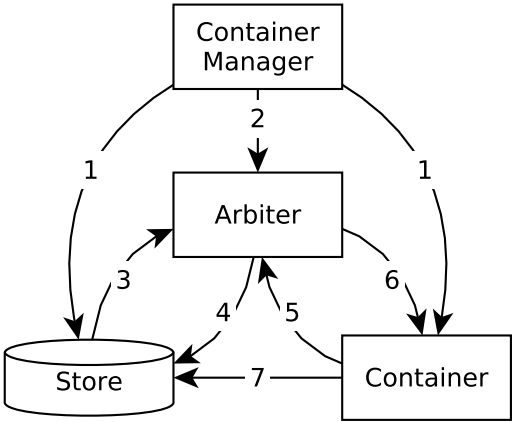
1. CM passes unique tokens
2. CM updates permissions
3. Store registers itself
4. Arbiter responds with shared secret

Implementation
Authorisation Flow



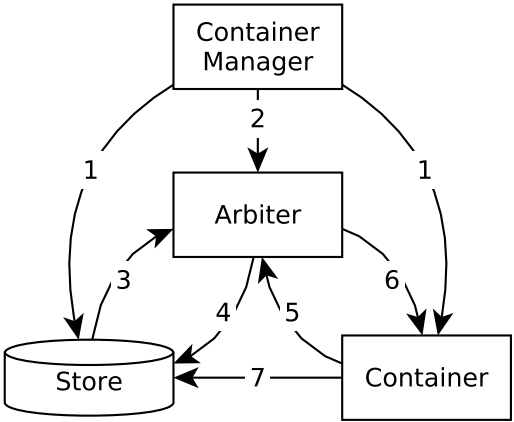
1. CM passes unique tokens
2. CM updates permissions
3. Store registers itself
4. Arbiter responds with shared secret
5. Container requests bearer token

Implementation
Authorisation Flow



1. CM passes unique tokens
2. CM updates permissions
3. Store registers itself
4. Arbiter responds with shared secret
5. Container requests bearer token
6. Arbiter checks and responds

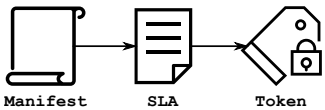
Implementation
Authorisation Flow



1. CM passes unique tokens
2. CM updates permissions
3. Store registers itself
4. Arbiter responds with shared secret
5. Container requests bearer token
6. Arbiter checks and responds
7. Container can now read/write to store

Implementation
Transcription of Permissions

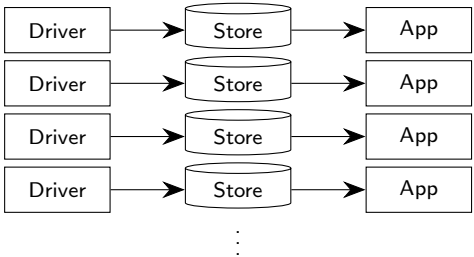
1. Drivers/apps come packaged with a *manifest*
   - Contain image metadata
   - Enumerate granular permissions for sources, concurrency, external access, and hardware
2. Users generate a Service-level Agreement (SLA)
3. The arbiter records granted permissions
4. Tokens are minted based on these




```
{
  "name": "app",
  "author": "amar",
  "permissions": [
    {
      "source": "twitter"
      "required": true
    },
    {
      "source": "gps"
    },
    {},
    {}
  ]
}
```

