Optimal Control of Wireless Computing Networks

Hao Feng, *Student Member, IEEE,* Jaime Llorca, *Member, IEEE,* Antonia M. Tulino, *Fellow, IEEE,* and Andreas F. Molisch, *Fellow, IEEE*

***Abstract*—Augmented information (AgI) services allow users to consume information that results from the execution of a chain of service functions that process source information to create real-time augmented value. Applications include real-time analysis of remote sensing data, real-time computer vision, personalized video streaming, and augmented reality, among others. We consider the problem of optimal distribution of AgI services over a wireless computing network, in which nodes are equipped with both communication and computing resources. We characterize the wireless computing network capacity region and design a joint flow scheduling and resource allocation algorithm that stabilizes the underlying queuing system while achieving a network cost arbitrarily close to the minimum, with a tradeoff in network delay. Our solution captures the unique chaining and flow scaling aspects of AgI services, while exploiting the use of the *broadcast approach* coding scheme over the wireless channel.**

***Index Terms*—Wireless computing network, edge computing, service function placement, service chaining, resource allocation, dynamic control, distributed algorithm, broadcast approach coding scheme.**

I. Introduction

Internet traffic will soon be dominated by the consumption of what we refer to as *augmented information (AgI) services*. Unlike traditional information services, in which users consume information that is produced or stored at a given source and is delivered via a communication network, AgI services provide end users with information that results from the real-time *processing* of source information via possibly multiple service functions that can be hosted anywhere in the network. Examples include real-time analysis of remote sensing data, real-time computer vision, personalized video streaming, and augmented reality, among others.

While today's AgI services are mostly implemented in the form of software functions instantiated over general purpose servers at centralized cloud data centers [2], the increasingly low latency requirements of next generation real-time AgI services is driving cloud resources closer to the end users in the form of small cloud nodes at the edge of the network, resulting in what is referred to as a *distributed cloud network*. This naturally raises the question of where to execute each service function and how to route network flows through the appropriate sequence of service functions, a question that is impacted both by the computation and the communication resources of the cloud network infrastructure. Given fixed service rates, linear programming formulations and heuristic procedures for joint function placement and routing were developed in [3]–[6], while polynomial-time algorithms with approximation guarantees were developed in [7]–[9]. The works of [3], [8] are based on a multi-commodity-chain (MCC) flow model that generalizes traditional multi-commodity flow to account for *flow chaining* via service processing and joint communication/computation resource allocation.

The study of *dynamic* control policies that adjust the configuration of service function chains in response to unknown changes in service demands was initiated by the present authors in [10], [11]. By extending the MCC model of [3], [8] to dynamic cloud network settings, these works provided the first characterization of a cloud network capacity region and the design of distributed throughput-optimal control policies that jointly schedule communication and computation resources while pushing overall network cost arbitrarily close to minimum.

A key aspect not considered in all previous works is the increasingly important role of the wireless access network for efficient service delivery. AgI services are increasingly sourced and accessed from wireless devices, and with the advent of mobile and fog computing [12], service functions can also be hosted at wireless computing nodes (*i.e.,* computing devices with wireless networking capabilities) such as mobile handsets, connected vehicles, compute-enabled access points or cloudlets [13]. When introducing the wireless network into the computing infrastructure, the often unpredictable nature of the wireless channel further complicates flow scheduling, routing, and resource allocation decisions. In the context of traditional wireless communication networks, the Lyapunov drift plus penalty (LDP) control methodology (see [14] and references therein) has been shown to be a promising approach to tackle these intricate stochastic network optimization problems. Ref. [15] extends the LDP approach to multi-hop, multi-commodity wireless ad-hoc networks, leading to the Diversity Backpressure (DIVBAR) algorithm. DIVBAR exploits the broadcast nature of the wireless medium without precise channel state information (CSI) at the transmitter, and it is shown to be throughput-optimal under the assumption that at most one packet can be transmitted in each transmission attempt, and that no advanced coding scheme is used. Ref.

Manuscript received on October 17, 2017; accepted on September 25, 2018; approved by IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS Editor Kai Zeng. This work was supported in part by the National Science Foundation, division of Network Technology and Systems, under Grant 1619129 and in part by the National Science Foundation, division of Computing and Communication Foundations, under Grant 1423140. We acknowledge Zheda Li and Pan Tang for constructive discussions and suggestions on the simulation settings. This paper was presented in part at 2017 IEEE ICC [1]. (Corresponding author: Hao Feng.)

