Application-Aware Resource Provisioning in a Heterogeneous Internet of Things

Eric Sturzinger*, Massimo Tornatore † *, and Biswanath Mukherjee*

*University of California, Davis  † Politecnico di Milano

16 May 2017
Outline

• Motivation
• Application profiles
• Mathematical formulation
• Simulation Results
• Conclusion
Motivation and objective

- Next big wave:
  Internet of Things (IoT) and Machine to Machine (M2M) traffic

Objective:
What is the impact of IoT traffic on optical network planning?
(Especially in metro!)

Enabler: hybrid fog-cloud computing

- **Drivers**
  - Bandwidth
  - Latency

Hybrid Fog-Cloud

Source: ns2-projects.org

Source: quora.com

Source: laroccasolutions
Challenges

1. Lack of quantitative characterization of IoT and M2M application
   • Identify application profiles

2. Network cost parameters

3. A model to assign resources (bandwidth, computing, storage):
   • tailored to application profile (slicing)
   • minimizing costs
Application Profile (1)

Source: GSMA Intelligence, 2015
Application Profile (2)

- Each application profile contains a unique combination of parameters
  - \( \Theta \): Latency budget (source to destination)
  - \( \kappa \): Bandwidth
  - \( \alpha \): Computational complexity (per unit of traffic)
  - \( \Lambda \): Storage time
  - \( \beta \): Compression factor (ratio of processed-to-raw data)

<table>
<thead>
<tr>
<th>Examples</th>
<th>( \Theta ) (ms)</th>
<th>( \kappa ) (Mbps)</th>
<th>( \alpha ) (CPU/Mbps)</th>
<th>( \beta )</th>
<th>( \Lambda ) (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - AR/VR</td>
<td>10</td>
<td>100</td>
<td>0.03</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>2 – Factory Automation</td>
<td>20</td>
<td>1</td>
<td>0.009</td>
<td>0.8</td>
<td>10</td>
</tr>
<tr>
<td>3 – Data Backup</td>
<td>1000</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4 – Smart Grid</td>
<td>50</td>
<td>0.4</td>
<td>0.007</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>5 – Smart Home</td>
<td>60</td>
<td>.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6 – Medical</td>
<td>40</td>
<td>2</td>
<td>0.02</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>7 – Environmental Mon.</td>
<td>1000</td>
<td>1</td>
<td>0.02</td>
<td>0.1</td>
<td>100</td>
</tr>
<tr>
<td>8 – Tactile Internet</td>
<td>1</td>
<td>200</td>
<td>.005</td>
<td>0.8</td>
<td>0</td>
</tr>
</tbody>
</table>


Application Profile (3)

- Additional note: we classify applications in four categories

\[ A = A_p \cup A_s \cup A_{sp} \cup A_n \]

- Processing only (cloud gaming, virtual reality, …)
- Storage only (analytics repositories, …)
- Processing & Storage (smart grid, …)
- Point to Point (basic M2M)
Hybrid Fog-Cloud Network Architecture

- 3 tiers of Central Offices (COs)
  - Access CO [2,3,4]
  - Metro CO [1,5,9,13]
  - Core CO [17,18]
- Data Centers
  - [19,20]

**P**

- **P**:
  - Tier-2 processing

**S**

- **S**:
  - Tier-2 storage

**Access**

**Metro**

**Core**

**Enterprise**

**Mobile**

**Wi-Fi, EPON**

**WSN**
Network (Cost) Parameters

- Costs (in COs)
  - $\mu$: processing cost
  - $\nu$: storage cost
  - $\Lambda$: metro bandwidth cost
  - $\varepsilon_{up}$, $\varepsilon_{down}$: core bandwidth cost

<table>
<thead>
<tr>
<th>Tier</th>
<th>$\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>($/CPU/Mo)</td>
</tr>
<tr>
<td></td>
<td>($/GB/Mo)</td>
</tr>
<tr>
<td></td>
<td>($/Mbps/Mo)</td>
</tr>
<tr>
<td>1 – Access CO</td>
<td>90</td>
</tr>
<tr>
<td>2 – Metro CO</td>
<td>70</td>
</tr>
<tr>
<td>3 – Core CO</td>
<td>50</td>
</tr>
<tr>
<td>4 - DC</td>
<td>25</td>
</tr>
</tbody>
</table>

https://cloud.google.com/compute/pricing
https://cloud.google.com/storage/pricing#pricing
http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php#
M. Tornatore: Application-Aware Resource Provisioning
Problem Statement

• **Inputs**
  - Offered traffic (per s-d pair, per application)
  - Application profiles:
    - \( \Theta, \kappa, \alpha, \beta, \Delta \)
    - Hybrid fog-cloud network topology \( G(N,L) \)

