Application-Sliced Resource Provisioning for Next-Generation Metro Networks

Eric Sturzinger, Massimo Tornatore, and Biswanath Mukherjee

ONDM

16 May 2017

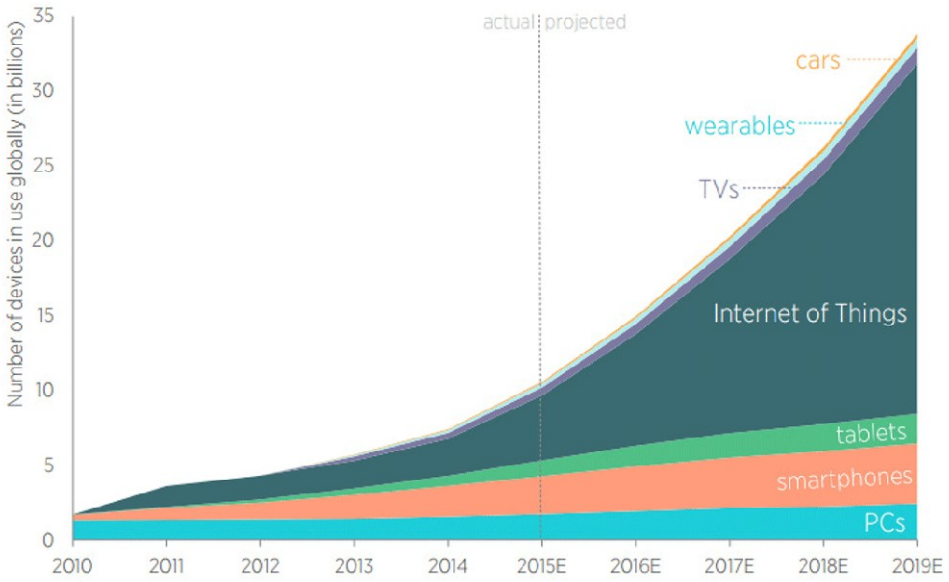
UCDAVIS

Outline

- Purpose/motivation
- Application profiles
- Network cost parameters
- Mathematical formulation
- Flow scenarios
- Simulation Results
- Slicing/reslicing
- Conclusion

Purpose/Motivation

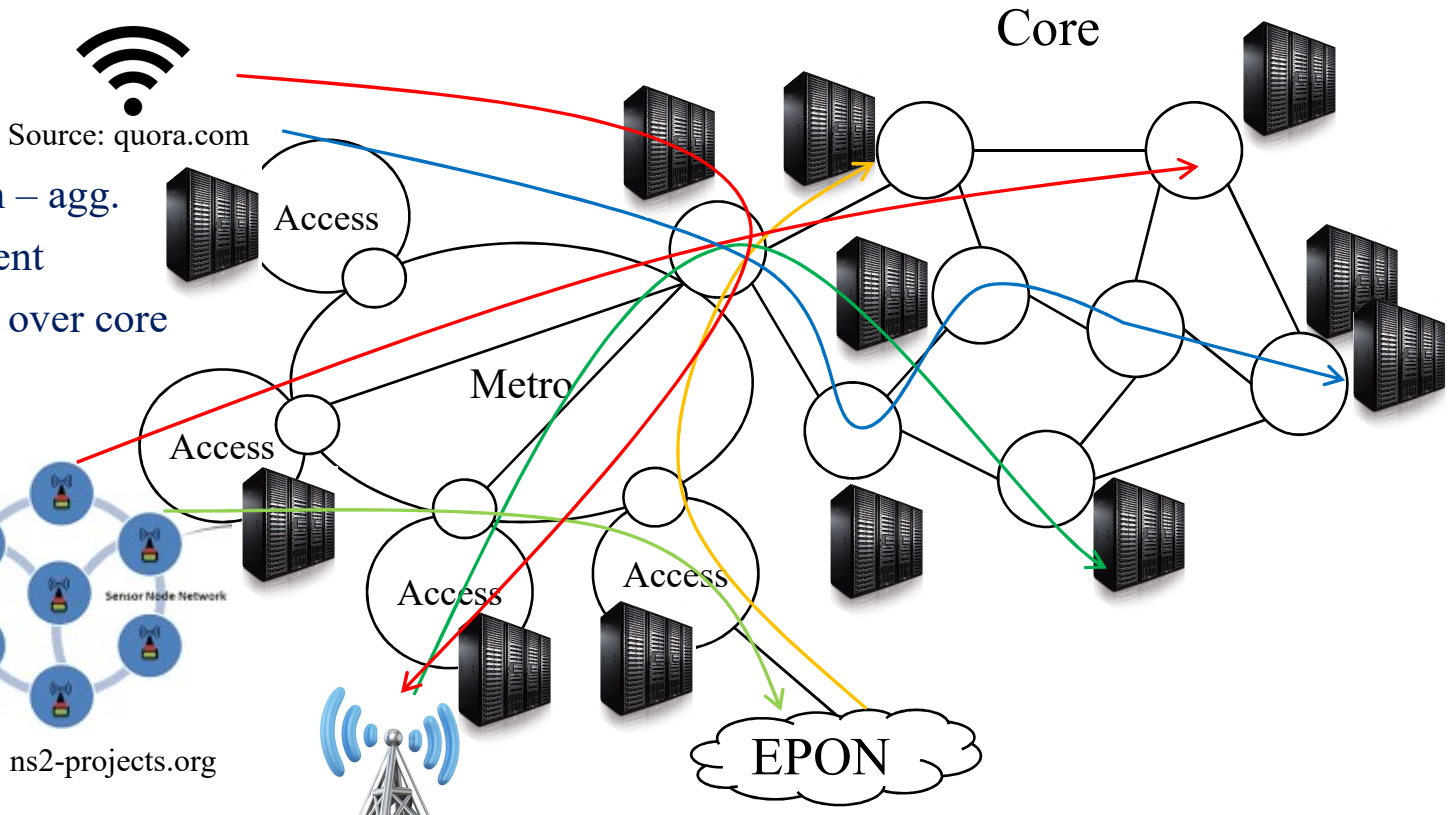
- Rapid growth in Internet of Things (IoT)/Machine to Machine (M2M) traffic
- Lack of quantitative application characterization
- What is the impact to metro/core networks, resources?
- Must define a new resource provisioning approach that adapts to traffic properties to maintain performance while minimizing costs



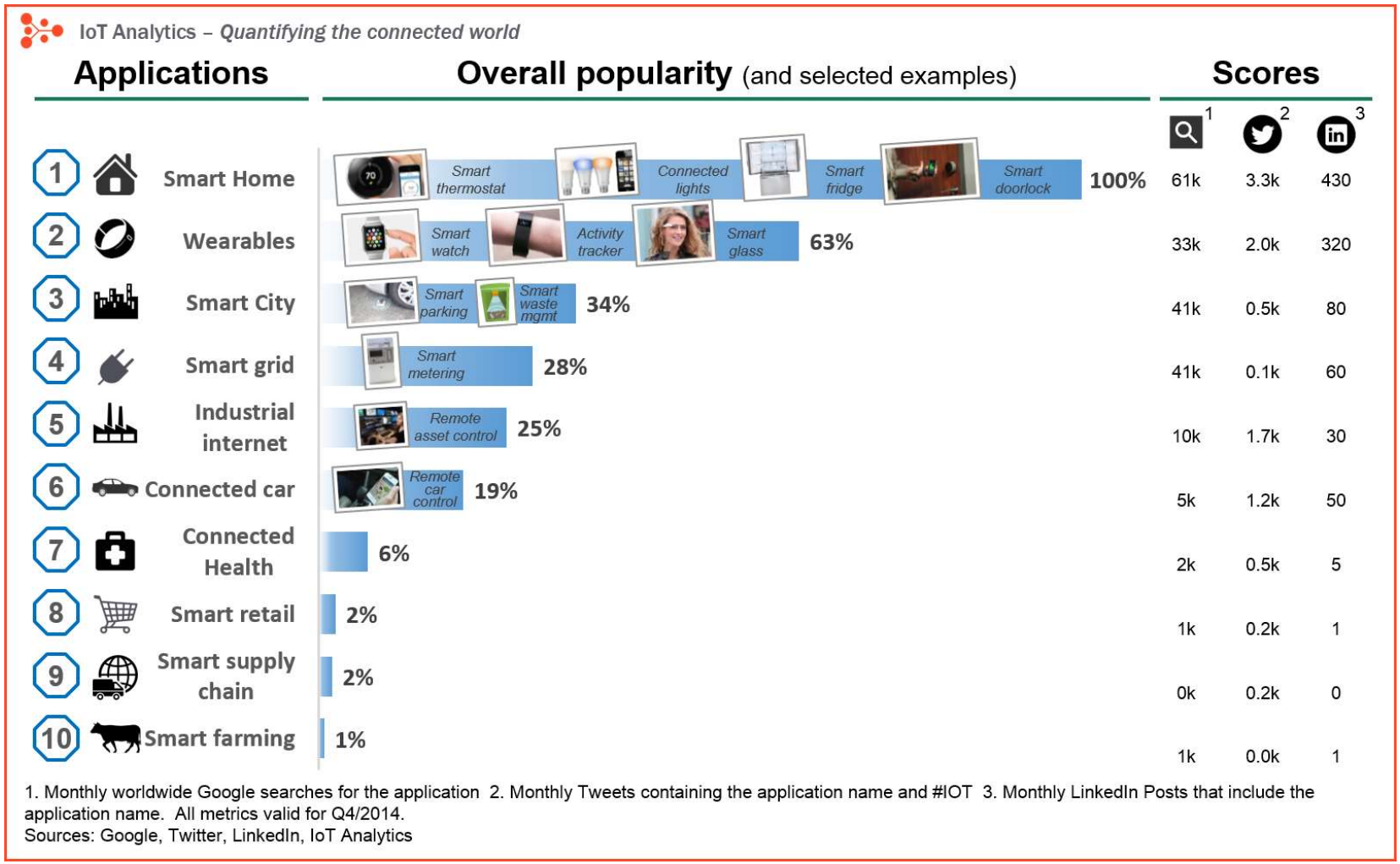
Source: John Greenough, "The Internet of Everything 2015," *Business Insider Intelligence*. Produced by Adam Thierer and Andrea Castillo, Mercatus Center at George Mason University, 2015.