H. Feng and A. F. Molisch are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089-2565 USA (e-mail: haofeng@usc.edu; molisch@usc.edu).

J. Llorca is with Nokia Bell Labs, Holmdel, NJ 07733 USA (e-mail: jaime.llorca@nokia-bell-labs.com).

A. M. Tulino is with Nokia Bell Labs, Holmdel, NJ 07733 USA, and also with the University degli Studi di Napoli Federico II, 80138 Naples, Italy (e-mail: a.tulino@nokia-bell-labs.com; antoniamaria.tulino@unina.it).

[16] extends DIVBAR by incorporating rateless coding in the transmissions of a single packet, yielding enhanced throughput performance.

Motivated by the important role of wireless networks in the delivery of AgI services, in this paper, we address the problem of distribution of AgI services over a multi-hop *wireless computing network*, which is composed of nodes with communication and computing capabilities.

Our contributions can be summarized as follows:

1) We extend the MCC flow model of [3], [8], [11] to the delivery of AgI services over wireless computing networks, taking into account the routing diversity created by the inherent broadcast nature of the wireless channel. In the wireless MCC model, the queue backlog of a given commodity builds up from receiving the commodity information units via broadcast transmissions from neighbor nodes, as well as from the generation of commodity information units via local service function processing.
2) We incorporate the use of broadcast approach coding scheme [17], [18] into the scheduling of AgI service flows over wireless computing networks in order to exploit routing diversity and enhance transmission efficiency. By applying the broadcast approach coding scheme at a transmitter with no precise channel state information (CSI), the source information is encoded into superposition layers. Then, multiple receivers can decode different amount of layers according to their channel states after overhearing the transmitted signal, enabling opportunities and challenges for enhancing routing diversity in wireless computing networks.
3) For a given set of AgI services, we characterize the capacity region of a wireless computing network in terms of the set of exogenous input rates that can be processed through the required service functions and delivered to the required destinations. Unlike the capacity region of a traditional communications network, which only depends on the network topology, the capacity region of a wireless computing network also depends on the AgI service structure, and it is shown to be enlarged via the use of the broadcast approach coding scheme, as opposed to the traditional single-layer (outage approach) coding scheme.
4) We design a dynamic wireless computing network control (DWCNC) algorithm that makes local transmission, processing, and resource allocation decisions without knowledge of service demands or their statistics. The local transmission scheduling takes the broadcast approach coding scheme into consideration to explore routing diversity. DWCNC is throughput optimal and allows pushing total resource cost arbitrarily close to minimum with a tradeoff in network delay.

The remainder of the paper is organized as follows: Section II presents the system model. Section III characterizes the network capacity region of a wireless computing network. Section IV constructs the DWCNC algorithm, and Section V proves the optimal performance of DWCNC. The paper is concluded in Section VI.

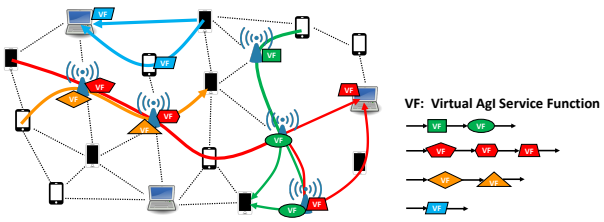


Fig. 1. Illustration of the delivery of AgI services over a wireless computing network.

II. System Model

Before describing the mathematical model in detail, we first introduce the main concepts around the delivery of AgI services over wireless computing networks via an illustrative example depicted in Fig. 1. As shown in the figure, a wireless computing network is composed of wireless nodes, e.g., user equipments and access points, equipped with both communication and computation resources. AgI services, on the other hand, are described by a chain of service functions that determines the sequence in which source information flows must be processed in order to generate the final information flows that get consumed by the end users. For example, a multimedia streaming service may be described by a sequence of functions such as video/audio mixing, media transcoding, and media acceleration [19]. Fig. 1 shows a wireless computing network offering four generic AgI services, each described by a chain of virtual functions (VFs). Each service is requested by a client, defined by a source-destination pair. Source flows are routed through the network to wireless computing nodes where they can be processed via the appropriate service functions, before being delivered to the corresponding destinations. Note that flows can be split into multiple paths to exploit routing diversity, and get processed by multiple instances of the same service function at different locations. As we will show in the following, the proposed DWCNC algorithm uses the broadcast approach coding scheme to optimally exploit routing diversity without precise CSI knowledge at the transmitters.