• **Objective function**
  - Minimize total resource provisioning cost

• **Constraints**
  - Node (processing, storage) and link capacity
  - Latency

• **Outputs**
  - “Slice” (per s-d pair, per application) consisting of:
    - Path(s) (with required bandwidth)
    - Required processing and storage resources at each node
Core Network Example

Local - cloud processing
15 + 10 + 15 = 40 ms

Global - cloud processing
17.25 + 15 + 10 = 42.25 ms

Global - fog processing
15 + 14.75 = 29.75 ms
Mathematical Formulation

Objective Function:

\[ \min(Cost_p + Cost_s + Cost_u + Cost_d + Cost_c) \]

1. **Processing**

\[
Cost_p = \sum_{m \in \mathcal{M}_p} \mu_m \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \alpha_a \sum_{f \in \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m}
\]

2. **Storage**

\[
Cost_s = \sum_{a \in \mathcal{A}_s} \Delta_a \sum_{f \in \mathcal{F}_c} \nu_f \sum_{s \in \mathcal{N}_g} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_s} \beta_a \Delta_a \sum_{f \in \mathcal{F}_c} \nu_f \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_p} x_{a,f}^{s,m} v_{a,f}^{s,m}
\]

3. **Upstream core BW**

\[
Cost_u = \epsilon_{up} \left[ \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_p} \sum_{f \in \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_p} \sum_{f \in \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} \right]
\]

4. **Downstream core BW**

\[
Cost_d = \epsilon_{down} \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_d} \sum_{f \in \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m}
\]

5. **Metro BW**

\[
\begin{align*}
\text{Cap}_{1,i,j} &= \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_d} \sum_{f \in \mathcal{F}_c} r_{a,k,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_p} \sum_{m \in \mathcal{M}_d} \sum_{k \in \mathcal{R}_{1,i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m} \\
\text{Cap}_{2,i,j} &= \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_d} \sum_{f \in \mathcal{F}_c} \sum_{k \in \mathcal{R}_{1,i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m} \\
\text{Cap}_{3,i,j} &= \sum_{a \in \mathcal{A}_p} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{M}_d} \sum_{f \in \mathcal{F}_c} \sum_{k \in \mathcal{R}_{1,i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_p} \sum_{m \in \mathcal{M}_d} \sum_{k \in \mathcal{R}_{1,i,j}} r_{a,k,f}^{s,m,d} v_{a,f}^{s,m} \\
\text{Cost}_c &= \Lambda \sum_{i,j \in \mathcal{E}_1} \sum_{h} \text{Cap}_{h,i,j}
\end{align*}
\]

Variables

\[
\begin{align*}
x_{a,f}^{s,m} &\in \{0, 1\} \\
r_{a,k,f}^{s,m} &\in \{0, 1\}
\end{align*}
\]
Mathematical Formulation (cont.)

Constraints:

Processing/Storage Assignments

\[
\sum_{m \in \mathcal{N}_p} x_{a,f}^{s,m} = 1, \forall (a \in \mathcal{A}_p, s \in \mathcal{N}_g, f \in \mathcal{N}_g \cup \mathcal{F}_c)
\]

\[
\sum_{f \in \mathcal{N}_g} x_{a,f}^{s,m} = 1, \forall (a \in \mathcal{A}_s, s \in \mathcal{N}_g, m = f)
\]

\[
\sum_{m \in \mathcal{N}_p} \sum_{f \in \mathcal{N}_g} x_{a,f}^{s,m} = 1, \forall (a \in \mathcal{A}_s, s \in \mathcal{N}_g)
\]

Solenoidality

\[
\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} = x_{a,f}^{s,m}, \forall (a \in \mathcal{A}_p \cup \mathcal{A}_s, s \in \mathcal{N}_g, m \in \mathcal{N}_p, f \in \mathcal{N}_g \cup \mathcal{N}_s \cup \mathcal{F}_c)
\]

\[
\sum_{k \in \mathcal{R}_{i,j}} r_{a,s,m}^{t,s,m} = x_{a,f}^{s,m}, \forall (a \in \mathcal{A}_p \cup \mathcal{A}_s, s \in \mathcal{N}_g, m \in \mathcal{N}_p, f \in \mathcal{N}_g \cup \mathcal{N}_s)
\]

\[
\sum_{m \in \mathcal{N}_c, k \in \mathcal{R}_{i,j}} r_{a,k,f}^{t,s,m,d} = x_{a,f}^{s,m}, \forall (a \in \mathcal{A}_p, s \in \mathcal{N}_g, f \in \mathcal{F}_c)
\]

Processing Delay

\[
\gamma_{a,m} = \alpha_a \kappa_a t_m
\]