Purpose/Motivation (cont.)

- Metro – Physical path – agg.
- Must become intelligent
- Prevent excess traffic over core
- Satisfy latency reqs



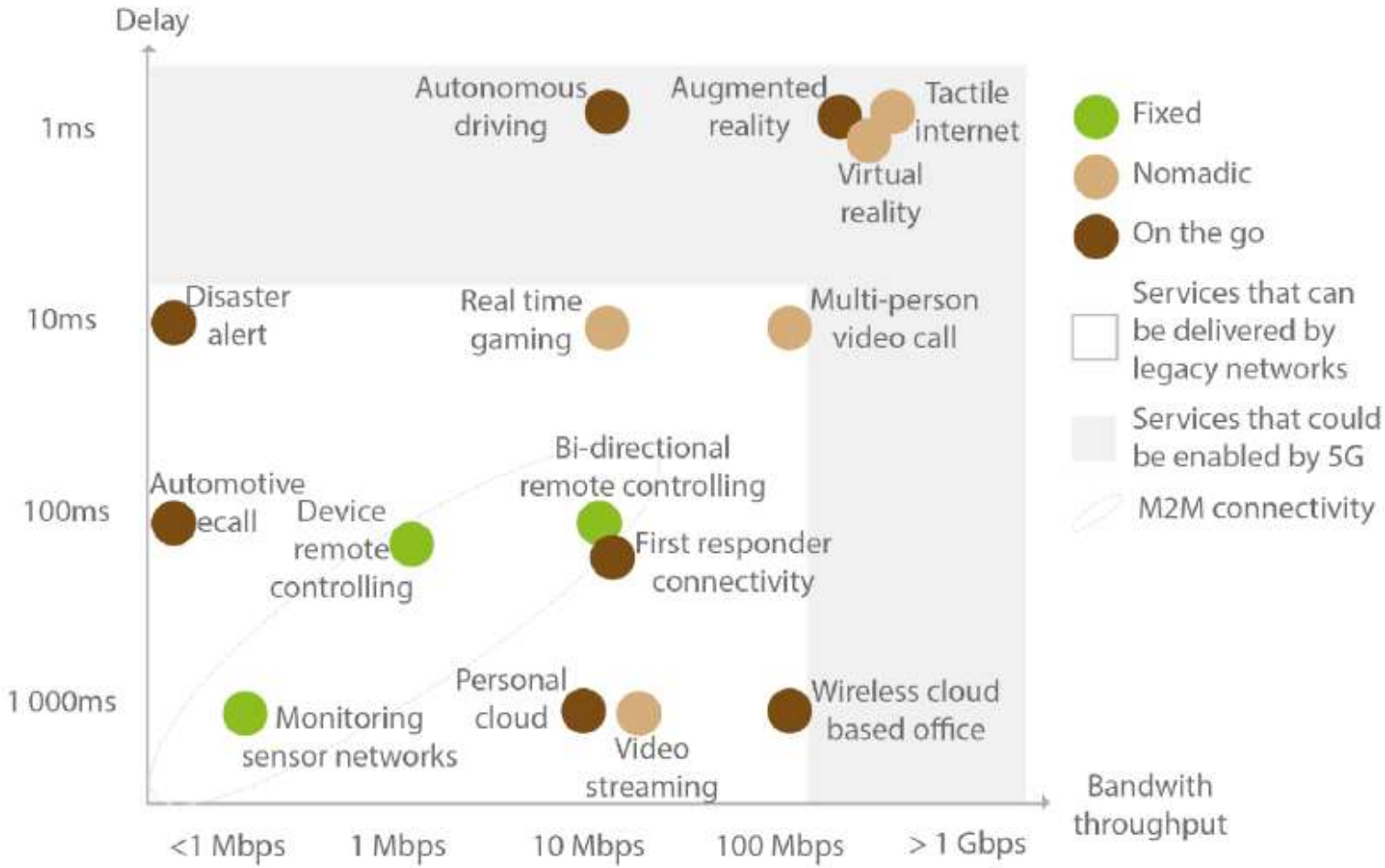
IoT/M2M Application Popularity



UC DAVIS
UNIVERSITY OF CALIFORNIA

Source: Google, Twitter, IoT Analytics, 2014.

Application Characteristics



UCDAVIS
UNIVERSITY OF CALIFORNIA

Source: GSMA Intelligence, 2015

Application Profile

- Each application profile contains a unique combination of parameters:
- Θ: Uni-directional latency budget from source to destination
- κ: bandwidth
- α: Computational complexity per unit of traffic
- β: Ratio of processed to raw data at processing node
- Λ: Minimum storage time

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

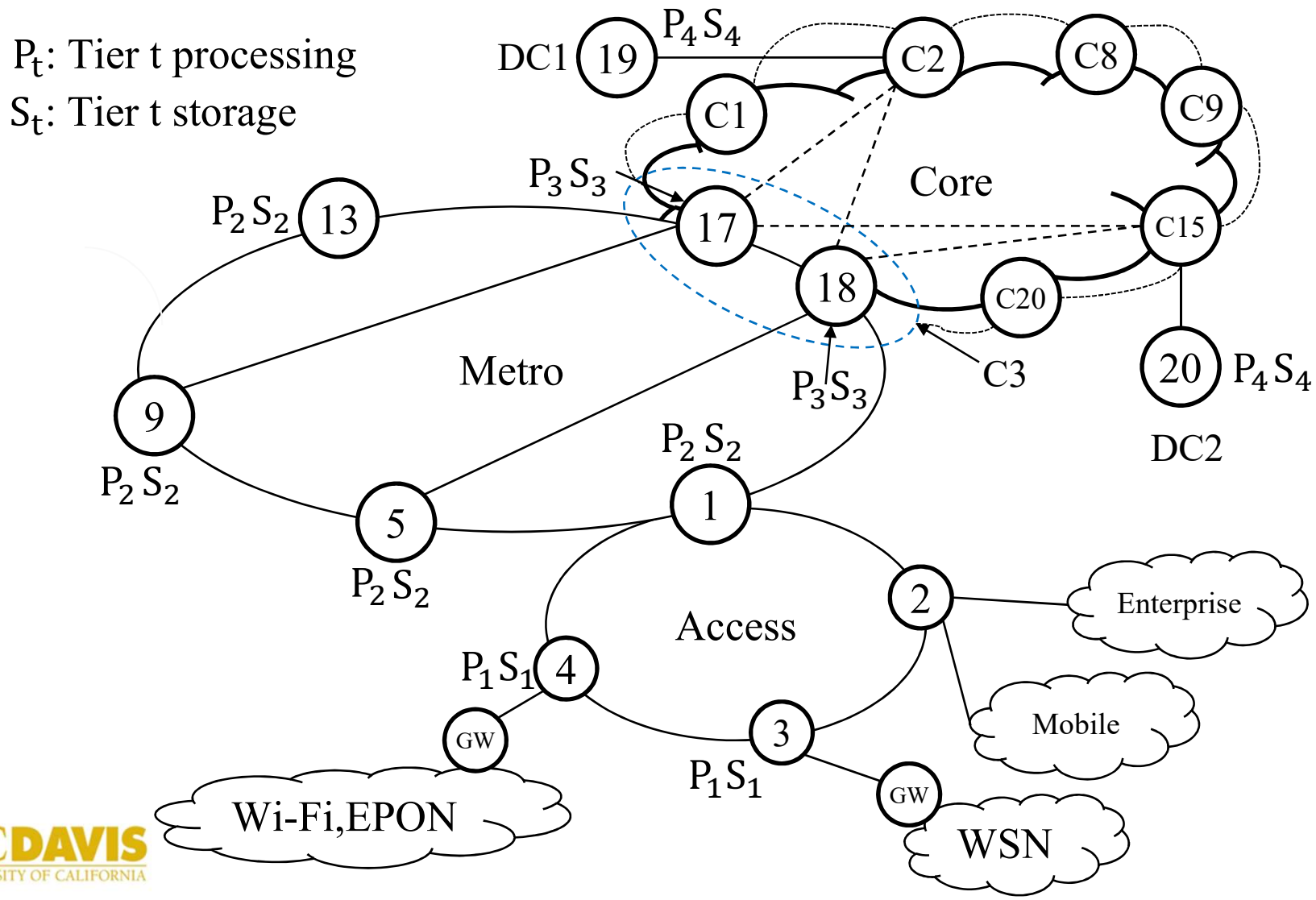
A_p: Processing only; A_s: Storage only; A_{sp}: P & S; A_n: Pt. to Pt.