A. Network Model

We consider a wireless computing network composed of $N = |\mathcal{N}|$ nodes representing distributed computing devices that communicate over wireless links labeled according to node pairs (i, j) for $i, j \in \mathcal{N}$. Node $i \in \mathcal{N}$ is equipped with K_i^{tr} transmission resource units (*e.g.,* transmission power) that it can use to transmit information over the wireless channel. In addition, node i is equipped with K_i^{pr} processing resource units (*e.g.,* central processing units or CPUs) that it can use to process information as part of an AgI service (see Sec. II-B).

Time is slotted with slots normalized to integer units $t \in \{0, 1, 2, \dots\}$. We use the binary variable $y_{i,k}^{\text{tr}}(t) \in \{0, 1\}$ to indicate the allocation or activation of $k \in \mathcal{K}_i^{\text{tr}} \triangleq \{0, \dots, K_i^{\text{tr}}\}$ transmission resource units at node i at time t, which incurs $w_{i,k}^{\text{tr}}$ cost units. Analogously, $y_{i,k}^{\text{pr}}(t) \in \{0, 1\}$ indicates the allocation of $k \in \mathcal{K}_i^{\text{pr}} \triangleq \{0, \dots, K_i^{\text{pr}}\}$ processing resource units at node i at time t, which incurs $w_{i,k}^{\text{pr}}$ cost units. Notice that the binary resource allocation variables $y_{i,k}^{\text{tr}}(t)$, $y_{i,k}^{\text{pr}}(t)$ must satisfy $\sum_{k=0}^{K_i^{\text{tr}}} y_{i,k}^{\text{tr}}(t) \leq 1$, $\sum_{k=0}^{K_i^{\text{pr}}} y_{i,k}^{\text{pr}}(t) \leq 1$.

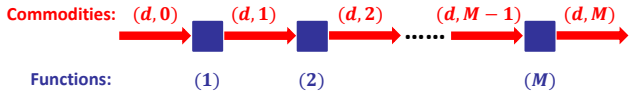


Fig. 2. Illustration of an AgI service chain for destination $d \in \mathcal{D}$. There are M functions and $M+1$ commodities. The AgI service takes the source commodity $(d,0)$ and delivers the final commodity (d,M) after going through the sequence of functions $\{1,2,\ldots,M\}$. Function m takes commodity $(d,m-1)$ and generates commodity (d,m).

B. Augmented Information Service Model

While the analysis in this paper readily applies to an arbitrary number of services, for ease of exposition, we focus on the distribution of single *augmented information service*, described by a chain of functions $\mathcal{M} = \{1,2,\ldots,M\}$. A service request is described by a source-destination pair $(s,d) \in \mathcal{N} \times \mathcal{N}$, indicating the request for source flows originating at node s to go through the sequence of functions $\mathcal{M}$ before exiting the network at destination node d. We adopt a MCC flow model, in which commodity $(d,m) \in \mathcal{D} \times \{\mathcal{M},0\}$ identifies the information units generated by function $m \in \mathcal{M}$ for destination $d \in \mathcal{D} \subseteq \mathcal{N}$, where $|\mathcal{D}| \triangleq D$. We assume information units have arbitrary fine granularity (*e.g.,* packets or bits). Commodity $(d,0)$ denotes the source commodity for destination d, which identifies the information units arriving exogenously at each source node s that have node d as their final destination. (see Fig. 2).

Each service function has (possibly) different processing requirements. We denote by $r^{(m)}$ the *processing complexity factor* of function m, which indicates the number of operations required by function m to process one input information unit. Another key aspect of AgI services is the fact that information flows can change size as they go through service functions. Let $\xi^{(m)} > 0$ denote the *scaling factor* of function m. Then the size of the function's output flow is $\xi^{(m)}$ times as large as its input flow.

