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

UCDAVIS

ONDM 2017
21th International Conference on Optical Network Design and Modeling
May 15-17, 2017 | Budapest, Hungary

POLITECNICO
MILANO

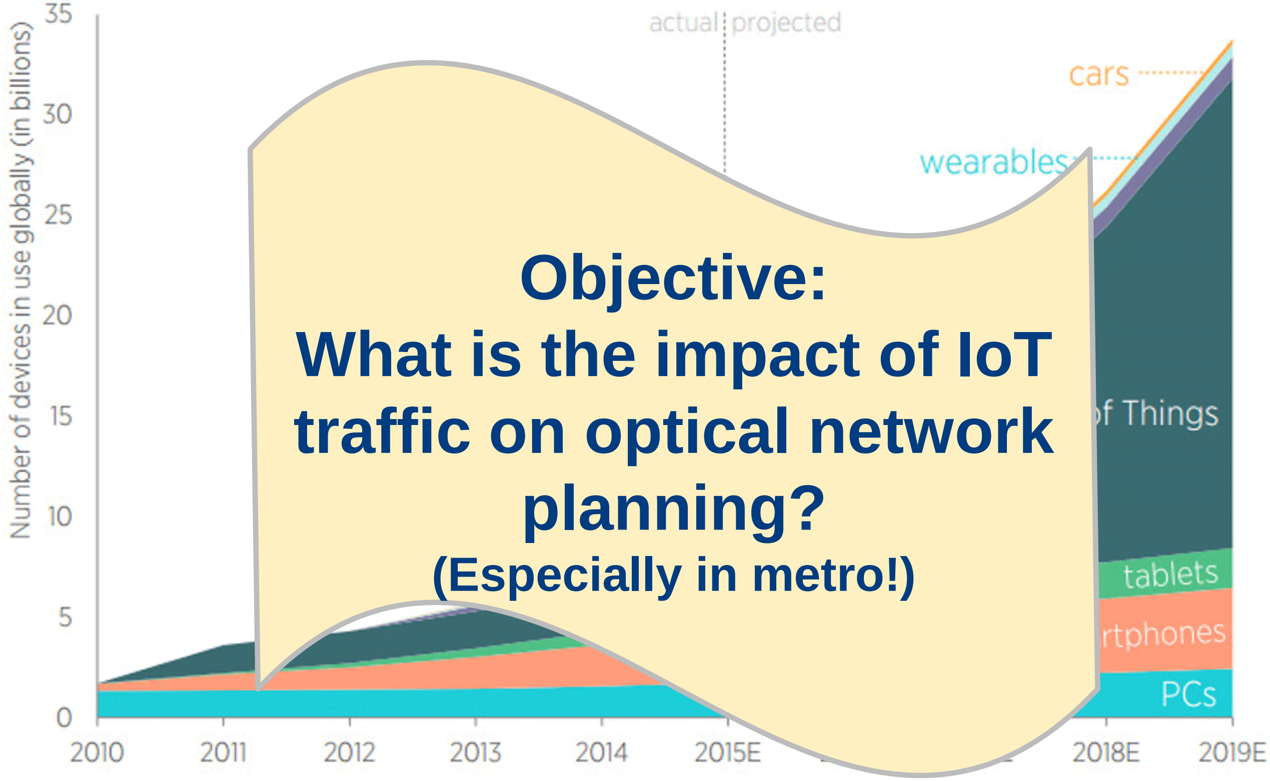
Outline

- Motivation
- Application profiles
- Mathematical formulation
- Simulation Results
- Conclusion

Motivation and objective

- Next big wave:

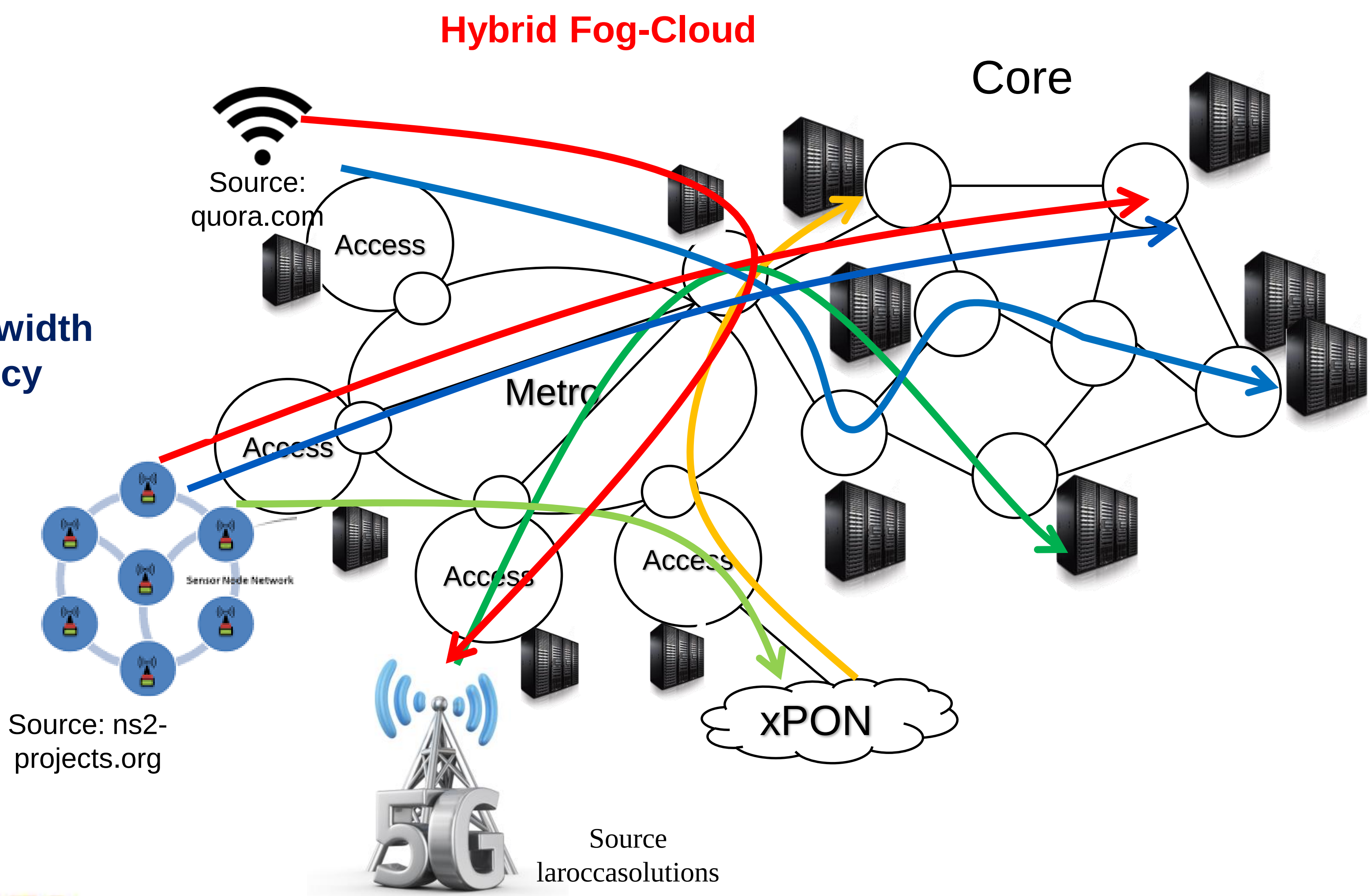
Internet of Things (IoT) and Machine to Machine (M2M) traffic



Source: John Greenough, "The Internet of Everything 2015," *Business Insider Intelligence*. Produced by Adam Thierer and Andrea Castillo, Mercatus Center at George Mason University, 2015.