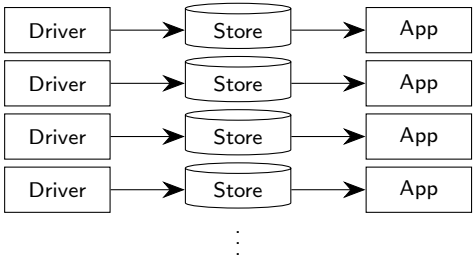
Scalability Evaluation

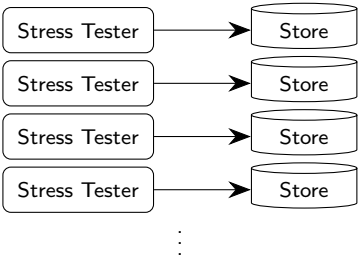
Procedure



Scalability Evaluation

Procedure





Scalability Evaluation

Results

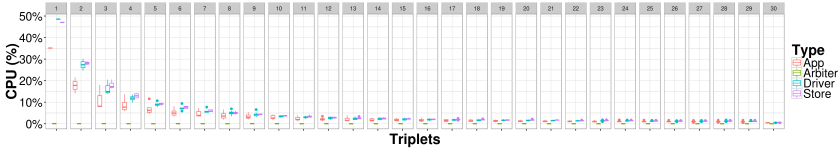


Figure: Percentage CPU Usage by Container Type

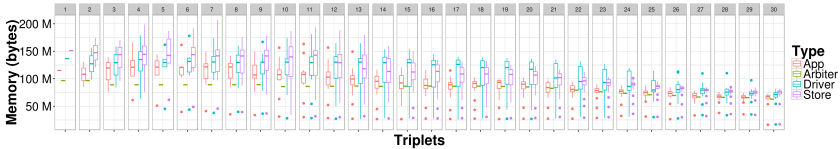


Figure: Memory Usage by Container Type

Scalability Evaluation

Results

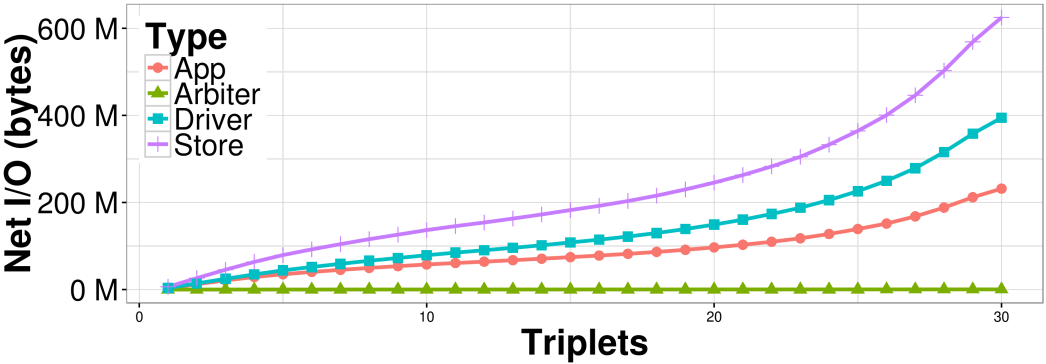


Figure: Sum Net I/O by Container Type

Scalability Evaluation

Results

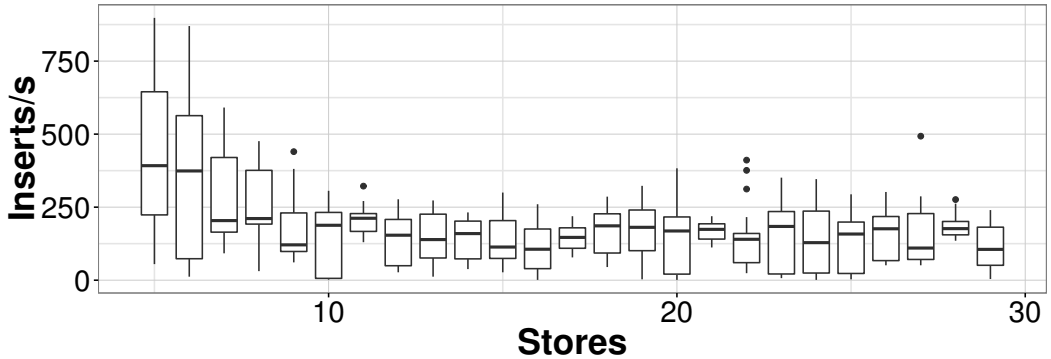


Figure: Inserts/s over Stores under Maximum Load

Scalability Evaluation

Results

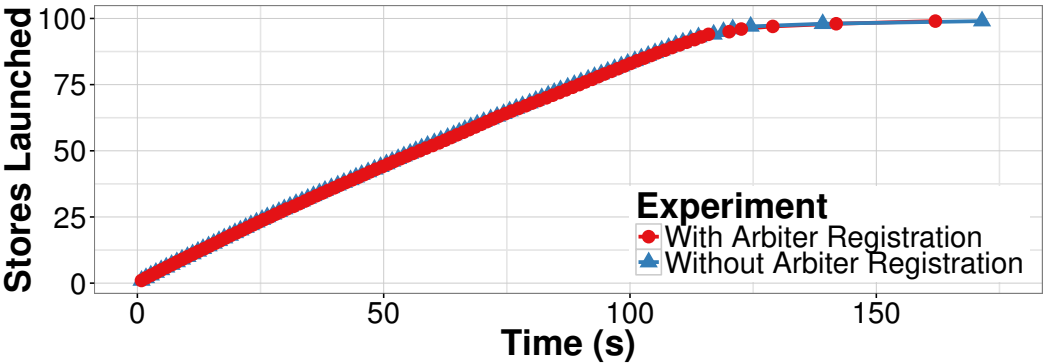


Figure: Stores Launched over Time

Topological Evaluation
Procedure and Results

Differences in Time to Availability (TTA)