C. Computing Model

As is shown in Fig. 3, we represent the processing capabilities of wireless computing nodes via a processing element (*e.g.,* CPU in a cloudlet node) co-located with each network node. A static, dedicated *computing channel* is considered, where the achievable processing rate at node i with the allocation of k processing resource units is given by $R_{i,k}$ in operations per timeslot. We use $\mu_{i,\mathrm{pr}}^{(d,m)}(t)$ to denote the flow rate (in information units per timeslot) of commodity (d,m) $(0 \le m < M)$ from node i to its processing unit at time t, and $\mu_{\mathrm{pr},i}^{(d,m)}(t)$ to denote the flow rate of commodity (d,m) $(0 < m \le M)$ from the processing unit back to node i (see Fig. 3). We then have the following MCC and maximum *processing rate* constraints:

$$\mu_{\mathrm{pr},i}^{(d,m)}(t) = \xi^{(m)} \mu_{i,\mathrm{pr}}^{(d,m-1)}(t), \qquad \forall i,d,m{>}0,t, \tag{1}$$

$$\sum_{(d,m>0)} \mu_{i,\mathrm{pr}}^{(d,m-1)}(t)\; r^{(m)} \le \sum_{k=0}^{K_i^{\mathrm{pr}}} R_{i,k}\; y_{i,k}^{\mathrm{pr}}(t), \quad \forall i,t. \tag{2}$$

Note that function m at node i processes input commodity $(d,m-1)$ at a rate $\mu_{i,\mathrm{pr}}^{(d,m-1)}(t)$ information units per timeslot, using $\mu_{i,\mathrm{pr}}^{(d,m-1)}(t) r^{(m)}$ operations per timeslot, and generates output commodity (d,m) at a rate $\mu_{\mathrm{pr},i}^{(d,m)}(t)$ information units per timeslot.

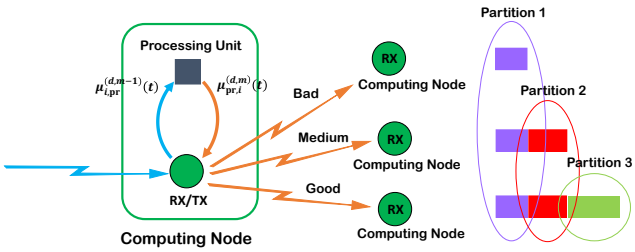


Fig. 3. Illustration of a wireless computing node equipped with computation and communication resources. Commodity $(d,m-1)$ can be processed via the processing unit hosting function m to generate commodity (d,m). The use of the broadcast approach at the TX leverages multi-receiver diversity. The information decoded by the RX with the "bad" channel is a subset of the information decoded by the RX with the "medium" channel, which is further a subset of the information decoded by the RX with the "good" channel. The transmitted information can therefore be grouped into three partitions.

D. Wireless Transmission Model

We assume that multiple transmitters (TXs) may transmit simultaneously to overlapping receivers (RXs) via the use of orthogonal broadcast channels of fixed bandwidth, a priori allocated by a given policy, whose design is outside the scope of this paper. On the other hand, due to the broadcast nature of the wireless medium, multiple RXs may overhear the transmission of a given TX. We model the channel between node i and all other nodes in the network as a physically degraded Gaussian broadcast channel, where the *network state process* (the vector of all channel gains), denoted by $\mathbf{S}(t) \triangleq \{s_{ij}(t), \forall i,j \in \mathcal{N}\}$, evolves according to a Markov process with state space $\mathcal{S}$ and whose steady-state probabilities exist. We assume that the statistical CSI is known at the TX, while the instantaneous CSI can only be learned after the transmission has taken place and is thereby outdated (delayed).

It is well-known that superposition coding is optimal (capacity achieving) for the physically degraded broadcast channel with independent messages [21]. In particular, in this work we adopt the *broadcast approach* coding scheme (see [17], [18] and references therein), which consists of sending incremental information using superposition layers, such that the number of decoded layers at any RX depends on its own channel state, and the information decoded by a given RX is a subset of the information decoded by any other RX with no worse channel gain. That is, for a given transmitting node i, if we sort the $N-1$ potential RX nodes in non-decreasing order of their channel gains $\{q_{i,1},\ldots,q_{i,N-1}\}$, such that $q_{i,n}$ with $n \in \{1,\ldots,N-1\}$ denotes the receiver with the n-th lowest channel gain, then the information decoded by receiver $q_{i,n}$ is also decoded by receiver $q_{i,u}$, for $u > n$. Moreover, let $\Omega_{i,n} \triangleq \{q_{i,n},\cdots,q_{i,N-1}\}$ be the set of receivers with the $N-n$ highest channel gains. Then, we can partition the information transmitted by node i during a given timeslot into $N-1$ disjoint groups, with the n-th partition being the information whose successful receiver set is exactly $\Omega_{i,n}$, *i.e.,* the information that is decoded by the nodes in $\Omega_{i,n}$, but not

by the nodes in $\mathcal{N}\backslash\{i\}\backslash\{\Omega_{i,n}\}$. Fig. 3 illustrates the use of the broadcast approach for multi-receiver diversity.