Enabler: hybrid fog-cloud computing

- **Drivers**
  - **Bandwidth**
  - **Latency**



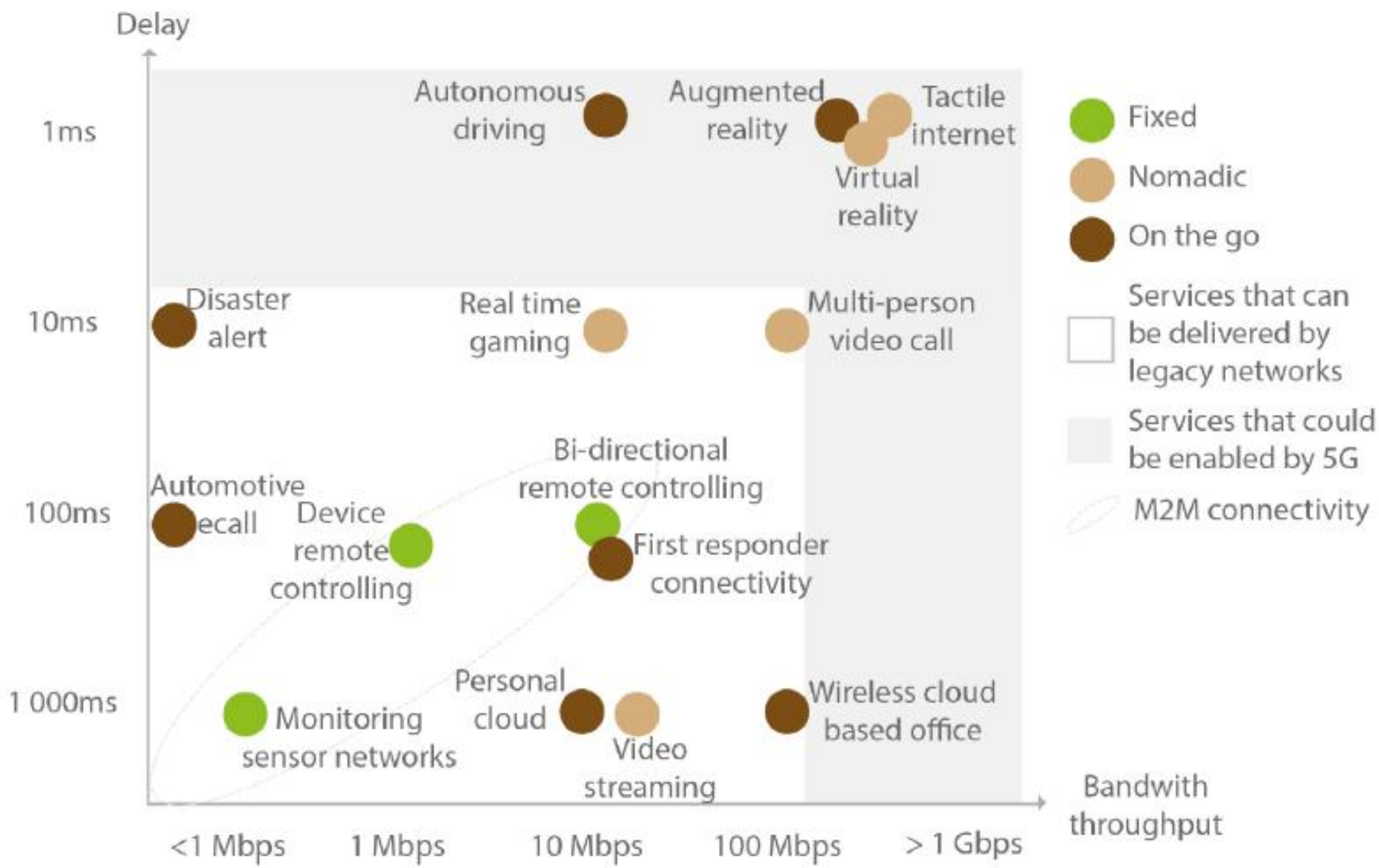
Challenges

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
  - **Θ: Latency budget** (source to destination)
  - **κ: Bandwidth**
  - **α: Computational complexity** (per unit of traffic)
  - **Λ: Storage time**
  - **β: Compression factor** (ratio of processed-to-raw data)

Examples	Θ (ms)	κ (Mbps)	α (CPU/Mbps)	β	Λ (hrs)
1- AR/VR	10	100	0.03	0.6	0
2 – Factory Automation	20	1	0.009	0.8	10
3 – Data Backup	1000	1	0	0	4
4 – Smart Grid	50	0.4	0.007	0.3	0
5 – Smart Home	60	.001	0	0	0
6 – Medical	40	2	0.02	0.2	0.1
7 – Environmental Mon.	1000	1	0.02	0.1	100
8 – Tactile Internet	1	200	.005	0.8	0

5G PPP, “5G Automotive Vision,” white paper, 2015.

A. Frotzscher *et al*., “Requirements and Current Solutions of Wireless Communication in Industrial Automation,” *Proc.IEEE ICC Wksps.*, Sydney, Australia, 2014, pp. 67–72.

Application Profile (3)

- Additional note: we classify applications in four categories

$$A = A_p \cup A_s \cup A_{sp} \cup A_n$$

Point to Point (basic M2M)

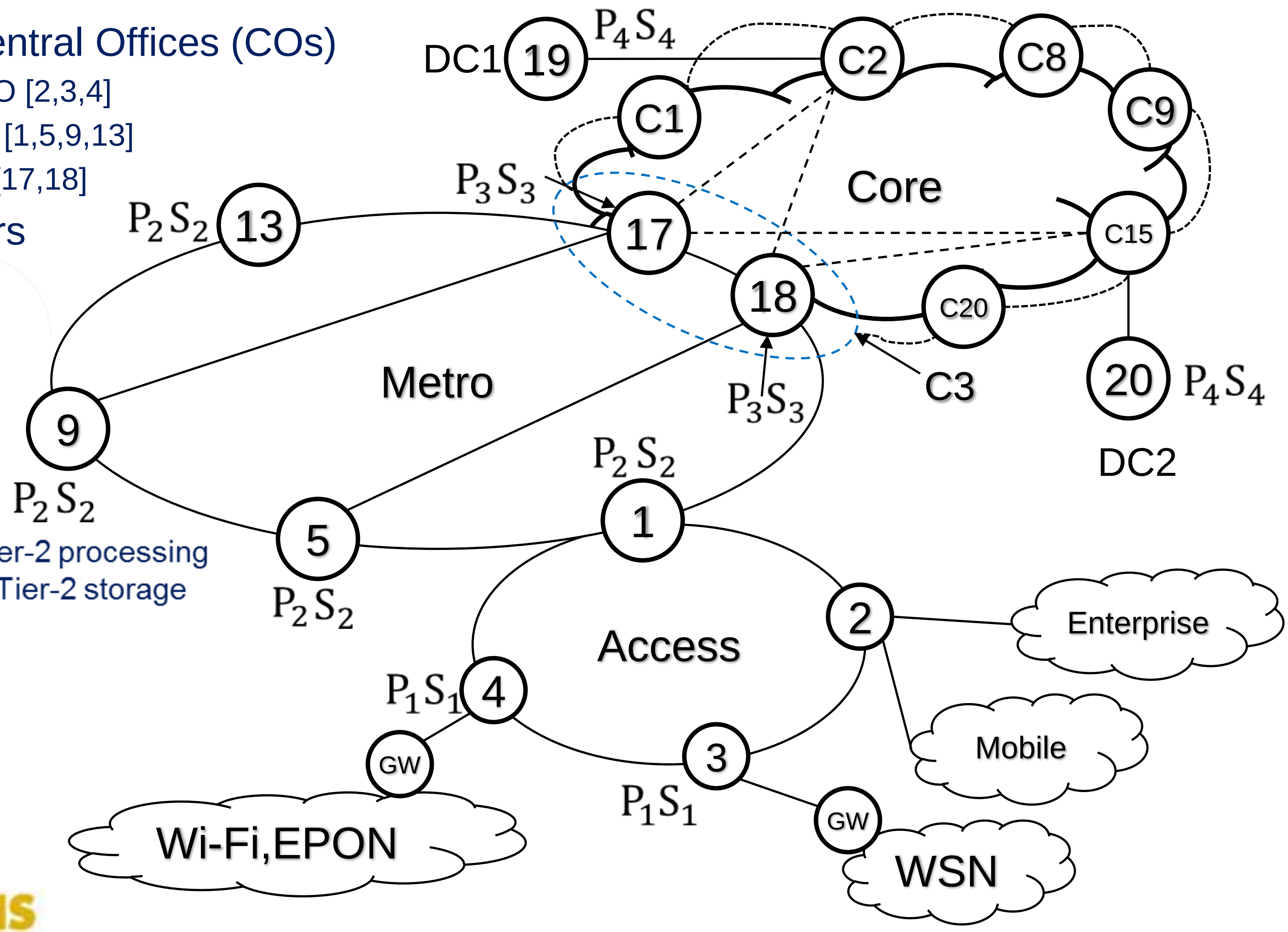
Processing & Storage (smart grid, ...)

Storage only (analytics repositories, …)

Processing only (cloud gaming, virtual reality, …)

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_2: Tier-2 processing
S_2: Tier-2 storage

Network (Cost) Parameters

- Costs (in COs)
  - μ: processing cost
  - ν: storage cost
  - Λ: metro bandwidth cost
  - $\varepsilon_{up}, \varepsilon_{down}$: core bandwidth cost