Latency – Pt to Pt, Storage

\[
\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_{k}^{s,m} \leq \theta_{a,f}, \forall (a \in \mathcal{A}_p \cup \mathcal{A}_s, s \in \mathcal{N}_g, m \in \mathcal{N}_f \cup \mathcal{N}_s \cup \mathcal{F}_c)
\]

Latency – Local Proc, Proc/Storage

\[
\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_{k}^{s,m} + \gamma_{a,m} + \sum_{k} r_{a,k,f}^{t,s,m} D_{k}^{m,d} \leq \theta_{a,f,m}, \forall (a \in \mathcal{A}_p \cup \mathcal{A}_s, s \in \mathcal{N}_g, m \in \mathcal{N}_p, f \in \mathcal{N}_g \cup \mathcal{N}_s)
\]

Processing Capacity

\[
\sum_{a \in \mathcal{A}_p \cup \mathcal{A}_s} \alpha_a \sum_{s \in \mathcal{N}_g} \sum_{f \in \mathcal{N}_g \cup \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} \leq C_m, \forall m \in \mathcal{N}_p
\]

Storage Capacity

\[
\sum_{a \in \mathcal{A}_s} \Delta_a \sum_{s \in \mathcal{N}_g} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_s} \beta_a \Delta_a \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_p} x_{a,f}^{s,m} v_{a,f}^{s,m} \leq S_f, \forall f \in \mathcal{N}_s
\]

Latency – Global Destination

\[
\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_{k}^{s,m} + \gamma_{a,m} + \sum_{k} r_{a,k,f}^{t,s,m,d} D_{k}^{m,d} \leq \theta_{a,f}, \forall (a \in \mathcal{A}_p, s \in \mathcal{N}_g, m \in \mathcal{N}_p, d \in \mathcal{N}_c, f \in \mathcal{F}_c)
\]

M. Tornatore: Application-Aware Resource Provisioning
Simulation Setup

- **Traffic Volume – 5 Tbps**

- **Profiles**
  - Computational complexity: 0.005 - 0.03 CPU/Mbps
  - Compression factor: 0.1 – 1
  - Latency: 10 – 100 ms
    - Real-time ~ 10-50 ms
    - Near real-time ~ 50-100 ms

- **Network resource costs**
  - Processing cost: 25, 50, 70, 90 $/CPU/Mo
    - Tier 4 (DC) costs [Google]
  - Storage cost: 2.50, 5, 7, 9 $/TB/Mo
  - Metro/Core BW ~ 1$ / Mbps

https://cloud.google.com/compute/pricing
https://cloud.google.com/storage/pricing#pricing-example-simple
http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php#
Results: latency effect

**Take-Away 1**: Total cost decreases and stabilizes as application traffic migrates to cloud thanks to less restrictive latency budget.

**Take-Away 2**: With increasing latency budgets, cloud processing for high complexity applications increases faster than for low complexity apps.
Simulation Results (cont.)

**Take-Away 4:** Total cost is stable, but processing location depends on compression factor!

**Take-Away 3:** For local traffic, cost increases with computation factor due to increase of WAN and MAN BW. **Processing location is unaffected**
Conclusion

• Motivation
  • Lack of quantitative analysis of how specific application traffic affects resource provisioning

• Modeling work
  • a parameterized application profile: \( A = A_p \cup A_s \cup A_{sp} \cup A_n \)
    • \( \Theta, \kappa, \alpha, \beta, \Lambda \)
  • Network costs for 4-tier hybrid fog-cloud architecture

• Developed a model for resource assignment
  • High flexibility: decoupling of storage and computing

• Simulation Results
  • Show impact not only of latency&BW, but also other aspects (compression factor, etc…)
massimo.tornatore@polimi.it
Slice Per Application

Slice 1 - $\alpha \in A_p$

Slice 2 - $\alpha \in A_{sp}$

Slice 3 - $\alpha \in A_s$

Physical

1KCPU/PB

M. Tornatore: App.-Aware Res. Prov. in a Heterog. IoT
Slice Priority

- Previous works categorize IoT/M2M slices/usage scenarios as:
  - Ultra-reliable and low latency communications (URLLC): autonomous driving, emergency services, automated manufacturing, remote medical surgery
  - Enhanced Mobile Broadband (eMBB): streaming video, high capacity multimedia, AR/VR
  - Massive Machine Type Communication (mMTC): (low power) sensor networks, smart metering, city, home (huge number of devices), less latency constrained
- Specific applications with parameterized profiles are assigned a slice of resources, which is then prioritized in a certain class
- Critical – Emerg. services, life/health/safety, remote surgery, auto. driving, factory automation/actuation
- Standard – AR/VR, gaming, Pokemon, smart grid/metering
- Best Effort – sensor data with no real-time actuation

Slice Priority (cont.)