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	0.003	0.2	0.1
7 – Geothermal Event	1000	1	0.02	0.3	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.

Hybrid Fog-Cloud Architecture



Topological Cost Properties

- Each CO has per unit costs of compute, storage, metro and core bandwidth
  - μ: compute power
  - ν: storage volume
  - Λ: metro bandwidth
  - ε_{up}, ε_{down} : core bandwidth
  - τ: Processing time constant (normalized to DC)

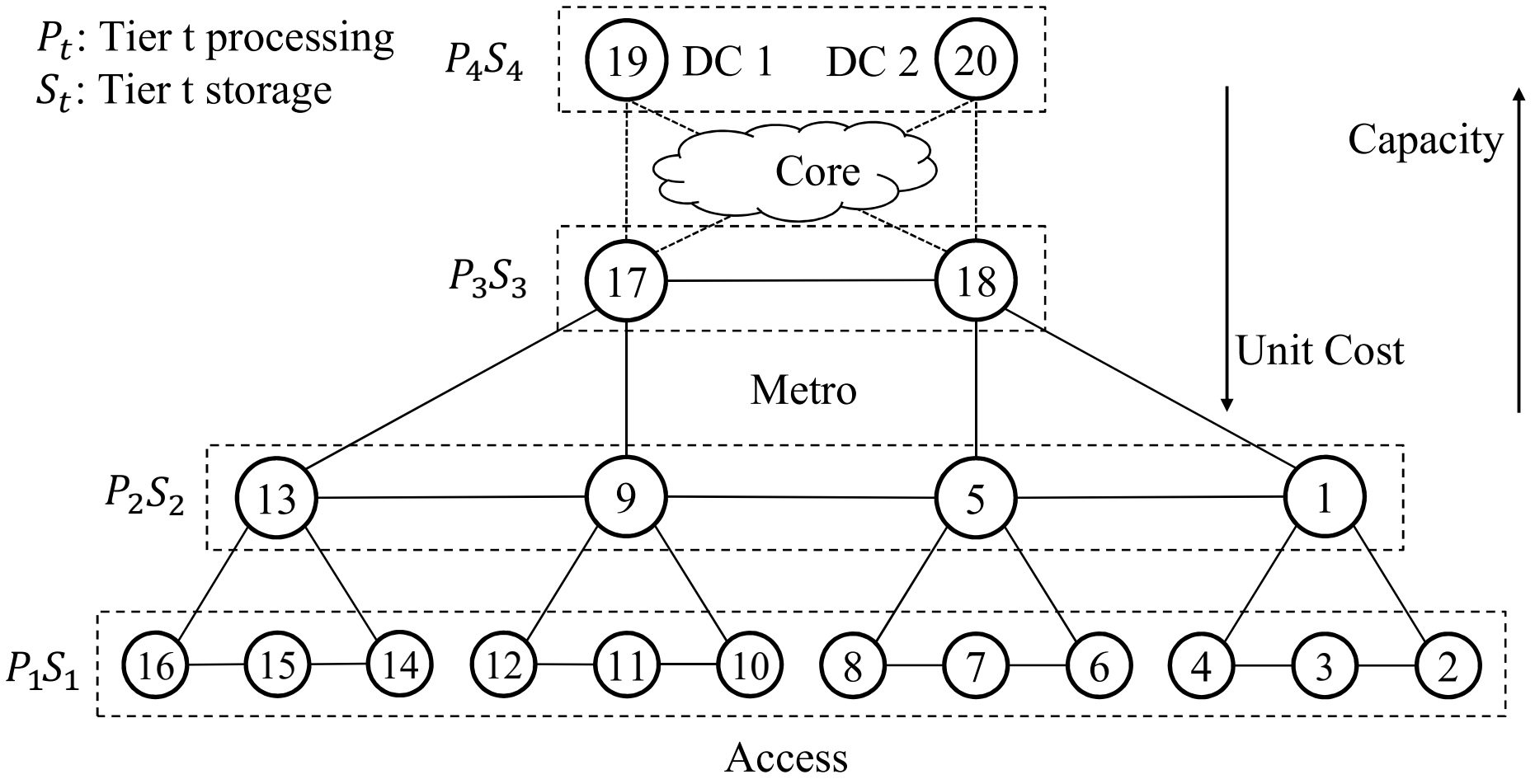
Tier	μ ($/CPU/Mo)	ν ($/GB/Mo)	Λ ($/Mbps/Mo)	$\varepsilon_{up}, \varepsilon_{down}$ ($/Mbps/Mo)	τ (/CPU)
1 – Access CO	90	0.0042	~1	~1/1	1.2
2 – Metro CO	70	0.004	~1	~1/1	1.15
3 – Core CO	50	0.0035	~1	~1/1	1.1
4 - DC	25	0.0025	~1	~1/1	1

https://cloud.google.com/compute/pricing

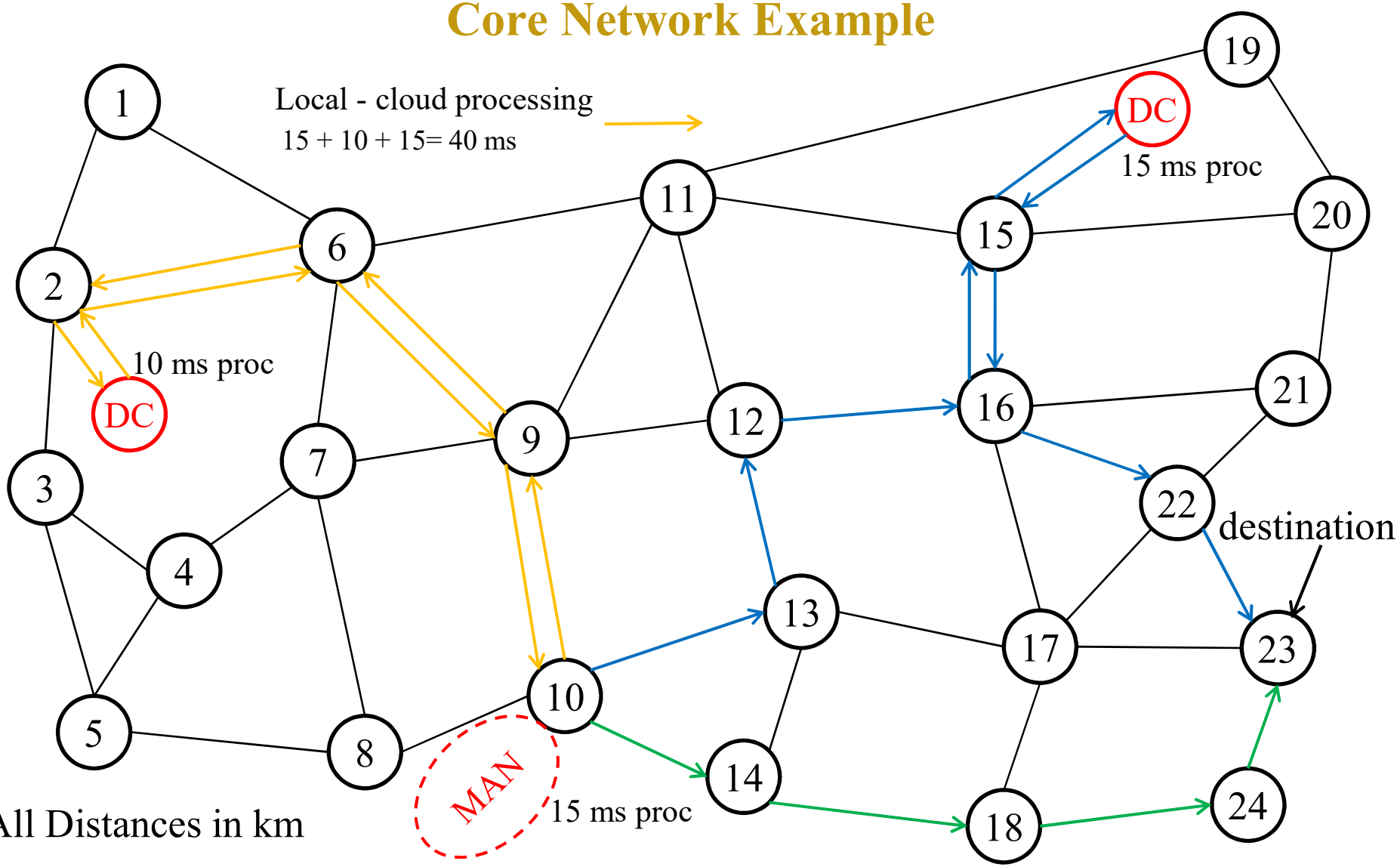
https://cloud.google.com/storage/pricing#pricing-example-simple