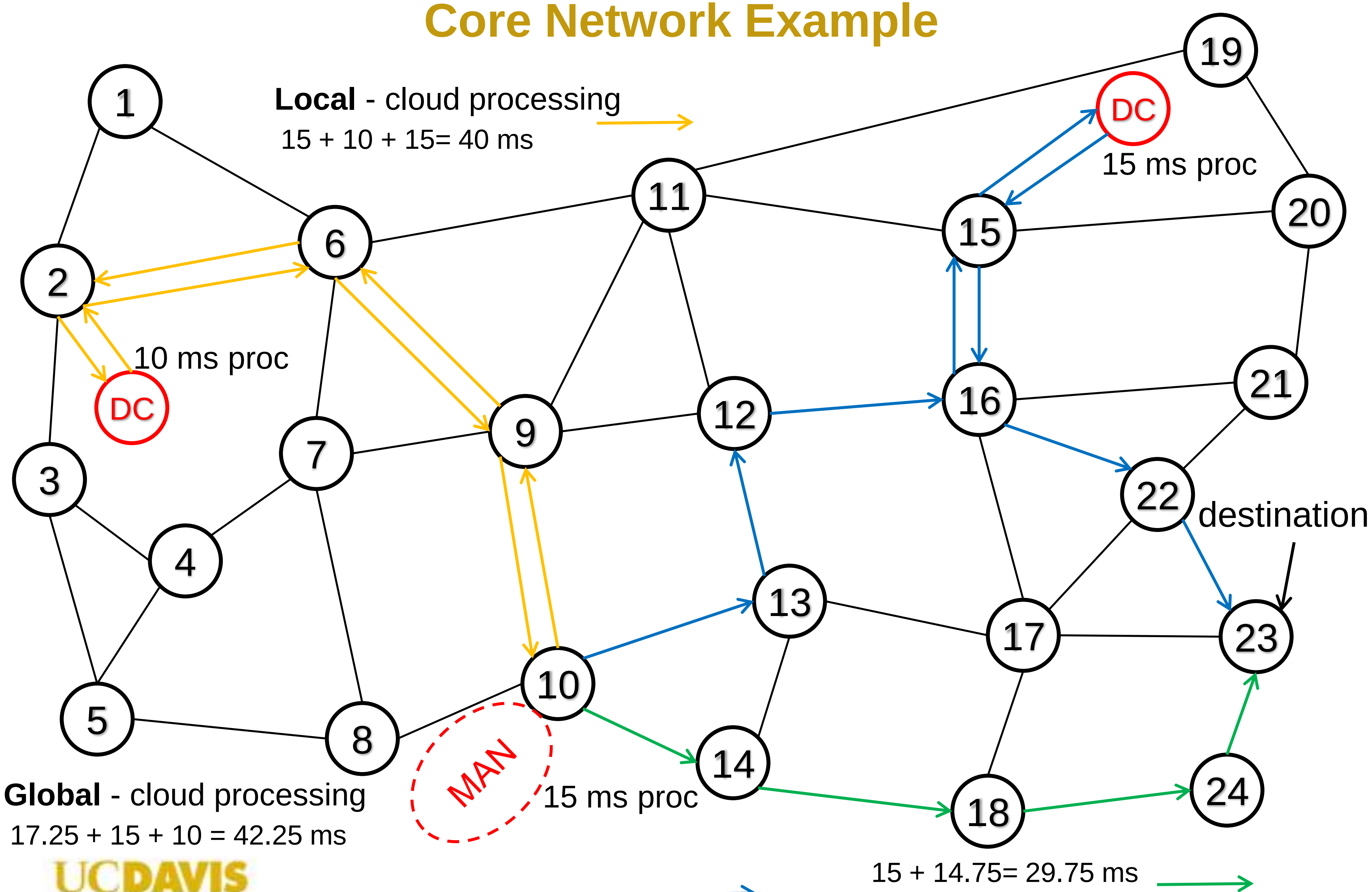
Tier	μ ($/CPU/Mo)	ν ($/GB/Mo)	Λ ($/Mbps/Mo)	$\varepsilon_{up}, \varepsilon_{down}$ ($/Mbps/Mo)
1 – Access CO	90	0.0042	~1	~1/1
2 – Metro CO	70	0.004	~1	~1/1
3 – Core CO	50	0.0035	~1	~1/1
4 - DC	25	0.0025	~1	~1/1

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#

Problem Statement

- **Inputs**
  - Offered traffic (per s-d pair, per application)
  - Application profiles:
    - Θ, κ, α, β, Δ
  - 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



Mathematical Formulation

Objective Function:

$$min(Cost_p + Cost_s + Cost_u + Cost_d + Cost_c)$$

Variables

$x_{a,f}^{s,m} \in \{0,1\}$

$r_{a,k,f}^{s,m} \in \{0,1\}$

$r_{a,k,f}^{\prime,s,m} \in \{0,1\}$

$r_{a,k,f}^{\prime\prime,s,m} \in \{0,1\}$

1 *Processing*

$$Cost_p = \sum_{m \in \mathcal{N}_p} \mu_m \sum_{a \in \mathcal{A}_p \cup \mathcal{A}_{sp}} \alpha_a \sum_{s \in \mathcal{N}_g} \sum_{f \in \mathcal{N}_g \cup \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m}$$

2 *Storage*

$$Cost_s = \sum_{a \in \mathcal{A}_s} \Delta_a \sum_{f \in \mathcal{N}_s} \nu_f \sum_{s \in \mathcal{N}_g} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_{sp}} \beta_a \Delta_a \sum_{f \in \mathcal{N}_s} \nu_f \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_p} x_{a,f}^{s,m} v_{a,f}^{s,m}$$

3 *Upstream core BW*

$$Cost_u = \epsilon_{up} \Big[\sum_{a \in \mathcal{A}_p \cup \mathcal{A}_{sp}} \beta_a \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_{pl}} \sum_{f \in \mathcal{N}_{DC} \cup \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} + \textstyle\sum_{a \in \mathcal{A}_p \cup \mathcal{A}_s \cup \mathcal{A}_{sp}} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_{DC}} \sum_{f \in \mathcal{N}_g \cup \mathcal{N}_s \cup \mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a \in \mathcal{A}_n} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_c} \sum_{f \in \mathcal{F}_c} v_{a,f}^{s,m} \Big]$$

Downstream core BW
4

$$Cost_d = \epsilon_{down} \textstyle\sum_{a \in \mathcal{A}_p \cup \mathcal{A}_{sp}} \beta_a \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_{DC}} \sum_{f \in \mathcal{N}_g \cup \mathcal{N}_{sl}} x_{a,f}^{s,m} v_{a,f}^{s,m}$$

5 *Metro BW*

$$Cap_{1,i,j} = \sum_{a \in \mathcal{A}_n \cup \mathcal{A}_s} \sum_{s \in \mathcal{N}_g} \Big[\sum_{f \in \mathcal{N}_g \cup \mathcal{N}_{sl}} \sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m} + \sum_{f \in \mathcal{F}_c} \sum_{m \in \mathcal{N}_c} \sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m} \Big]$$

$$Cap_{2,i,j} = \sum_{a \in \mathcal{A}_p \cup \mathcal{A}_{sp}} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_p} \sum_{f \in \mathcal{N}_g \cup \mathcal{N}_s \cup \mathcal{F}_c} \sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} v_{a,f}^{s,m}$$

$$Cap_{3,i,j} = \sum_{a \in \mathcal{A}_p \cup \mathcal{A}_{sp}} \beta_a \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_p} \sum_{f \in \mathcal{N}_g \cup \mathcal{N}_s \cup \mathcal{F}_c} \Big[\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{\prime s,m} + \sum_{d \in \mathcal{N}_c} \sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{\prime\prime s,m,d} \Big] v_{a,f}^{s,m}$$

$$Cost_c = \Lambda \sum_{i,j \in \mathcal{L}_l} \sum_{h} Cap_{h,i,j}$$

Mathematical Formulation (cont.)

Constraints:

Processing/Storage Assignments

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

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

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

Latency – Pt to Pt, Storage

$$\sum_{k\in\mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_k^{s,m} \le \theta_{a,f}, \forall (a\in A_n\cup\mathcal{A}_s, s\in\mathcal{N}_g, m\in\mathcal{N}_l\cup\mathcal{N}_s, f\in\mathcal{N}_g\cup\mathcal{N}_s\cup\mathcal{F}_c)$$

Latency – Local Proc, Proc/Storage

$$\sum_{k\in\mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_k^{s,m} + \gamma_{a,m} + \sum_k r_{a,k,f}^{\prime s,m} D_k^{m,f} \le \theta_{a,f,m}, \forall (a\in\mathcal{A}_p\cup\mathcal{A}_{sp}, s\in\mathcal{N}_g, m\in\mathcal{N}_p, f\in\mathcal{N}_g\cup\mathcal{N}_s)$$

Processing Capacity

$$\sum_{a\in\mathcal{A}_p\cup\mathcal{A}_{sp}} \alpha_a \sum_{s\in\mathcal{N}_g}\sum_{f\in\mathcal{N}_g\cup\mathcal{N}_s\cup\mathcal{F}_c} x_{a,f}^{s,m} v_{a,f}^{s,m} \le C_m, \forall m\in\mathcal{N}_p$$

Solenoidality

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

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

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

Storage Capacity

$$\sum_{a\in\mathcal{A}_s} \Delta_a \sum_{s\in\mathcal{N}_g} x_{a,f}^{s,m} v_{a,f}^{s,m} + \sum_{a\in\mathcal{A}_{sp}} \beta_a \Delta_a \sum_{s\in\mathcal{N}_g}\sum_{m\in\mathcal{N}_p} x_{a,f}^{s,m} v_{a,f}^{s,m} \le S_f,\ \forall f\in\mathcal{N}_s$$