Let $p_{i,k}(a)$ denote the optimal power density function over the continuum of superposition layers resulting from the allocation of k transmission resource units at node i. Then, based on the broadcast approach [18], when allocating k transmission resource units, the maximum achievable rate over link (i,j) at time t is given by

$$R_{ij,k}(t) = B_W \int_0^{g_{ij}(t)} \frac{a p_{i,k}(a)}{1 + a\int_a^{\infty} p_{i,k}(s)ds} da, \qquad (3)$$

where $g_{ij}(t)$ is the channel gain over link (i,j) at time t, and B_W is the available bandwidth.

In practice, instead of a continuum of superposition layers, each node i is assumed to use a set of L_i discrete superposition code layers. In this case, we use $P_{i,k}^{\text{tot}}$ to denote the total power associated with the allocation of k transmission resource units at node i, and $P_{i,k}(l)$ to denote the power allocated to code layer l, with $\sum_{l=1}^{L_i} P_{i,k}(l) = P_{i,k}^{\text{tot}}$. In addition, the L_i code layers are respectively associated with L_i channel gain thresholds $\{\bar{g}_{i,1}, \cdots, \bar{g}_{i,L_i}\}$, $\bar{g}_{i,1} \leq \cdots \leq \bar{g}_{i,L_i}$.[1] Then, the maximum achievable rate over link (i,j) at time t, $R_{ij,k}(t)$, can take L_i+1 possible values, given by the sum of the rates associated with the code layers whose channel gain thresholds are no higher than the channel realization at time t:

$$R_{ij,k}(t) = \begin{cases} \bar{R}_{i,k}^0, & \text{if } g_{ij}(t) < \bar{g}_{i,1}, \\ \bar{R}_{i,k}^l, & \text{if } \bar{g}_{i,l} \leq g_{ij}(t) < \bar{g}_{i,l+1},\ 1 \leq l < L_i - 1, \\ \bar{R}_{i,l}^{L_i}, & \text{if } g_{ij}(t) \geq \bar{g}_{i,L_i}, \end{cases}$$

where $\bar{R}_{i,k}^0 = 0$, and

$$\bar{R}_{i,k}^l = B_W \sum_{l' \leq l} \log\left(1 + \frac{P_{i,k}(l')\,\bar{g}_{i,l'}}{1 + \bar{g}_{i,l'} \sum_{l''>l'} P_{i,k}(l'')}\right), \quad \text{for } 1 \leq l \leq L_i. \qquad (4)$$

Correspondingly, the channel realization of link (i,j) at time t, $s_{ij}(t)$, has L_i+1 possible *channel states* $\{\bar{s}_{i,0}, \cdots, \bar{s}_{i,L_i}\}$:

$$s_{ij}(t) = \begin{cases} \bar{s}_{i,0}, & \text{if } g_{ij}(t) < \bar{g}_{i,1}; \\ \bar{s}_{i,l}, & \text{if } \bar{g}_{i,l} \leq g_{ij}(t) < \bar{g}_{i,l+1},\ 1 \leq l < L_i - 1; \\ \bar{s}_{i,L_i}, & \text{if } g_{ij}(t) \geq \bar{g}_{i,L_i}. \end{cases}$$

E. Communication Protocol

The communication protocol between each TX-RX pair is illustrated in Fig. 4. At the beginning of each timeslot, TX and RX exchange all necessary control signals, including queue backlog state information (see Sec. II-F). Then, the TX decides how many transmission resource units to allocate for the given timeslot and how much rate to allocate to each available commodity. Afterwards, the transmission starts and lasts for a fixed time period (within the timeslot); during that time, both data and pilot tones (whose overhead is neglected) are transmitted.

[1]We assume that the power allocated to each layer and the channel gain thresholds are given parameters, and leave their optimization to maximize the expected achievable rate out of the scope of this paper.

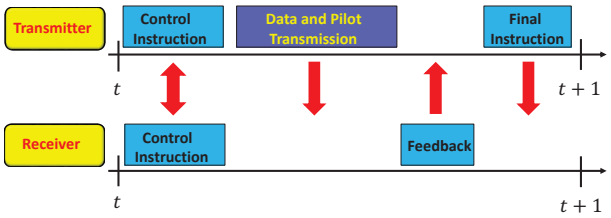


Fig. 4. Timing diagram of the communication protocol over a wireless link.