http://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php#

Hybrid Fog-Cloud Architecture – Hierarchical



Core Network Example



UC DAVIS
UNIVERSITY OF CALIFORNIA

Problem Explanation

- Inputs:
  - Offered traffic between s,d pairs – by application
  - Application Profiles:
    - Θ, κ, α, β, Δ
  - Hybrid Fog-Cloud Architecture: G(N,L)
  - Core Network SLAs: Residual latency by application & destination - $\theta_{a,f}$
- Objective function: Minimize total resource provisioning cost
  - Processing, Storage, core capacity up/down, metro capacity
- Constraints: Compute, storage capacity, latency
- Outputs: For all node pairs by application (or application alone):
  - Slice consisting of:
    - Path(s) with capacity (including core)
    - Required compute and storage resources at each node
    - Total required link, processing and storage capacities

Mathematical Formulation

Objective Function (costs):

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

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

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

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

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

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

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

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

$$\sum_{a \in \mathcal{A}_n} \sum_{s \in \mathcal{N}_g} \sum_{m \in \mathcal{N}_c} \sum_{f \in \mathcal{F}_c} v_{a,f}^{s,m} \Big]$$

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

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

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

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

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

Metro BW

Mathematical Formulation (cont.)

Constraints:

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

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

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

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

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

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

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

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

Solenoidality

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

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

Latency – Pt to Pt, Storage

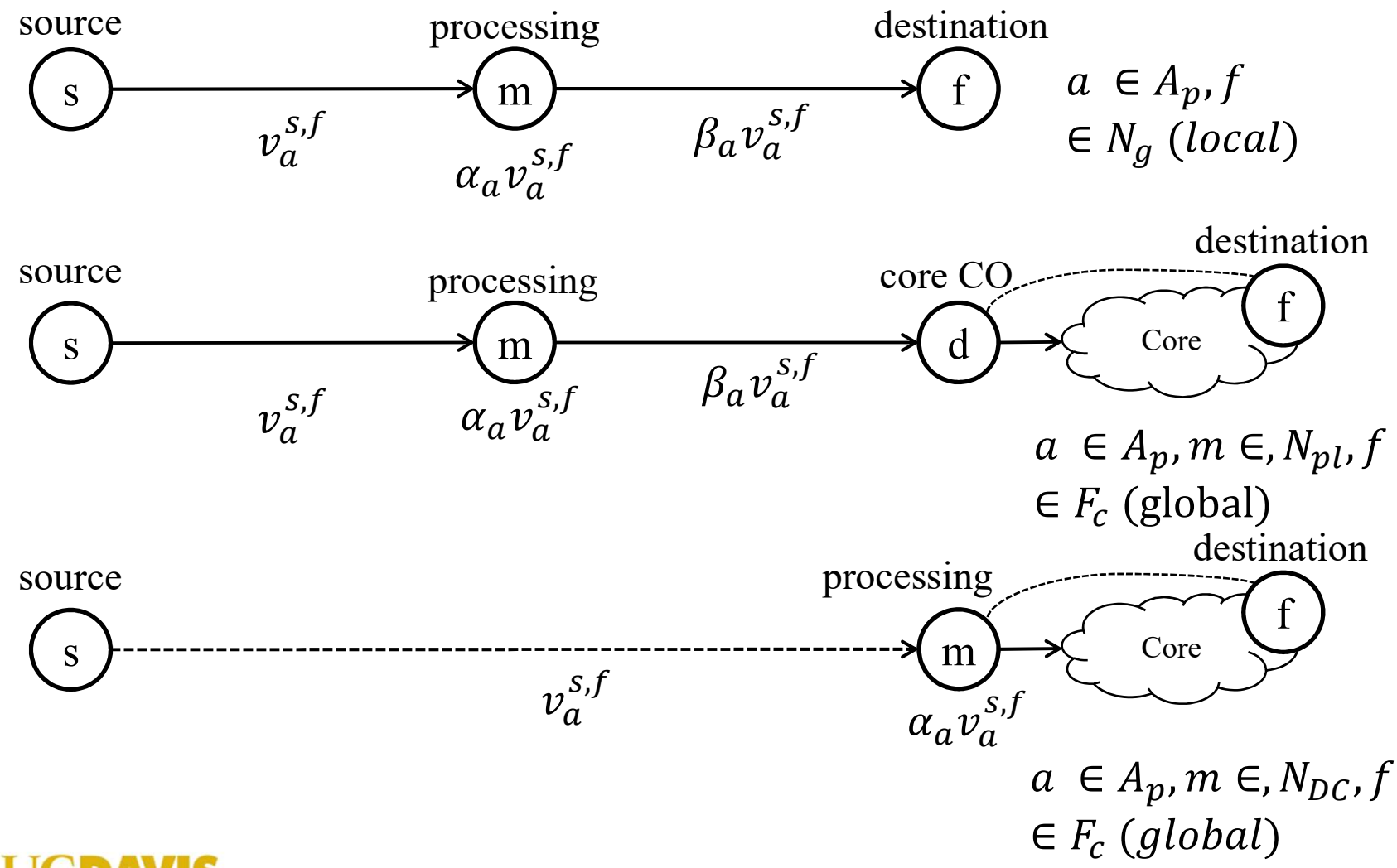
$$\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_k^{s,m} + \gamma_{a,m} + \sum_k r'^{s,m}_{a,k,f} D_k^{m,f} \leq \theta_{a,f,m}, \forall (a \in \mathcal{A}_p \cup \mathcal{A}_{sp}, s \in \mathcal{N}_g, m \in \mathcal{N}_p, f \in \mathcal{N}_g \cup \mathcal{N}_s)$$

Latency – Local Proc, Proc/Storage

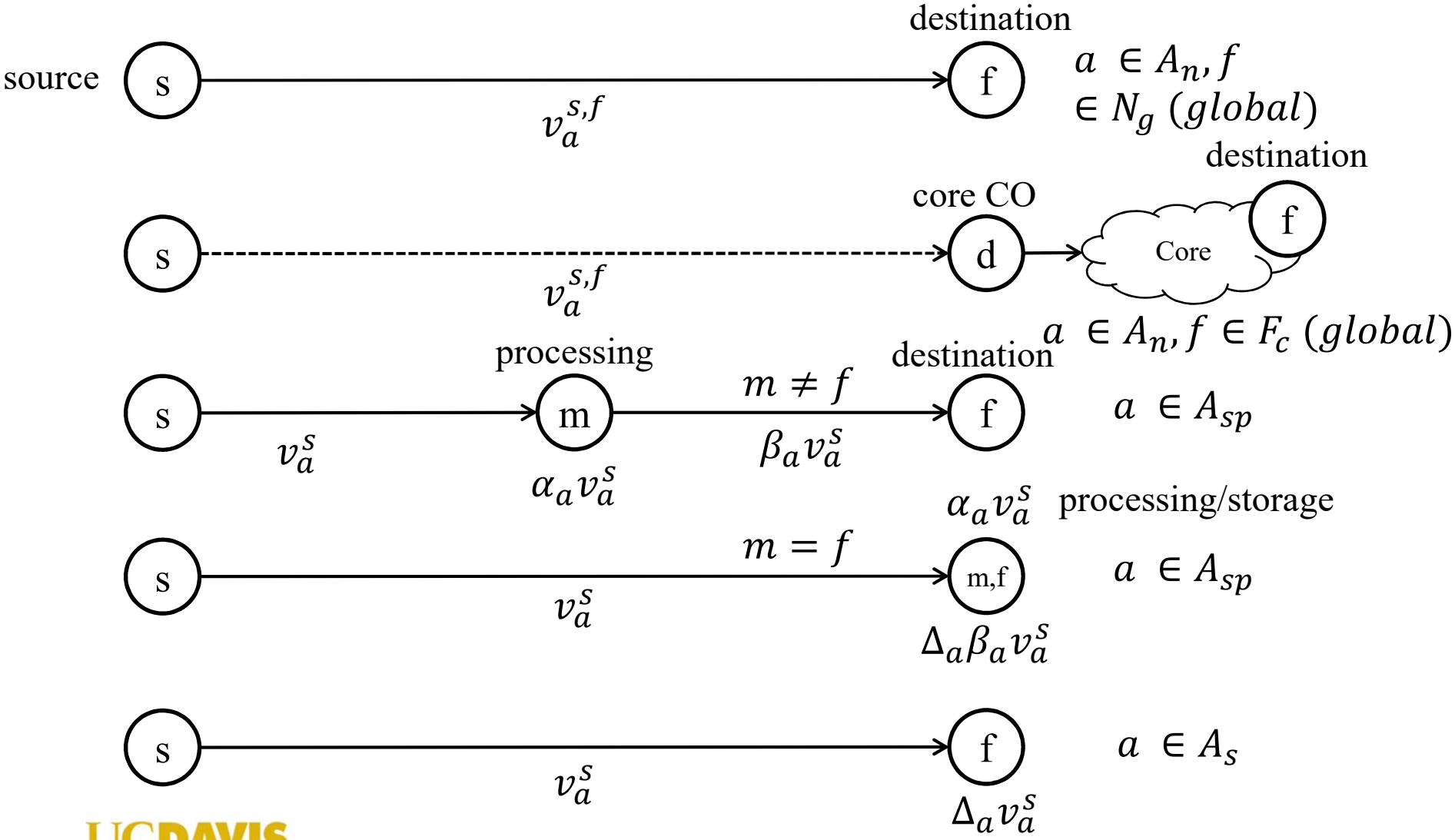
$$\sum_{k \in \mathcal{R}_{i,j}} r_{a,k,f}^{s,m} D_k^{s,m} + \gamma_{a,m} + \sum_k r''^{s,m,d}_{a,k,f} D_k^{m,d} \leq \theta_{a,f}, \forall (a \in \mathcal{A}_p, s \in \mathcal{N}_g, m \in \mathcal{N}_p, d \in \mathcal{N}_c, f \in \mathcal{F}_c)$$