Processing Delay

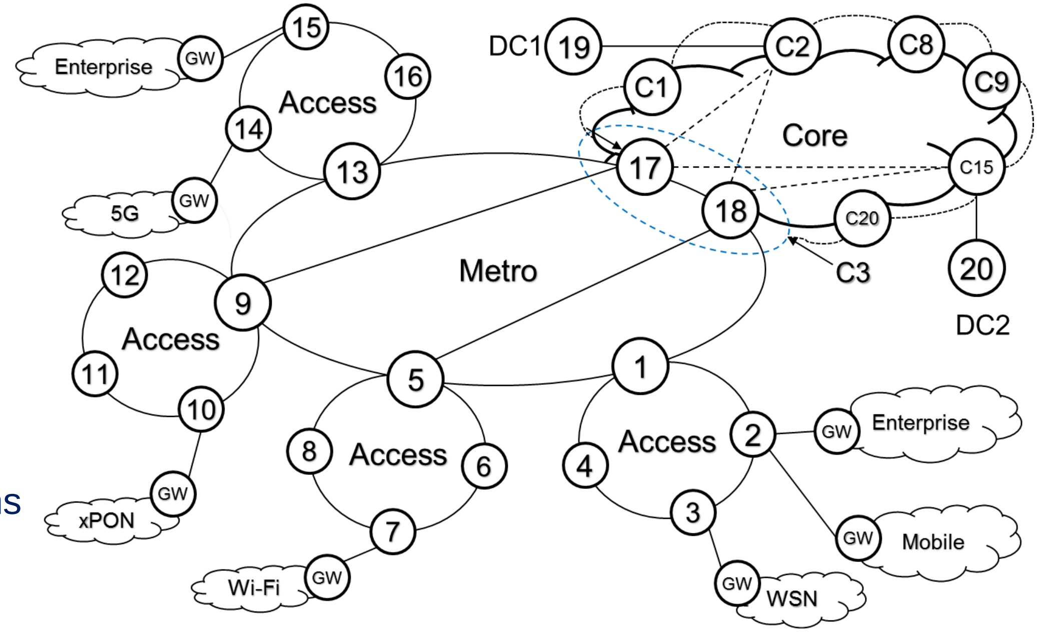
$$\gamma_{a,m} = \alpha_a \kappa_a \tau_m$$

Latency – Global Destination

$$\sum_{k\in\mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_k^{s,m} + \gamma_{a,m} + \sum_k r_{a,k,f}^{\prime\prime s,m,d} D_k^{m,d} \le \theta_{a,f}, \forall (a\in\mathcal{A}_p, s\in\mathcal{N}_g, m\in\mathcal{N}_p, d\in\mathcal{N}_c, f\in\mathcal{F}_c)$$

Simulation Setup

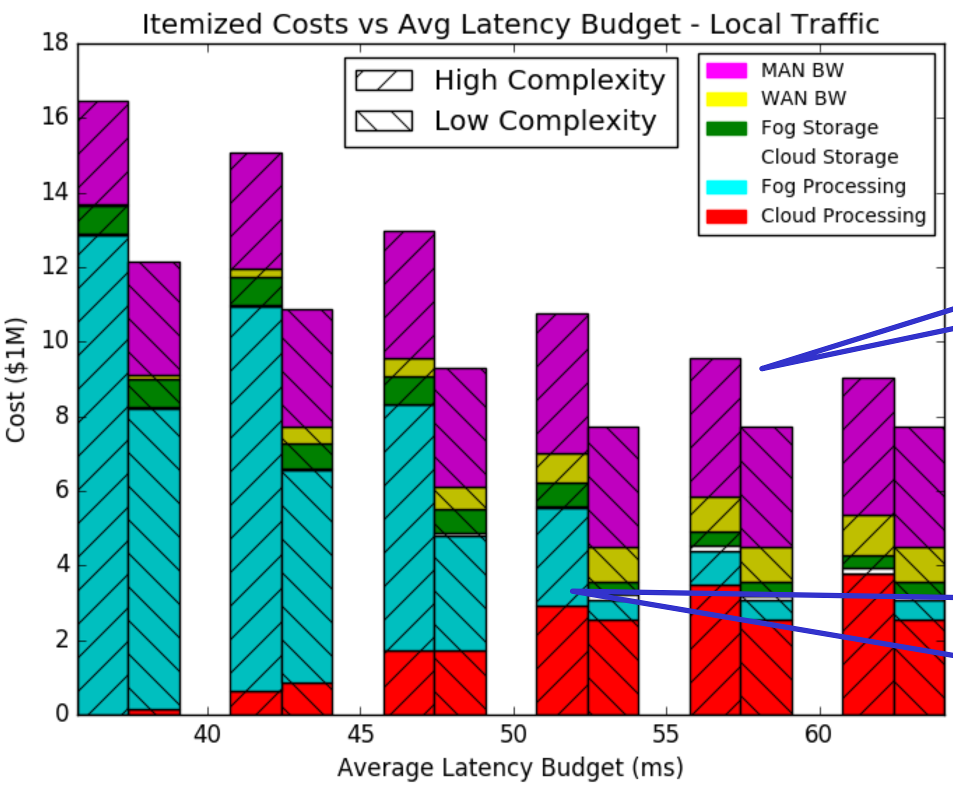
- Traffic Volume – 5 Tbps
- Profiles
  - Computational complexity:
    - 0.005 - 0.03 CPU/Mbps
  - Compression factor: 0.1 – 1
  - Latency: 10 – 100 ms
    - Real-time ~ 10-50 ms
    - Near real-time ~ 50-100 ms



- 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

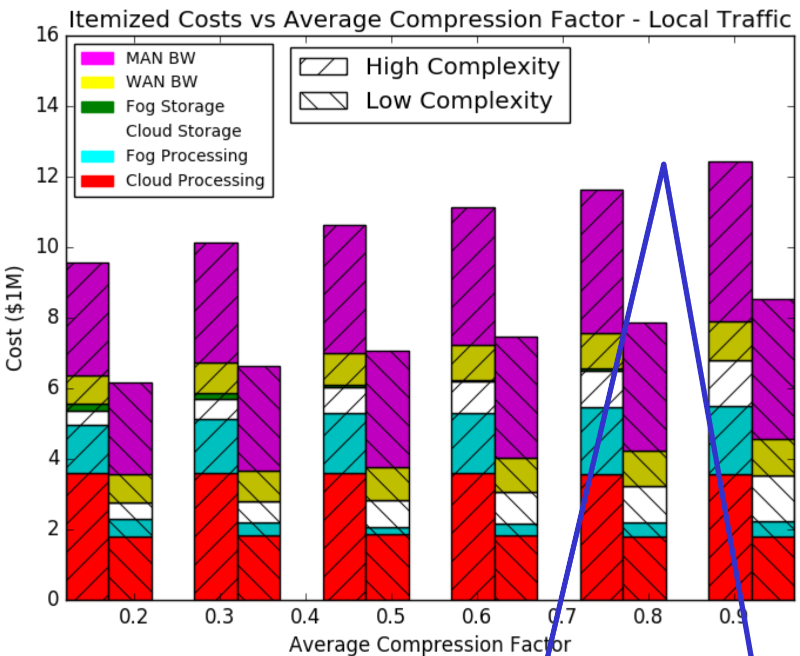


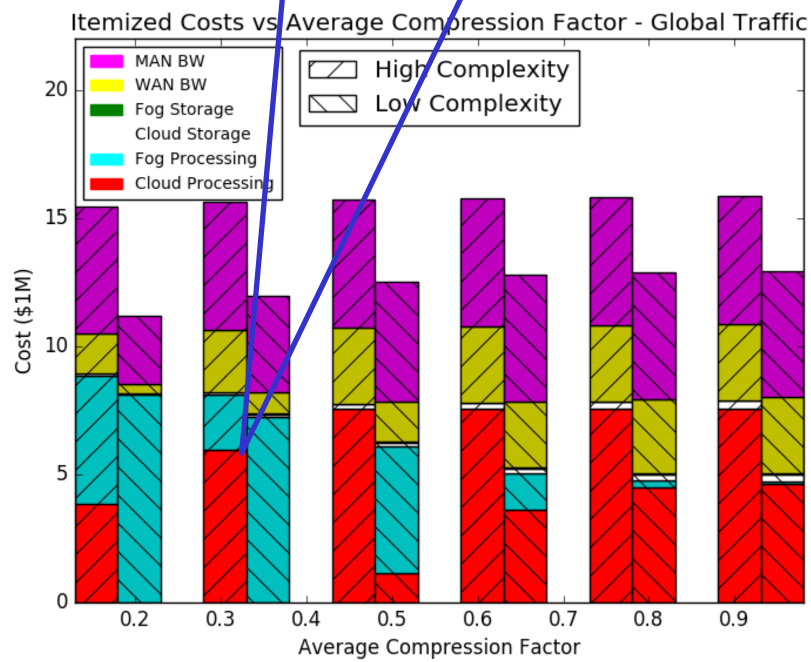
Take-Away 1: Total cost decreases and stabilizes as application traffic migrates to cloud thanks to less restrictive latency budget

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**!