Critical

Standard

Best Effort

M. Tornatore: App.-Aware Res. Prov. in a Heterog. IoT
1) Congestion threshold reached on slice 1:link 1-2, and node 2 (compute): 2 Gbps and 100 CPUs
2) Check for internal slice resources (new path, etc)
3) Find all BE slices with respective resources
4) Calc. donor subslice solution which minimizes impact
5) Transfer subslice(s)
6) Return subslice when lower threshold reached
Questions
# IoT/M2M Application Popularity

<table>
<thead>
<tr>
<th>Applications</th>
<th>Overall popularity</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Home</td>
<td>100%</td>
<td>61k</td>
</tr>
<tr>
<td>Wearables</td>
<td>63%</td>
<td>33k</td>
</tr>
<tr>
<td>Smart City</td>
<td>34%</td>
<td>41k</td>
</tr>
<tr>
<td>Smart grid</td>
<td>28%</td>
<td>41k</td>
</tr>
<tr>
<td>Industrial internet</td>
<td>25%</td>
<td>10k</td>
</tr>
<tr>
<td>Connected car</td>
<td>19%</td>
<td>5k</td>
</tr>
<tr>
<td>Connected Health</td>
<td>6%</td>
<td>2k</td>
</tr>
<tr>
<td>Smart retail</td>
<td>2%</td>
<td>1k</td>
</tr>
<tr>
<td>Smart supply chain</td>
<td>2%</td>
<td>0k</td>
</tr>
<tr>
<td>Smart farming</td>
<td>1%</td>
<td>1k</td>
</tr>
</tbody>
</table>

1. Monthly worldwide Google searches for the application
2. Monthly Tweets containing the application name and #IOT
3. Monthly LinkedIn Posts that include the application name. All metrics valid for Q4/2014.

Sources: Google, Twitter, LinkedIn, IoT Analytics

Source: Google, Twitter, IoT Analytics, 2014.
Hybrid Fog-Cloud Architecture – Hierarchical

$P_t$: Tier t processing
$S_t$: Tier t storage

$P_2S_2$

$P_3S_3$

$P_4S_4$

13

16 15 14 12 11 10 8 7 6 4 3 2

19 DC 1

17

18

Core

Metro

DC 2

Access

Capacity (+)

Unit Cost (+)

M. Tornatore: Application-Aware Resource Provisioning
M. Tornatore: Application-Aware Resource Provisioning
Functional Scenarios

source \( s \) → processing \( m \) → destination \( f \)

\[ \nu_{a}^{s,f} \]

\[ \alpha_{a} \nu_{a}^{s,f} \]

\[ \beta_{a} \nu_{a}^{s,f} \]

\[ a \in A_{p}, f \in N_{g} \text{ (local)} \]

source \( s \) → processing \( m \) → core CO \( d \) → destination \( f \)

\[ \nu_{a}^{s,f} \]

\[ \alpha_{a} \nu_{a}^{s,f} \]

\[ \beta_{a} \nu_{a}^{s,f} \]

\[ a \in A_{p}, m \in N_{pl}, f \in F_{c} \text{ (global)} \]

source \( s \) → processing \( m \) → core CO \( f \)

\[ \nu_{a}^{s,f} \]

\[ \alpha_{a} \nu_{a}^{s,f} \]

\[ a \in A_{p}, m \in N_{DC}, f \in F_{c} \text{ (global)} \]
Functional Scenarios (cont.)

source → processing → destination

- $s \rightarrow m \rightarrow f$
  - $\nu_{a}^{s,f}$
  - $\nu_{a}^{s}$
  - $\nu_{a}^{s,f}$
  - $\nu_{a}^{s}$

- $s \rightarrow m, f$
  - $\alpha_{a} \nu_{a}^{s}$
  - $\beta_{a} \nu_{a}^{s}$

- $s \rightarrow f$
  - $\Delta_{a} \nu_{a}^{s}$

destination

- $a \in A_{n}, f \in N_{g} \text{ (global)}$
- $a \in A_{n}, f \in F_{c} \text{ (global)}$
- $a \in A_{sp}$
- $a \in A_{s}$

processing/storage

storage only
**Take-Away 1:** For increasing computational complexity and real-time services, fog processing costs steadily increase (no other choice!)

**Take-Away 2:** Cloud processing costs of real-time traffic start to decrease at 0.02 while near real-time cloud processing costs continue to increase with complexity.

**Take-Away 1:** Cloud processing costs increase at much slower rate with increasing complexity for real-time traffic as DC compute locations restrict more RT traffic to fog processing.