Latency – Global Destination

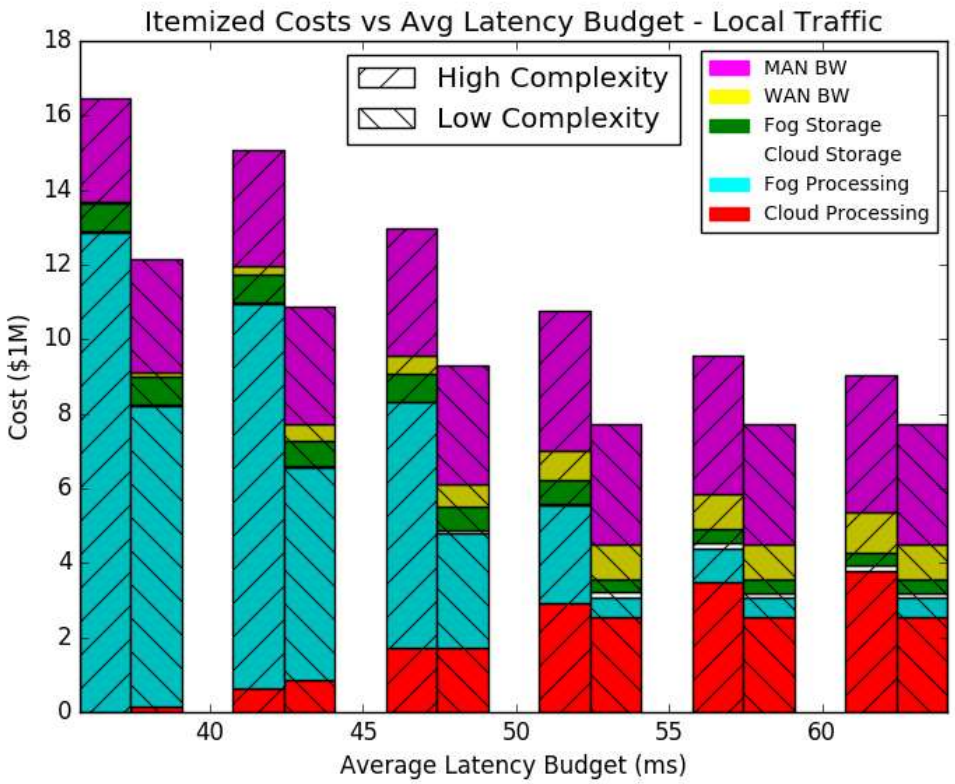
Functional Scenarios



Functional Scenarios (cont.)



Simulation Setup and Results



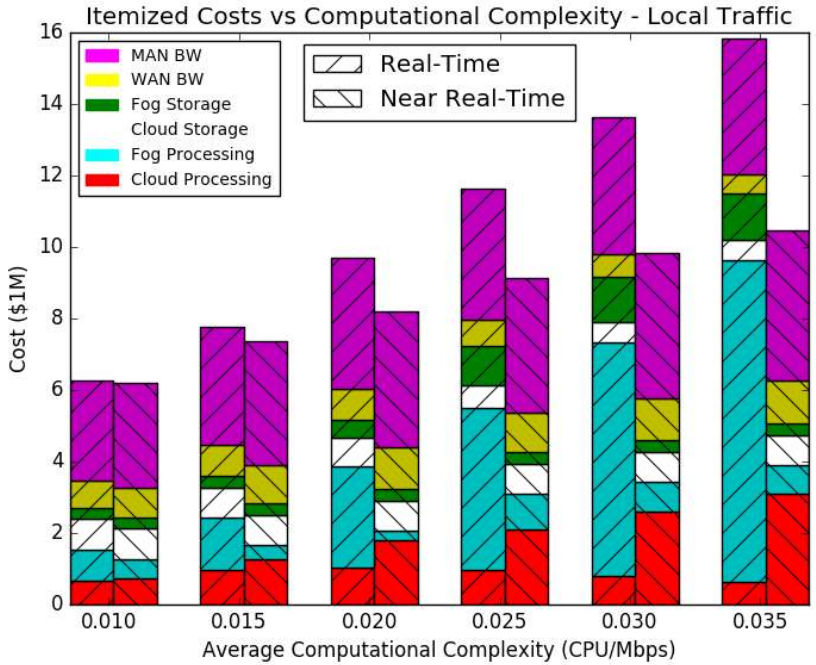
- Traffic Volume – 5 Tbps
- Complexity: .005 - .03 - CPU/Mbps
- Compression factor: 0.1 – 1
- Latency: 10 – 100 ms
- Real-time ~ 10-50 ms
- Near real-time ~ 50-100 ms
- Compute cost: 25, 50, 70, 90 $/CPU/Mo
- Storage cost: 2.50, 5, 7, 9 $/TB/Mo

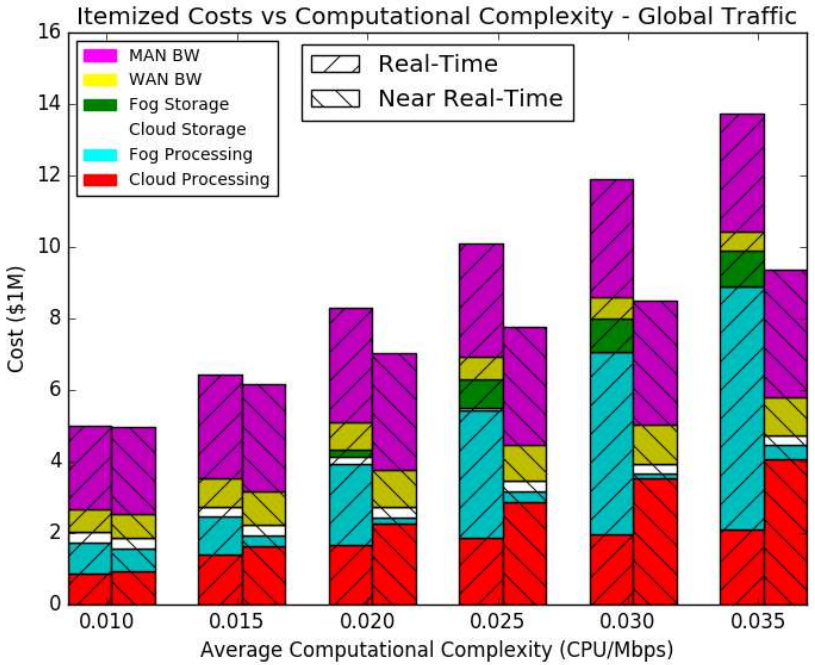
At low latency budgets, high complexity applications are slightly more restricted to fog processing due to higher processing delays.

As latency constraints are relaxed, high complexity applications can leverage inexpensive compute costs in the cloud, even though there is an insignificant increase in WAN bandwidth costs as a result.

UCDAVIS
UNIVERSITY OF CALIFORNIA

Simulation Results (cont.)