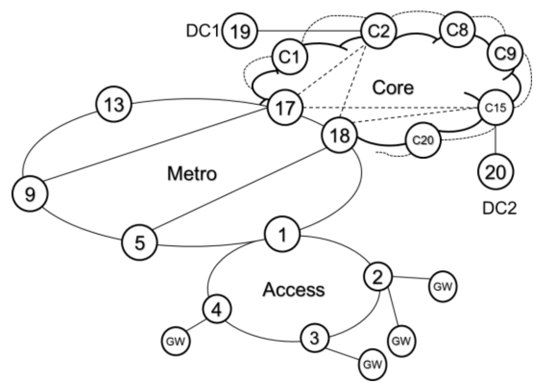


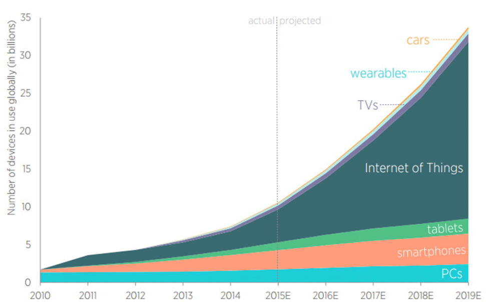


Take-Away 3: For local traffic, cost increases with computation factor due to increase of WAN and MAN BW**. Processing location is unaffected**

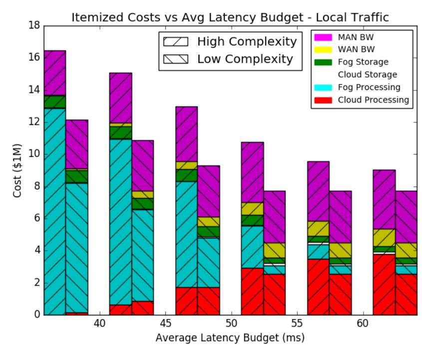
Conclusion

- Motivation
  - Lack of quantitative analysis of how specific application traffic affects resource provisioning
- Modeling work
  - a parameterized application profile: $A = A_p \cup A_s \cup A_{sp} \cup A_n$
    - Θ, κ, α, β, Λ
  - 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…)





Source: John Greenough, "The Internet of Everything 2015," *Business Insider Intelligence*. Produced by Adam Thierer and Andrea Castillo, Mercatus Center at George Mason University, 2015.