After the transmission ends, every potential RX provides immediate feedback, containing the identification of the information decoded by the RX, which allows the TX to derive the experienced channel states. The TX then makes a *forwarding decision* and sends it through a final instruction to all the RXs, instructing each RX which portion of its decoded information to keep for further processing and/or forwarding (hence assigning the processing/forwarding responsibility). Control information, feedbacks, and final instructions are sent through a stable control channel, whose overhead is neglected.

We use $\mu_{ij}^{(d,m)}(t)$ to denote the amount of information of commodity (d,m) retained by node j after the transmission from node i during timeslot t. In addition, it shall be useful to denote by $\mu_{iq_{i,u},n}^{(d,m)}(t)$ the information retained by node $q_{i,u}$ belonging to the n-th partition of node i's transmitted information. Then, since $q_{i,u} \in \Omega_{i,n}$ for all n satisfying $n \leq u$, we have

$$\mu_{iq_{i,u}}^{(d,m)}(t) = \sum\nolimits_{n=1}^{u} \mu_{iq_{i,u},n}^{(d,m)}(t), \quad \forall i,u,d,m,t. \qquad (5)$$

Moreover, given the allocation of k transmission resource units at time t, the maximum achievable rate for the n-th partition is $R_{iq_{i,n},k}(t) - R_{iq_{i,n-1},k}(t)$. Hence, we have

$$\sum_{(d,m)} \mu_{i,q_{i,u},n}^{(d,m)}(t) \leq \sum_{k=0}^{K_i^{\text{tr}}} \left[R_{iq_{i,n},k}(t) - R_{iq_{i,n-1},k}(t)\right] y_{i,k}^{\text{tr}}(t), \quad \forall i,t,u \geq n, \qquad (6)$$

where $R_{iq_{i,0},k}(t) = 0$, for all i,k,t. Note that Eqs. (5) and (6) lead to the following rate constraint on link (i,j) for all t: $\sum_{(d,m)} \mu_{ij}^{(d,m)}(t) \leq \sum_{k=0}^{K_i^{\text{tr}}} R_{ij,k}(t)\ y_{i,k}^{\text{tr}}(t)$.

F. Queuing Model

We denote by $a_i^{(d,m)}(t)$ the exogenous arrival rate of commodity (d,m) at node i at time t, and by $\lambda_i^{(d,m)}$ its expected value. We assume that $a_i^{(d,m)}(t)$ is independently and identically distributed (i.i.d.) across timeslots and its forth moment is upper bounded,[2] *i.e.*, $\mathbb{E}\{(\sum_{(d,m)} a_i^{(d,m)}(t))^4\} \leq (A_{\max})^4$. Recall that in an AgI service only the source commodity $(d,0)$ enters the network exogenously, while all other commodities are created inside the network as the output of a service function. Hence, $a_i^{(d,m)}(t) = 0$, for all i,t when $m > 0$.

During AgI service delivery, internal network queues buffer incoming data according to their commodities. We define the *queue backlog* of commodity (d,m) at node i, $Q_i^{(d,m)}(t)$, as the amount (in information units) of commodity (d,m) in the

[2]The upper bound of the fourth moment is used in the proof of convergence with probability 1 in Theorem 2.

queue of node i at the beginning of timeslot t, which evolves over time as follows:

$$Q_i^{(d,m)}(t+1) \leq \left[Q_i^{(d,m)}(t) - \sum\nolimits_{j:j\neq i} \mu_{ij}^{(d,m)}(t) - \mu_{i,\text{pr}}^{(d,m)}(t)\right]^+ + \sum\nolimits_{j:j\neq i} \mu_{ji}^{(d,m)}(t) + \mu_{\text{pr},i}^{(d,m)}(t) + a_i^{(d,m)}(t), \quad (7)$$

where $[x]^+$ denotes $\max\{x, 0\}$.

Note that, in an AgI service chain, only the final commodity (d, M) is allowed to exit the network once it arrives to its destination $d \in \mathcal{D}$, while any other commodity (d, m), $m < M$, can only get consumed by being processed into the next commodity $(d, m+1)$. Commodity (d, M) is assumed to leave the network immediately upon arrival/decoding at its destination, *i.e.,* $Q_d^{(d,M)}(t) = 0$, for all d, t.