-Locally destined, real-time and near real-time traffic: fog processing costs increase with computational complexity as processing delay becomes a larger proportion of total latency budget

-Cloud processing costs of real-time traffic start to decrease at .02 while near real-time cloud processing costs continue to increase with complexity

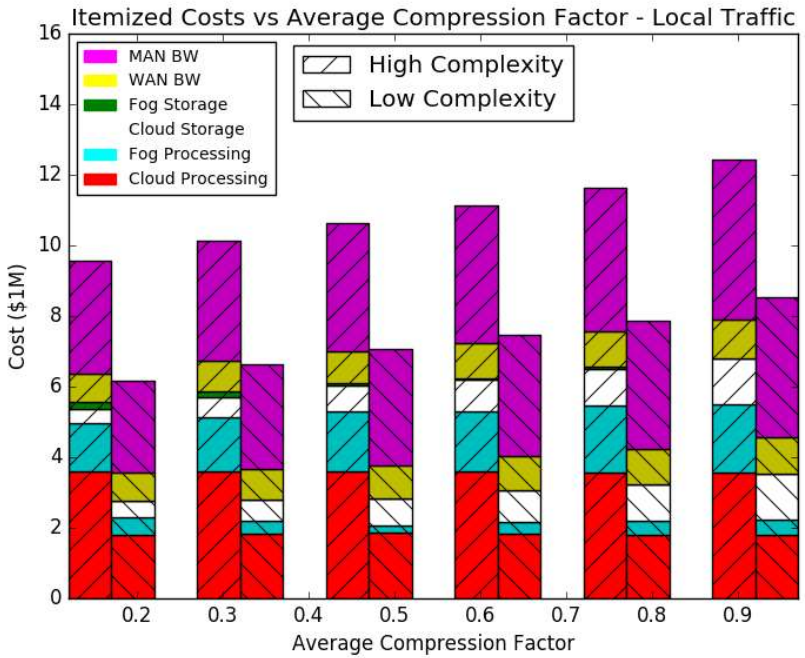
-Global traffic: higher cloud processing costs than local

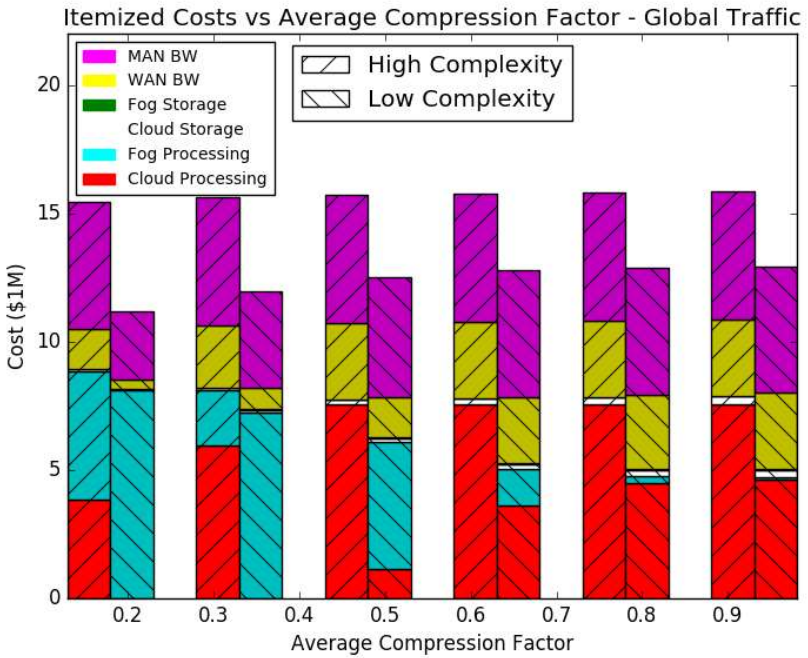
-Fog/cloud processing ratio increases with real-time traffic as processing delay consumes higher proportion of latency budget

-Cloud processing costs increase at much slower rate with increasing complexity for real-time traffic as DC compute locations restrict more applications to fog processing

UCDAVIS
UNIVERSITY OF CALIFORNIA

Simulation Results (cont.)





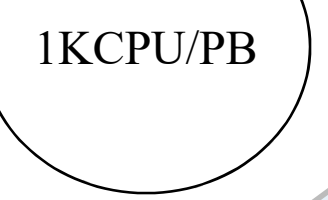
-Metro bandwidth cost increases as it is less expensive than moving fog processing to lower (more expensive) tiers
-WAN bandwidth and cloud storage increases with compression factor due to decoupling of processing and storage functions

-Metro bandwidth cost increases as it is less expensive than to move fog processing to lower tiers
- Increases in cloud processing (lower total processing costs) are offset by larger WAN bandwidth costs with increasing compression factor (minimizing total costs)