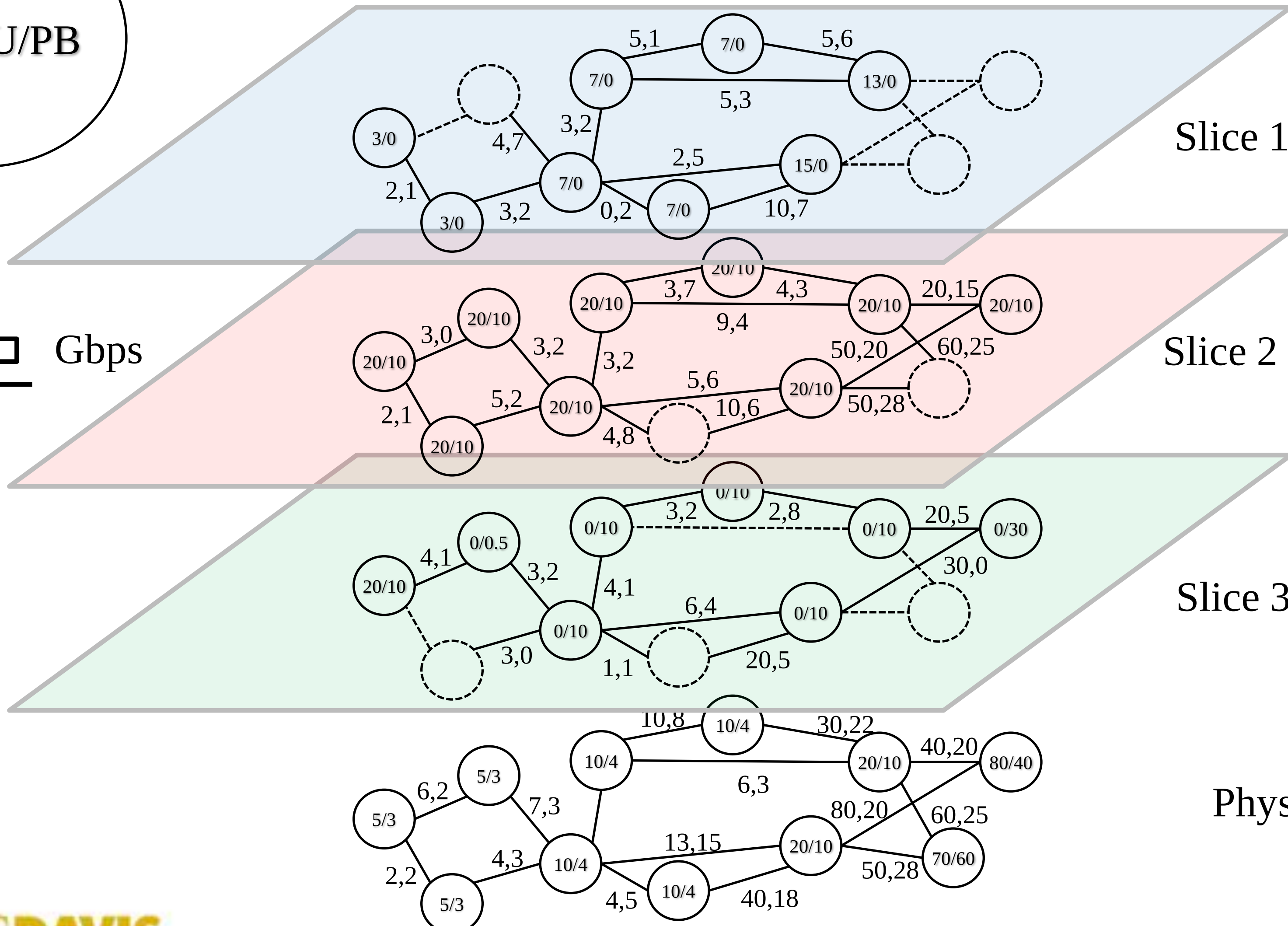


massimo.tornatore@polimi.it

Slice Per Application

1KCPU/PB

Gbps



Slice 1 - $a \in A_p$

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

Slice 3 - $a \in A_s$

Physical

UCDAVIS UNIVERSITY OF CALIFORNIA

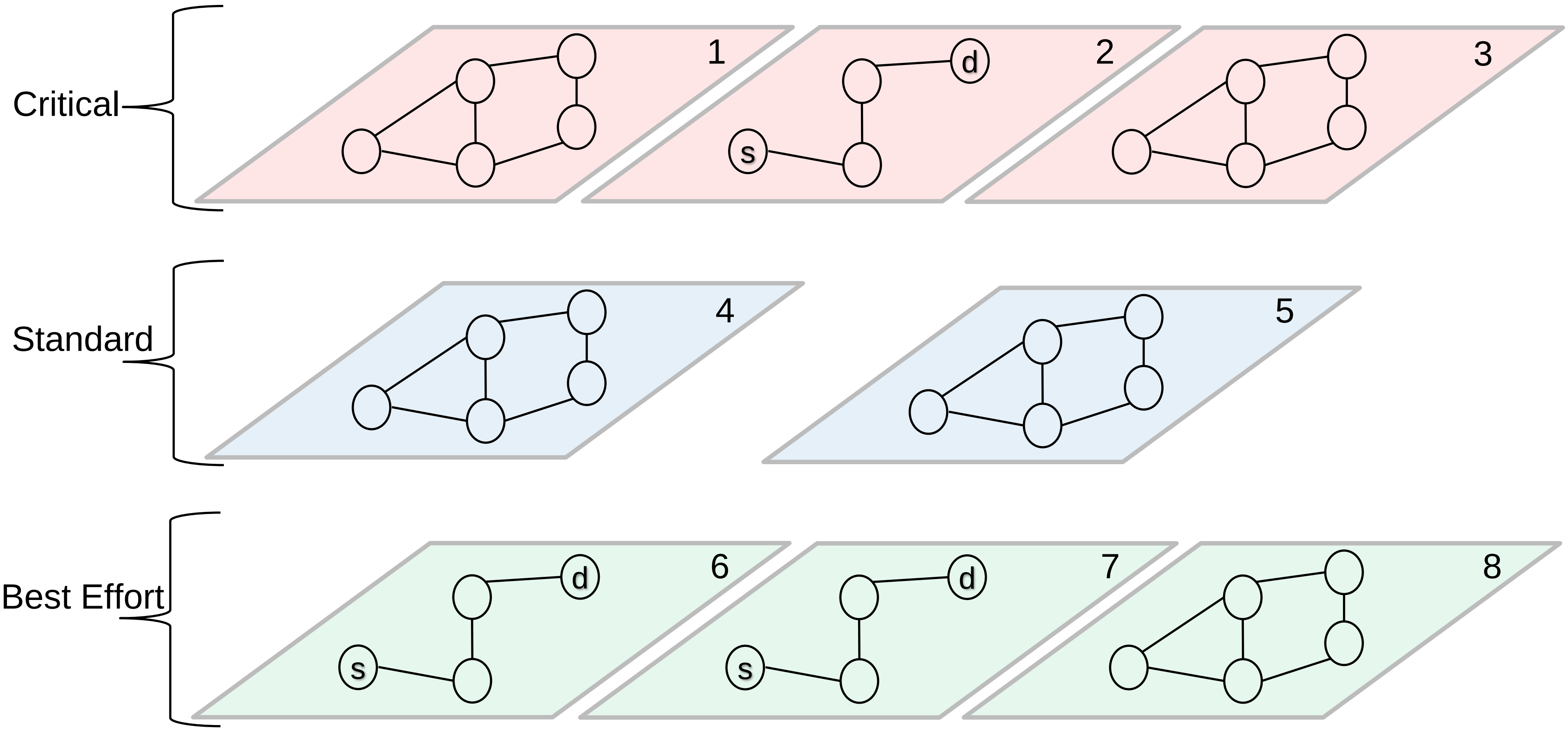
Slice Priority

- 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

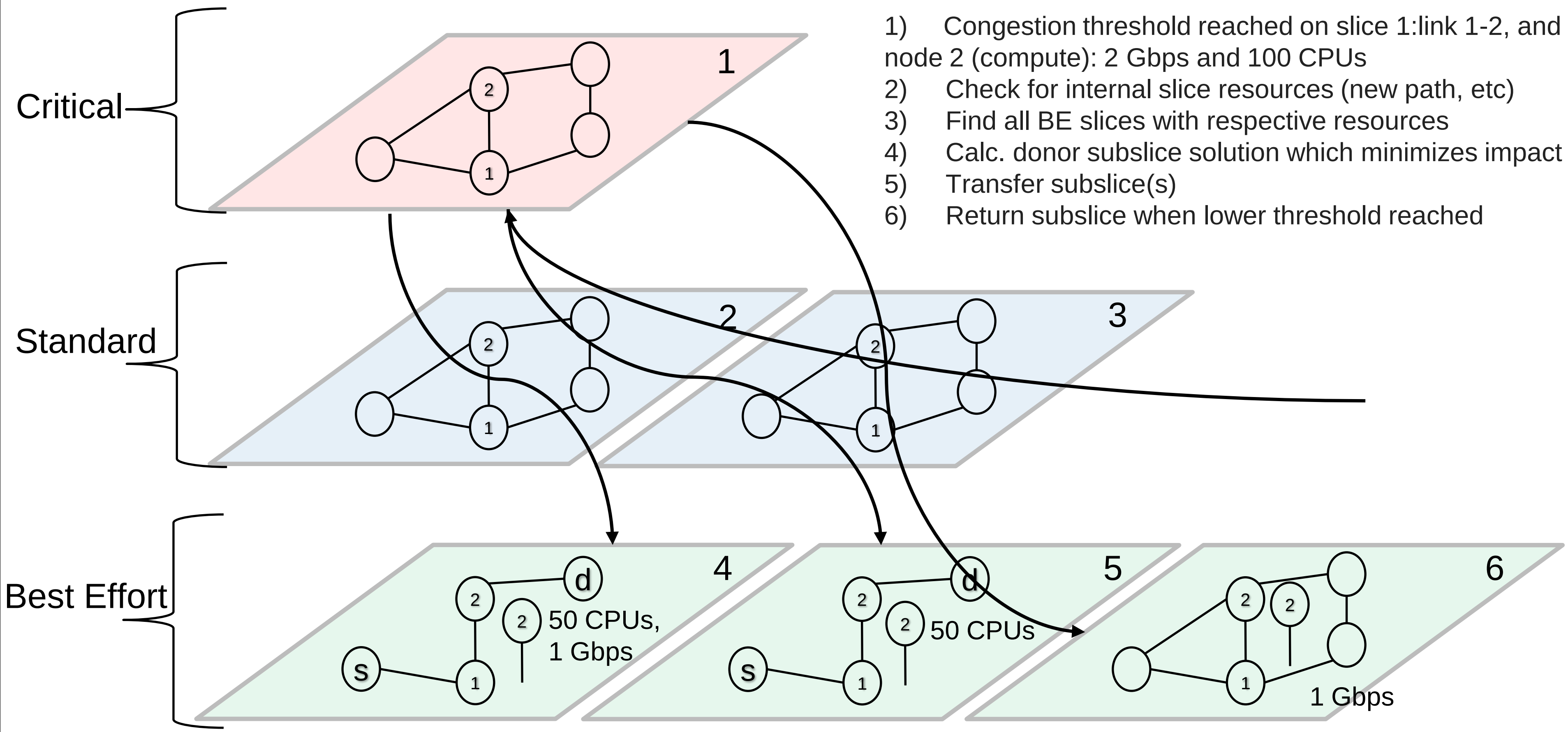
Nakao, A., Du, P., Kiriha, Y., Granelli, F., Gebremariam, A.A., Taleb, T. and Bagaa, M. End-to-End Network Slicing for 5G Mobile Networks. *Journal of Information Processing*, *25*, pp.153-163, 2017.

The Fifth Generation Mobile Communication Forum (5GMF) White Paper. “5G Mobile Communications for 2020 and Beyond.” July, 2016.

Slice Priority (cont.)