G. Network Objective

The goal is to design a control algorithm that dynamically schedules, routes, and process service flows over the wireless computing network with minimum total average resource cost,

$$\limsup_{t\to\infty} \quad \frac{1}{t} \sum\nolimits_{\tau=0}^{t-1} \mathbb{E}\left\{h(\tau)\right\}, \quad (8)$$

where $h(t)$ is the total cost of the network at time t,

$$h\left(t\right) \triangleq \sum_{i\in\mathcal{N}} \left[\sum_{k=0}^{K_i^{\text{pr}}} w_{i,k}^{\text{pr}}\ y_{i,k}^{\text{pr}}(t) + \sum_{k=0}^{K_i^{\text{tr}}} w_{i,k}^{\text{tr}}\ y_{i,k}^{\text{tr}}(t)\right]. \quad (9)$$

III. Wireless Computing Network Capacity Region

Given a set of AgI services, the wireless computing network capacity region Λ is defined as the closure of all service input rates $\{\lambda_i^{(d,m)}\}$ that can be stabilized by a control algorithm.

Theorem 1. *The wireless computing network capacity region Λ consists of all average exogenous input rates $\{\lambda_i^{(d,m)}\}$ for which there exist multi-commodity flow variables $f_{ij}^{(d,m)}$, $f_{\text{pr},i}^{(d,m)}$, $f_{i,\text{pr}}^{(d,m)}$, together with probability values $\alpha_{i,k}^{\text{pr}}$, $\alpha_{i,k}^{\text{tr}}(\mathbf{s})$, $\beta_{i,\text{pr}}^{(d,m)}(k)$, $\beta_{i,\text{tr}}^{(d,m)}(\mathbf{s},k)$, $\eta_{ij}^{(d,m)}(\mathbf{s},k,n)$, for all $i, j \neq i, k, d, m$, and all network states $\mathbf{s} \in \mathcal{S}$, such that:*

$$\sum\nolimits_j f_{ji}^{(d,m)} + f_{\text{pr},i}^{(d,m)} + \lambda_i^{(d,m)} \leq \sum\nolimits_j f_{ij}^{(d,m)} + f_{i,\text{pr}}^{(d,m)}, \quad \forall i, d, m < M \ \text{ or } \ \forall i \neq d, m = M, \quad (10a)$$

$$f_{\text{pr},i}^{(d,m+1)} = \xi^{(m+1)} f_{i,\text{pr}}^{(d,m)}, \quad \forall i, d, m < M, \quad (10b)$$

$$f_{i,\text{pr}}^{(d,m)} \leq \frac{1}{r^{(m+1)}} \sum\nolimits_{k=0}^{K_i^{\text{pr}}} \alpha_{i,k}^{\text{pr}} \beta_{i,\text{pr}}^{(d,m)}(k) R_{i,k}, \quad \forall i, d, m < M, \quad (10c)$$

$$f_{ij}^{(d,m)} \leq \sum\nolimits_{\mathbf{s}\in\mathcal{S}} \pi_{\mathbf{s}} \sum\nolimits_{k=0}^{K_i^{\text{tr}}} \alpha_{i,k}^{\text{tr}}(\mathbf{s}) \beta_{i,tr}^{(d,m)}(\mathbf{s},k) \times \sum\nolimits_{n=1}^{q_{i,\mathbf{s}}^{-1}(j)} \left[R_{iq_{in},k}(\mathbf{s}) - R_{iq_{in-1},k}(\mathbf{s})\right] \eta_{ij}^{(d,m)}(\mathbf{s},k,n), \quad \forall i, j, d, m, \quad (10d)$$

$$f_{i,\text{pr}}^{(d,M)} = 0, \ f_{\text{pr},i}^{(d,0)} = 0, \ f_{dj}^{(d,M)} = 0, \ f_{i,\text{pr}}^{(d,m)} \geq 0, \ f_{ij}^{(d,m)} \geq 0, \quad \forall i, j, d, m, \quad (10e)$$

$$\sum\nolimits_{k=0}^{K_i^{\text{pr}}} \alpha_{i,k}^{\text{pr}} \leq 1, \quad \sum\nolimits_{k=0}^{K_i^{\text{tr}}} \alpha_{i,k}^{\text{tr}}(\mathbf{s}) \leq 1, \quad \forall i, \mathbf{s}, \quad (10f)$$