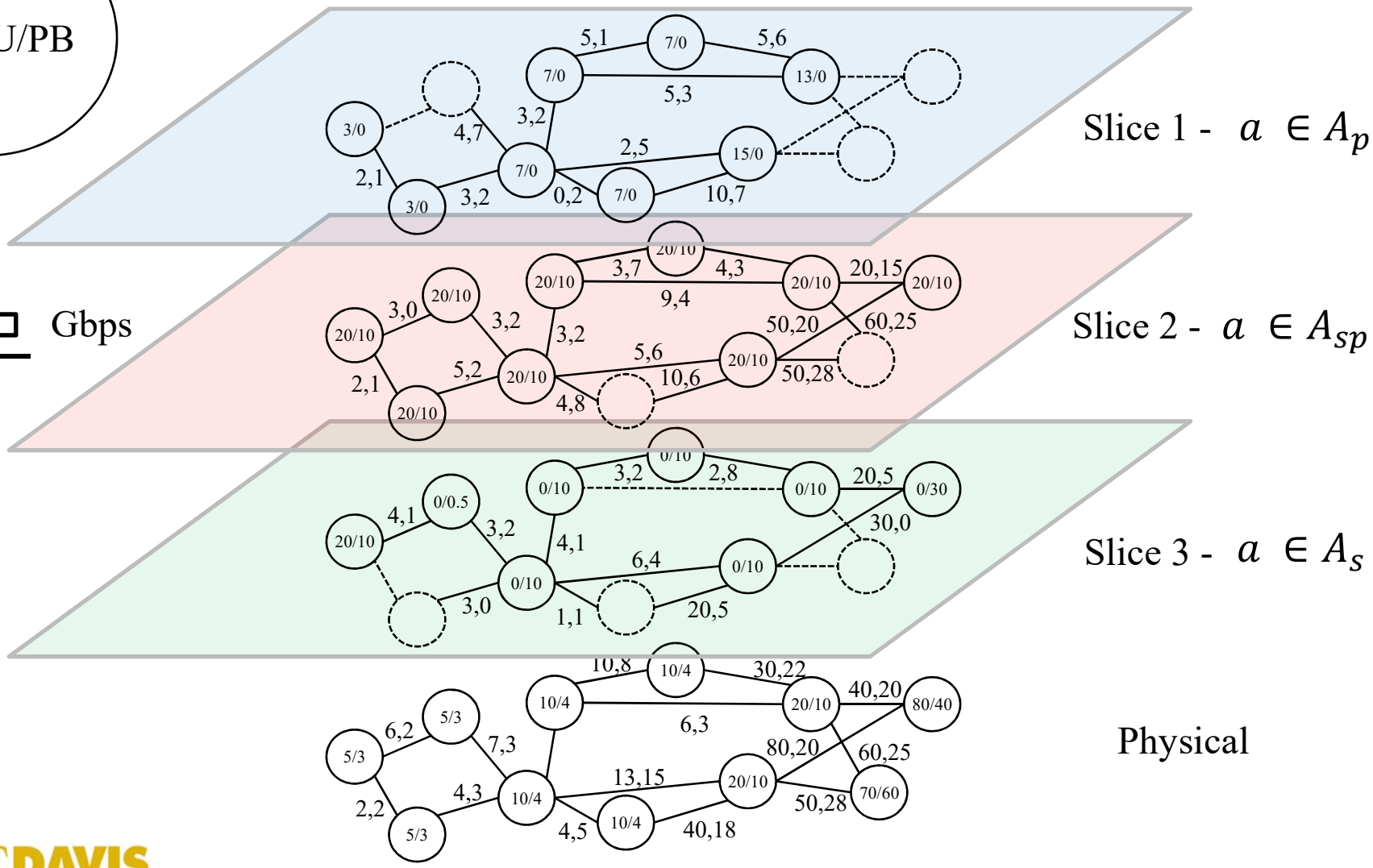
UCDAVIS
UNIVERSITY OF CALIFORNIA

Slice Per Application

1KCPU/PB



Gbps



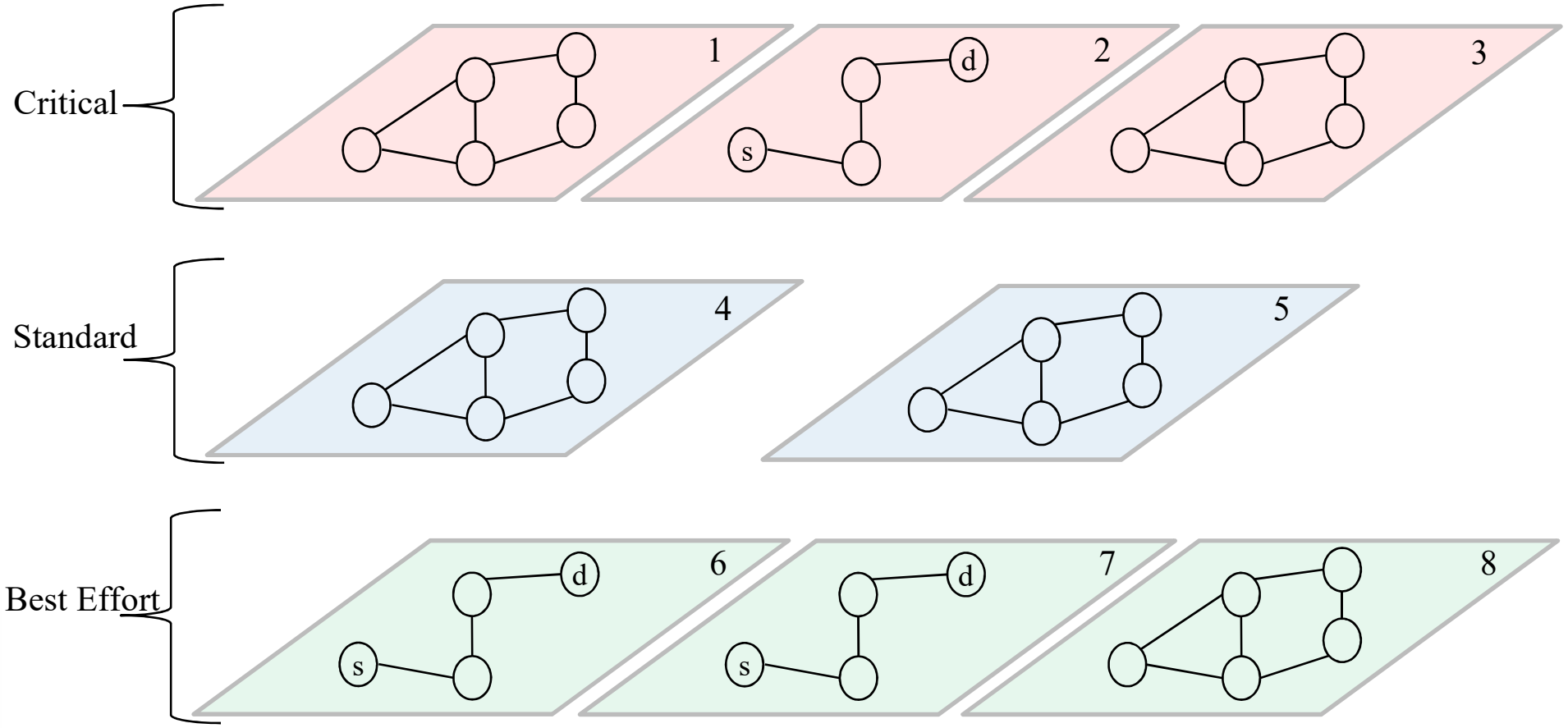
Slice Priority

- 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

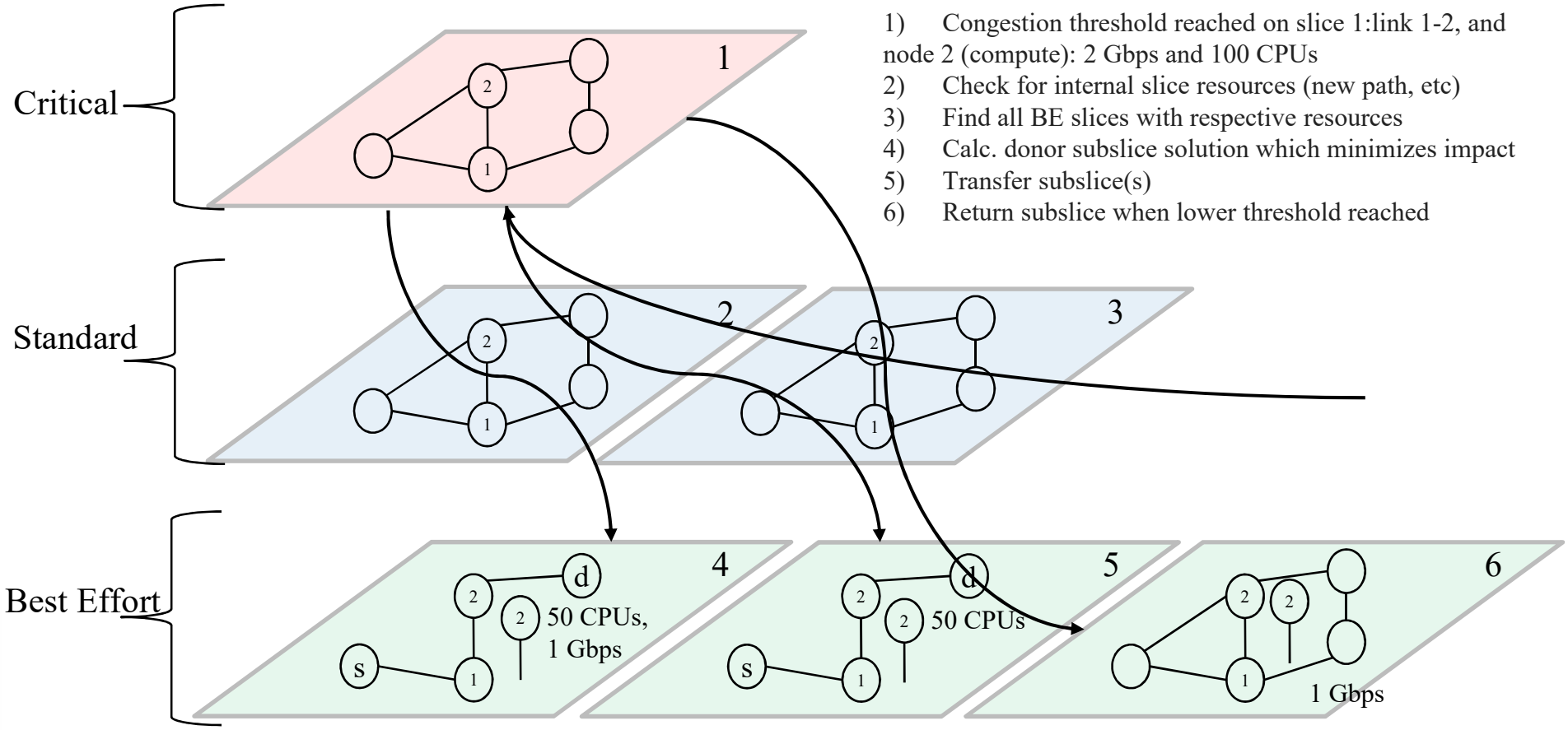
Nakao, A., Du, P., Kiriha, Y., Granelli, F., Gebremariam, A.A., Taleb, T. and Bagaa, M. End-to-End Network Slicing for 5G Mobile Networks. *Journal of Information Processing*, *25*, pp.153-163, 2017.

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

Conclusion

- Motivation: Lack of quantitative analysis of how specific application traffic affects resource provisioning: future (IoT/M2M) traffic
- Proposed a parameterized application profile: $A = A_p \cup A_s \cup A_{sp} \cup A_n$
  - Θ, α, β, κ, Λ
- 4-tier hybrid fog-cloud architecture
  - Increasing capacity/decreasing unit costs
- Flow scenarios: how profile parameters affect compute, storage, and link capacity
- Simulation Results: Θ, α, β
- Network Slicing
  - Granularity
  - Priority Slicing/Reslicing
- Future Work: model dynamic re-slicing algorithm and generate simulation results
  - Model tradeoffs between total traffic performance and higher priority application performance at multiple slicing granularities

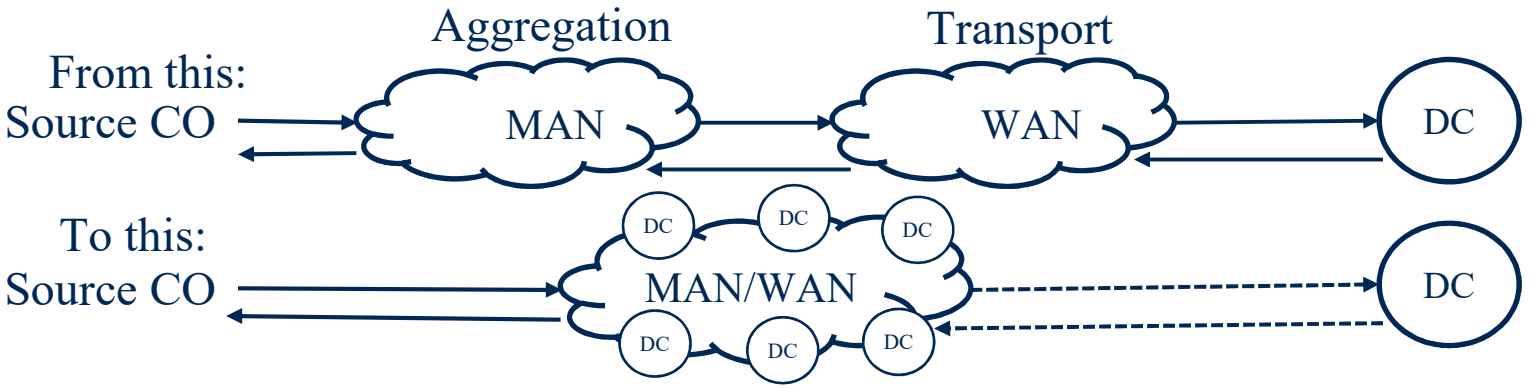
Questions

Purpose/Motivation (cont.)