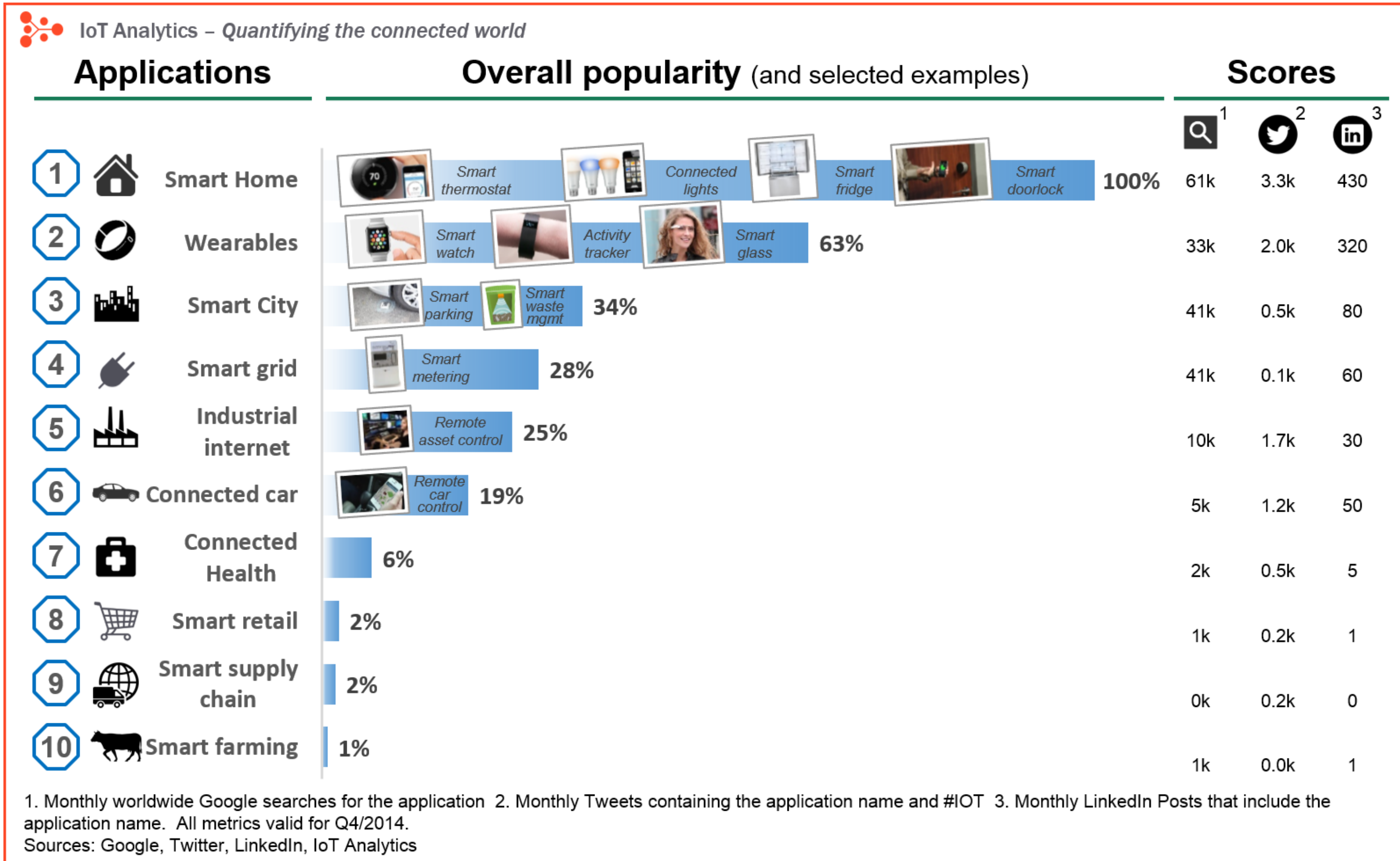
Reslicing



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

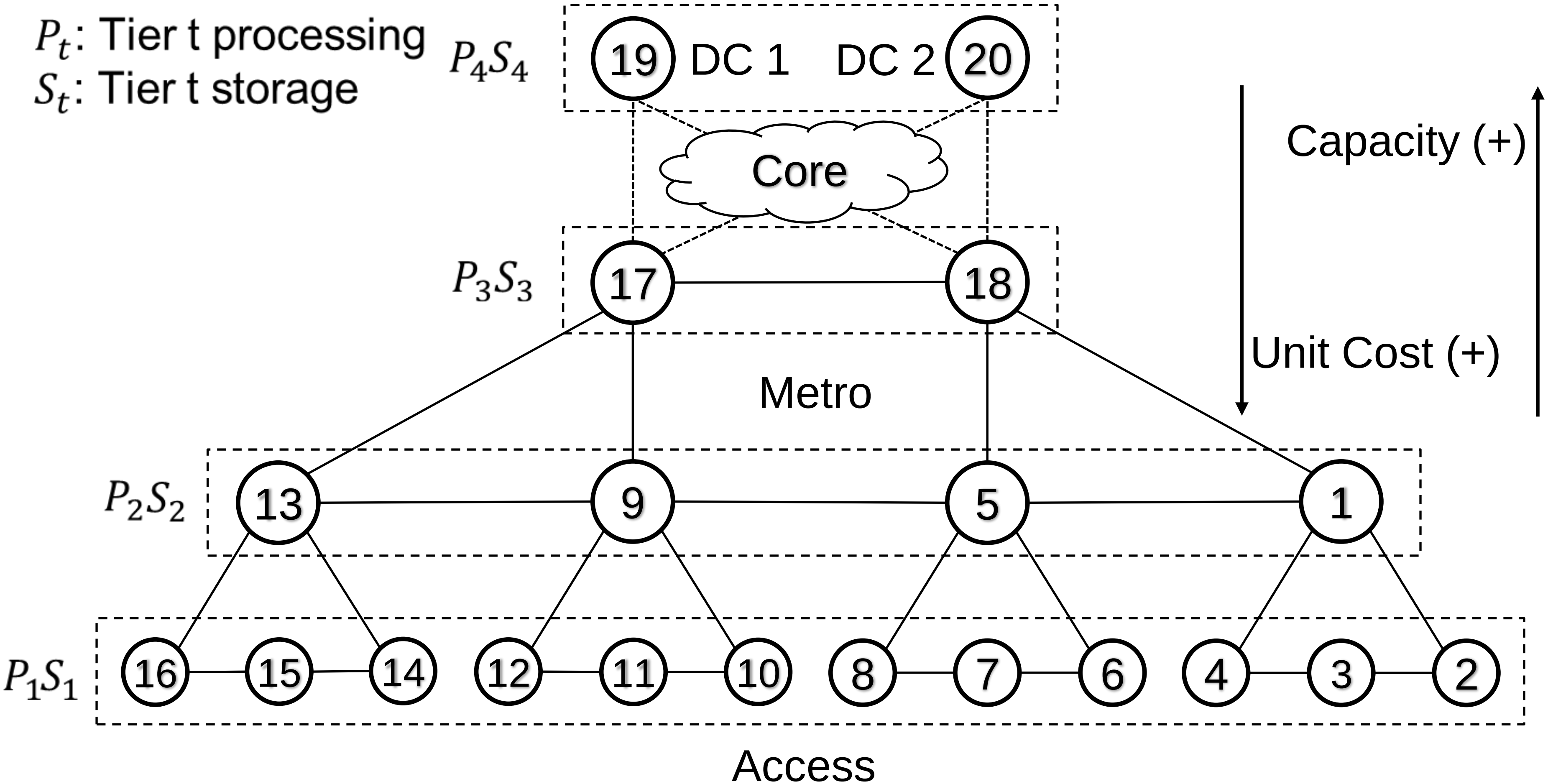
IoT/M2M Application Popularity

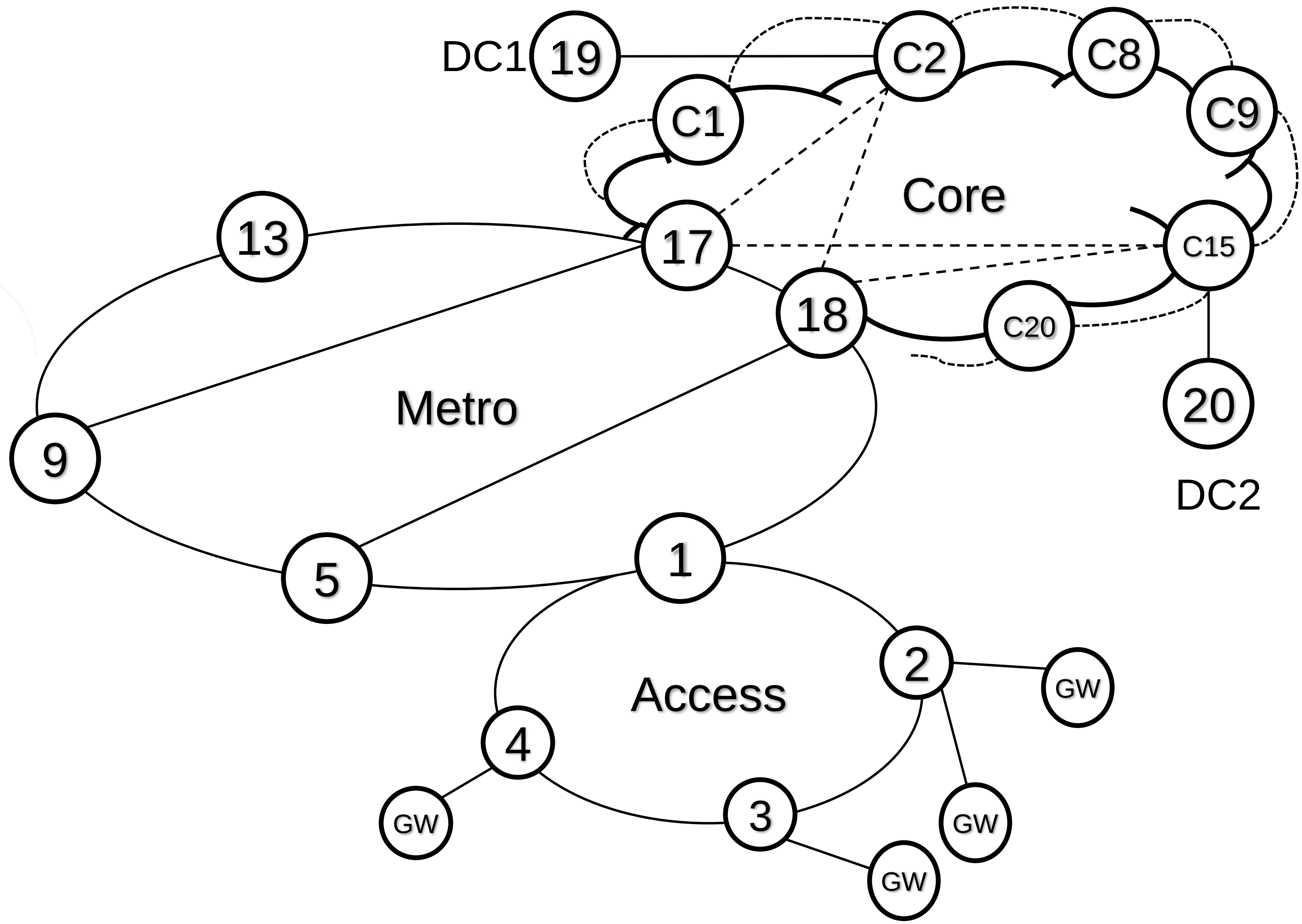


Source: Google, Twitter, IoT Analytics, 2014.

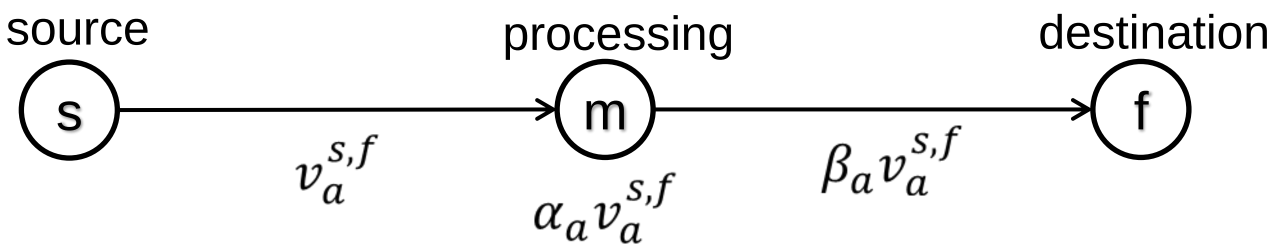
Hybrid Fog-Cloud Architecture – Hierarchical

P_t: Tier t processing
S_t: Tier t storage

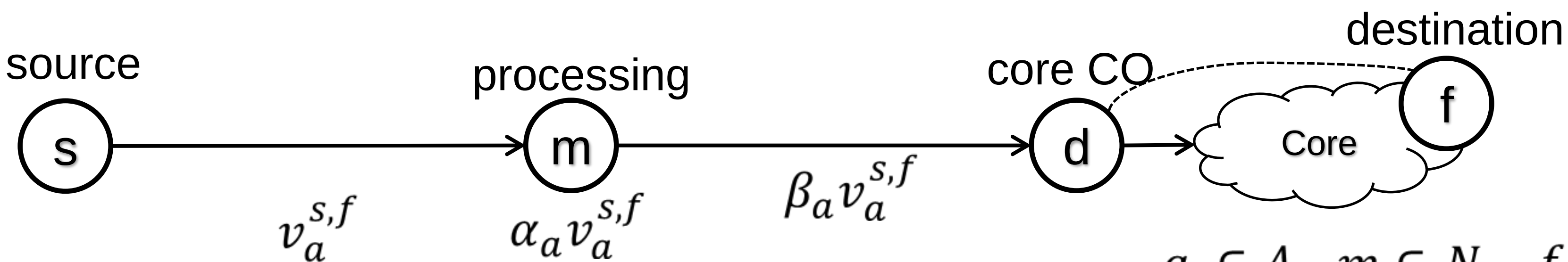



DC1
19
C2
C8
C9
C1
Core
13
17
C15
18
C20
Metro
20
9
DC2
1
5
Access
2
GW
4
3
GW
GW
GW