1. Device → Cloud: $65ms$
2. Device → Cloud → Home: $83ms$
3. Device → Home: $78ms$
4. Device → Home → Cloud: $80ms$

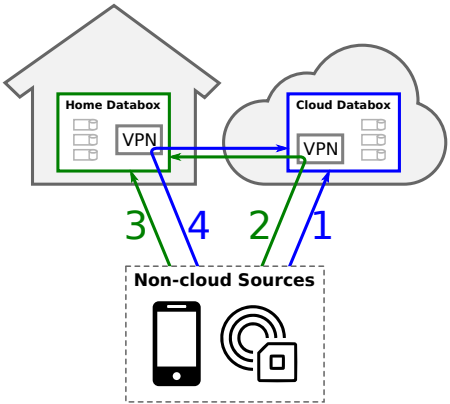


Figure: The four possible data flow scenarios tested

Topological Evaluation

Results

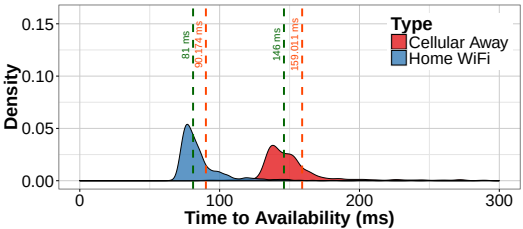


Figure: Data Time to Availability from Device to Cloud Databox Directly

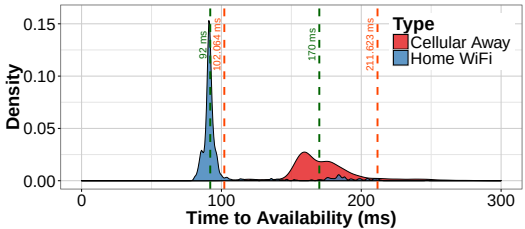


Figure: Data Time to Availability from Device to Home Databox Directly

Topological Evaluation

Results

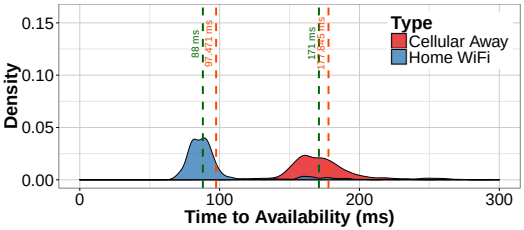


Figure: Data Time to Availability from Device to Home Databox via Cloud VPN

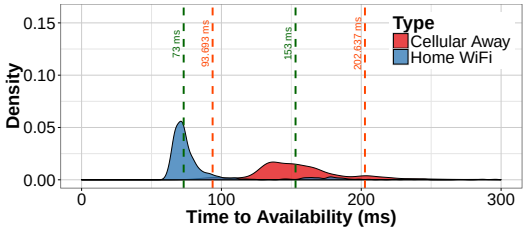


Figure: Data Time to Availability from Device to Cloud Databox via Home VPN

Topological Evaluation

Conclusions

- TTA source away from home $>$ source at home
- So minor, barely indistinguishable from NTP drift
- Based on performance alone, UX indifferent
- Scenarios through home (especially when source is away) have mean shifted right due to latency spikes
- Direct connections mean lower TTA, and cloud faster than home ceteris paribus
- Small difference for devices as sources vs cloud servers
- For devices, processing at home $>$ in the cloud $\pm$ NTP error even ignoring privacy advantages
- Home vs cloud — reliability vs cost
- Pure cloud only more advantageous for off-site processing (e.g. GPU-heavy image processing)

Next Steps

- Community Launch next Friday
- EuroSys 2017
- Full system evaluation for SOSP 2017
- ARM support — RPi
- Many areas to research, e.g. watermarking
- Many example apps and drivers, with multipurpose datavis and transformation

Thank you for your attention!

Questions?

More info:	`http://www.databoxproject.uk/`
Contribute:	`https://github.com/me-box`
Slides:	`https://github.com/yousefamar/dmsn-seminar-2017`