- Lack of quantitative modeling of application profile: how can traffic be parameterized? What effects will parameters have on network resources and associated costs?

- Goal is to model these effects in a hybrid fog-cloud architecture and show how proper network slicing can ensure satisfactory performance at minimal cost via variable granularities and reslicing

- How do we determine the optimal slice configuration?



Provisioning Solution by Application

Mathematical Formulation

Inputs:

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

A_s : Set of all application profiles requiring storage only

A_p : Set of all application profiles requiring processing only

A_{sp} : Set of all application profiles requiring processing and storage

A_n : Set of all application profiles requiring neither processing nor storage

$\theta_{a,f,m}$: Residual latency budget of traffic destined for core node f of application profile a,processed at node m

$v_a^{s,f}$: Offered traffic of application profile a, node pair (s, f), $a \in A_n \cup A_p$

v_a^s: Offered traffic of application profile a, sourced at node s, $a \in A_s \cup A_{sp}$

Mathematical Formulation (cont.)

Inputs:

N_c: set of nodes directly attached to core network
N_g: set of nodes that generate traffic (sources of data)g
N_{DC}: set of data center nodes
$N = N_c \cup N_g \cup N_{DC}$
$N_p = N_g \cup N_{DC}$: Set of all nodes capable of processing
$N_s = N_g \cup N_{DC}$: Set of all nodes capable of storage
$N_l = N_g$ Set of local nodes capable of processing (excluding DC)

F_c : Set of distant core nodes

$\eta_{s,d}$: set of all admissible paths between node pair (s, d)
$P_k^{s,m}$: Total prop delay on the k^{th} adm path between node pair (s, m)
$T_k^{s,m}$: Total trans delay on the k^{th} adm path between node pair (s, m)
C_m : processing capacity of node m, in CPUs
S_f: storage capacity of node f, in GB

Mathematical Formulation (cont.)

Variables:

$A_p : x_{a,f}^{s,m} = 1$ if traffic of app profile a, destined for node f, generated at source node s, is processed at node m

$A_s : x_{a,f}^{s,m} = 1$ if traffic of app profile a, generated at node s, is stored at node f, $m = f$

$A_{sp} : x_{a,f}^{s,m} = 1$ if traffic of app profile a, generated at source node s, is processed at node m and stored at node f

$A_p, A_{sp} : r_{a,k,f}^{s,m} = 1$ if traffic of application profile **a** is routed over the k^{th} admissible path between node pair (s, m), destined for node f

$A_s : r_{a,k,f}^{s,m} = 1$ if traffic of application profile **a** is routed over the k^{th} admissible path between node pair (s, m), $m = f$

$A_n : r_{a,k,f}^{s,m} = 1$ if traffic of application profile a is routed over the k^{th} admissible path between node pair (s, m), internal: $m = f$,external: $m \in N_c$

Mathematical Formulation (cont.)

Variables:

$A_p, A_{sp}, A_s : r^{s,m}_{a,k,f} = 1$ if traffic of application profile a is routed over the k^{th} admissible path between node pair (s, m), $m \in N_p, A_s : m \in N_s$

$A_p, A_{sp} : r'^{s,m}_{a,k,f} = 1$ if traffic of application profile a is routed over the k^{th} admissible path between node pair (m, f), $m \in N_p, f \in N_g, A_{sp} : f \in N_s$

$A_p, A_{sp} : r''^{s,m,d}_{a,k,f} = 1$ if traffic of application profile a is routed over the k^{th} admissible path between node pair (m, d),destined for core node f, $m \in N_p, f \in F_c, d \in N_c$

Mathematical Formulation (cont.)

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

$$sec = \left(\frac{CPU}{Mbps}\right)(Mb)\left(\frac{1}{CPU}\right)$$

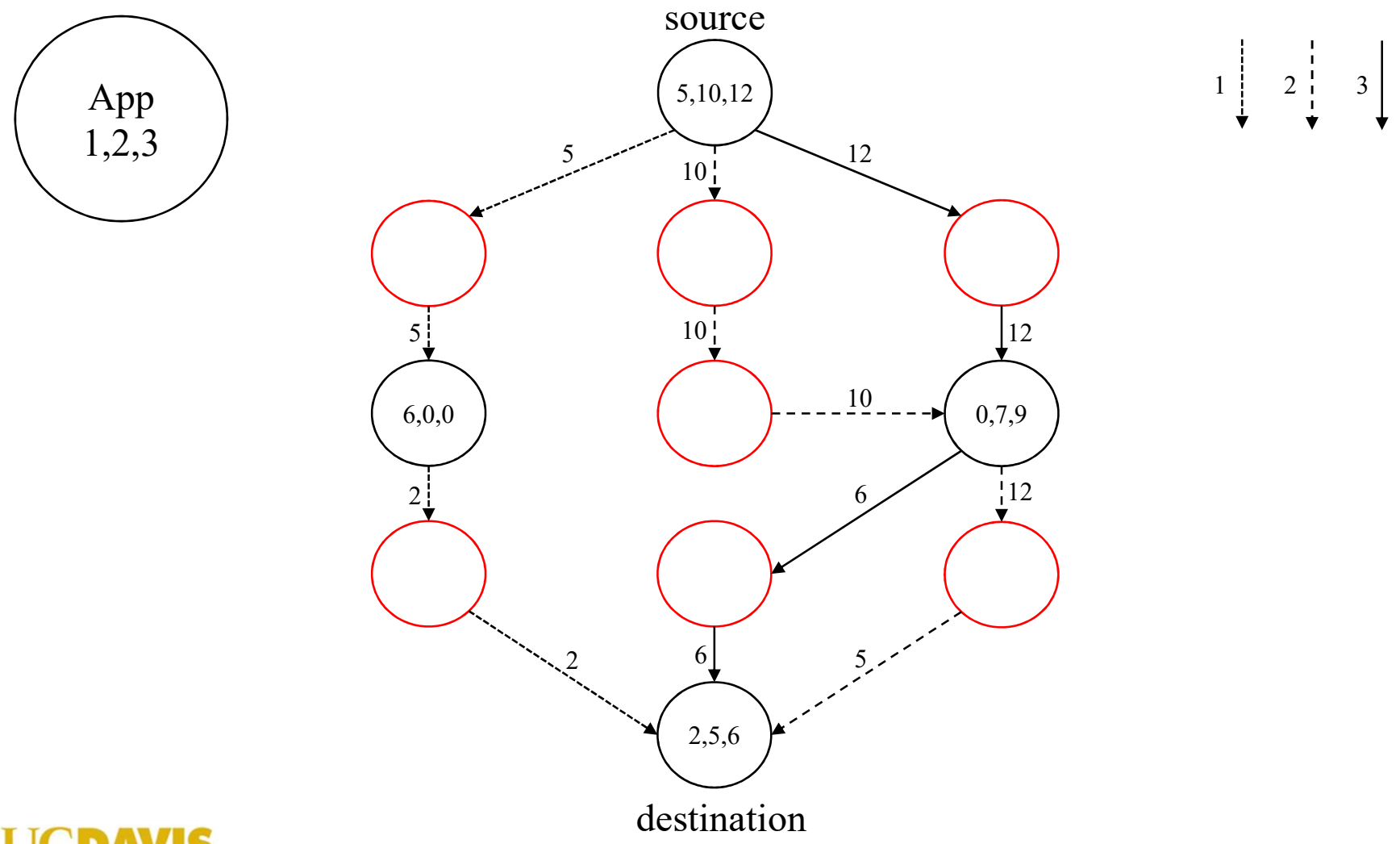
Processing Capacity: $C_m = \alpha_a v_{a,f}^{s,m}$

$$CPU = \left(\frac{CPU}{Mbps}\right)(Mbps)$$

Storage Capacity: $S_f = \Delta_a v_{a,f}^{s,m}, \beta_a$ if necessary

$$GB = (sec)\left(\frac{Mbit}{sec}\right)\left(\frac{GB}{Mbit}\right)$$

Slice Per Application, Node Pair



Online Heuristic Flow Chart