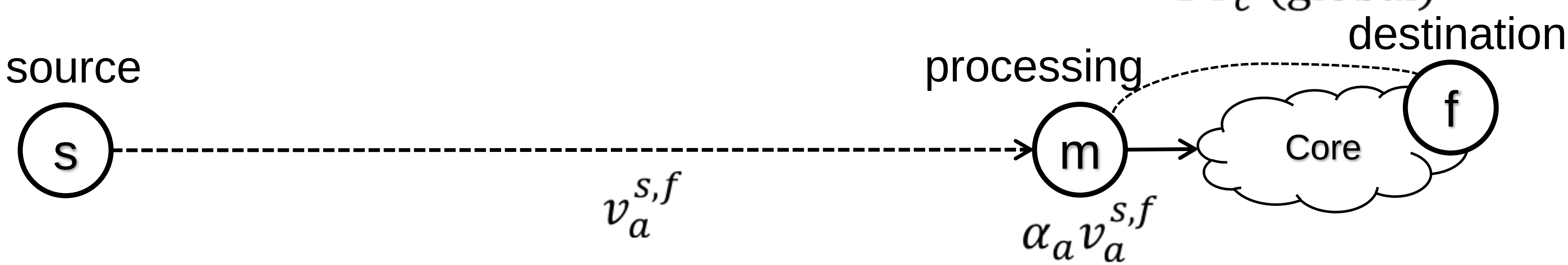
Functional Scenarios



$a \in A_p, f \in N_g$ (local)

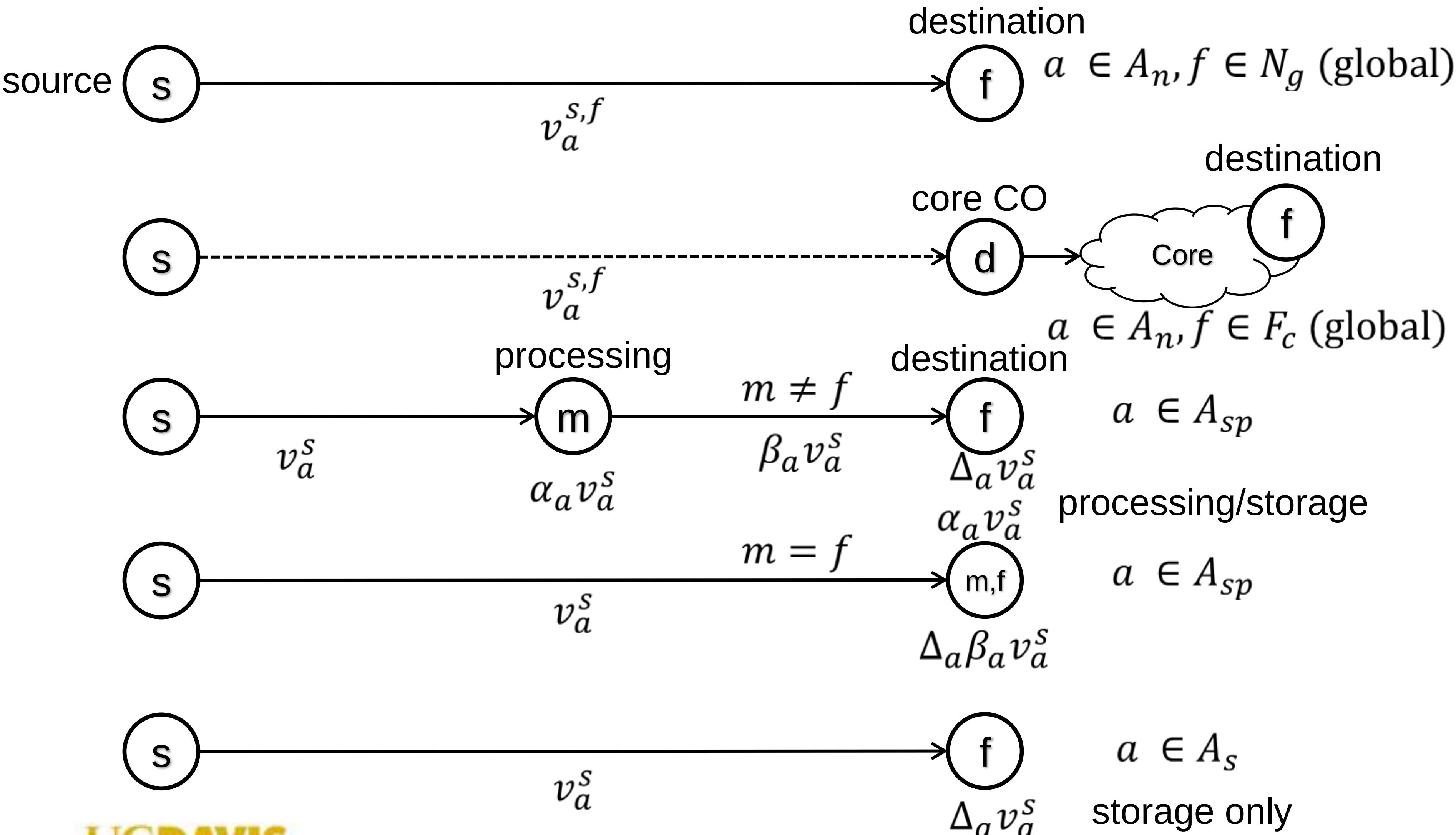


$a \in A_p, m \in, N_{pl}, f \in F_c$ (global)



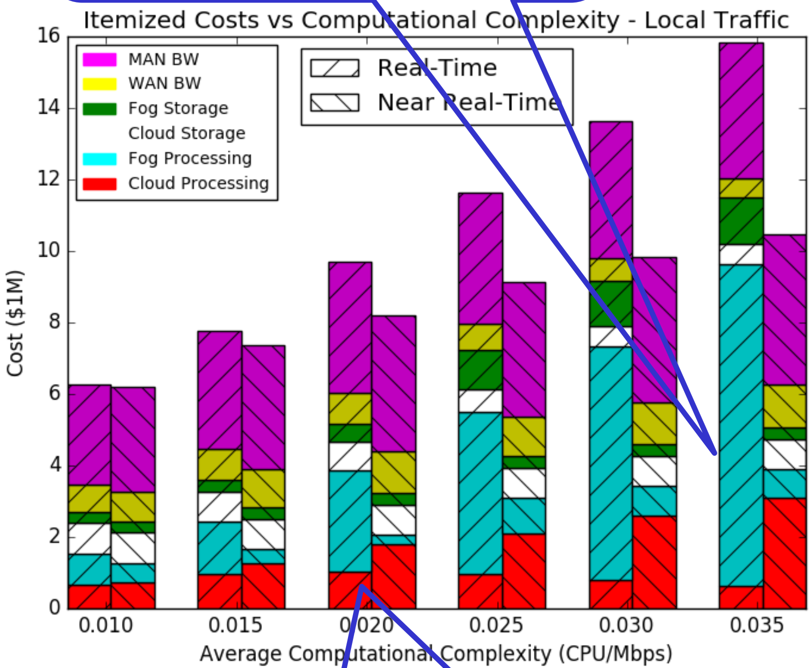
$a \in A_p, m \in, N_{DC}, f \in F_c$ (global)

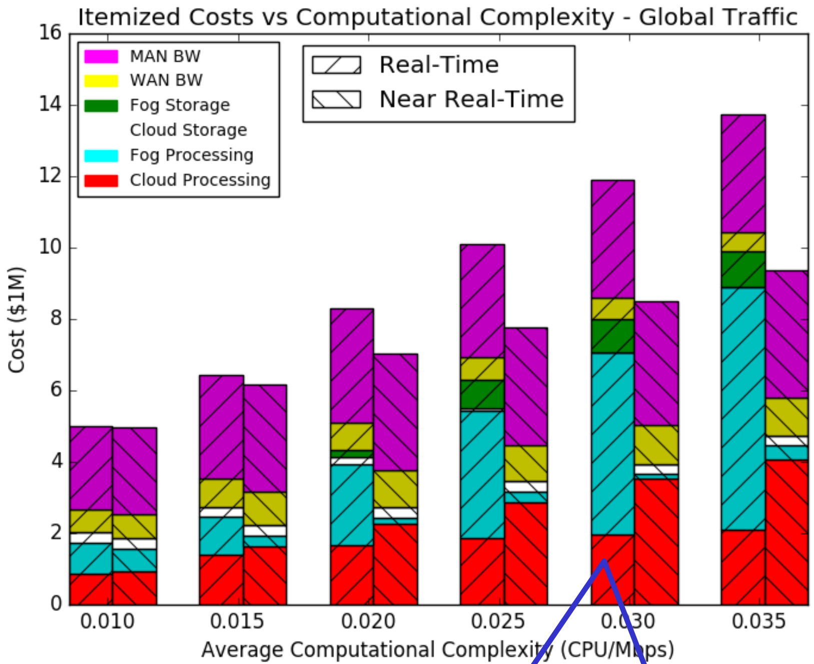
Functional Scenarios (cont.)



Results: Effect of Computation complexity

Take-Away 1: for increasing computational complexity and real time services, fog processing costs steadily increases (no other choice!)





Take-Away 2: Cloud processing costs of real-time traffic start to decrease at .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