$$\sum\nolimits_{(d,m)} \beta_{i,\text{pr}}^{(d,m)}(k) \leq 1, \ \sum\nolimits_{(d,m)} \beta_{i,\text{tr}}^{(d,m)}(\mathbf{s},k) \leq 1, \quad \forall i, \mathbf{s}, k, \quad (10g)$$

$$\sum\nolimits_j \eta_{ij}^{(d,m)}(\mathbf{s},k,n) \leq 1, \quad \forall i, \mathbf{s}, k, n, \quad (10h)$$

where $\mathbf{s} \in \mathcal{S}$ denotes the network state, whose (i,j)-th element $(\mathbf{s})_{ij}$ indicates the channel state of link (i,j), $\pi_{\mathbf{s}}$ denotes the steady state probability of the network state process $\mathbf{S}(t)$ for each $\mathbf{s} \in \mathcal{S}$, and $q_{i,\mathbf{s}}^{-1}(j)$ in (10d) is the index of node j in the sequence $\{q_{i,1}, \cdots, q_{i,N-1}\}$, given network state $\mathbf{s}$. Finally, with a slight abuse of notation, $R_{ij,k}(\mathbf{s})$ in (10d) denotes the maximum achievable rate over link (i,j), given network state $\mathbf{s}$ and the allocation of k transmission resource units.

Furthermore, the minimum average network cost required for network stability is given by

$$\bar{h}^* = \min \underline{h} \quad (11)$$

where

$$\underline{h} = \sum_{i\in\mathcal{N}} \left(\sum_{k=0}^{K_i^{\text{pr}}} \alpha_{i,k}^{\text{pr}} w_{i,k}^{\text{pr}} + \sum_{k=0}^{K_i^{\text{tr}}} w_{i,k}^{\text{tr}} \sum_{\mathbf{s}\in\mathcal{S}} \pi_{\mathbf{s}} \alpha_{i,k}^{\text{tr}}(\mathbf{s})\right), \quad (12)$$

and the minimization is over all $\alpha_{i,k}^{\text{pr}}$, $\alpha_{i,k}^{\text{tr}}(\mathbf{s})$, $\beta_{i,\text{pr}}^{(d,m)}(k)$, $\beta_{i,\text{tr}}^{(d,m)}(\mathbf{s},k)$, and $\eta_{ij}^{(d,m)}(\mathbf{s},k,n)$ satisfying (10a)-(10h). □

Proof. See Appendix A. □

In Theorem 1, Eq. (10a) are flow conservation constraints,[3] Eqs. (10c) and (10d) are rate constraints, and Eq. (10e) indicates non-negativity and flow efficiency constraints. The probability values $\alpha_{i,k}^{*\text{pr}}$, $\alpha_{i,k}^{*\text{tr}}(\mathbf{s})$, $\beta_{i,\text{pr}}^{*(d,m)}(k)$, $\beta_{i,\text{tr}}^{*(d,m)}(\mathbf{s},k)$ and $\eta_{ij}^{*(d,m)}(\mathbf{s},k,n)$ define a *stationary randomized policy* that uses *single-copy routing* – only one copy of each information unit is allowed to flow through the network - and it is optimal among all stabilizing algorithms (including algorithms that use *multi-copy routing*). Specifically, the parameters of the stationary randomized policy are defined as:

- $\alpha_{i,k}^{\text{pr}}$: the probability that k processing resource units are allocated at node i
- $\alpha_{i,k}^{\text{tr}}(\mathbf{s})$: the conditional probability that k transmission resource units are allocated at node i, given the network state $\mathbf{s}$;
- $\beta_{i,\text{pr}}^{(d,m)}(k)$: the conditional probability that node i processes commodity (d,m), given the allocation of k processing resource units;
- $\beta_{i,\text{tr}}^{(d,m)}(\mathbf{s},k)$: the conditional probability that node i transmits commodity (d,m), given network state $\mathbf{s}$ and the allocation of k transmission resource units;
- $\eta_{ij}^{(d,m)}(\mathbf{s},k,n)$: the conditional probability that node i forwards the information of commodity (d,m) in the n-th partition to node j, when the network state is $\mathbf{s}$ and k transmission resource units are allocated.

It is important to note that this optimal stationary randomized policy is hard to compute in practice, as it requires the knowledge of $\{\lambda_i^{(d,m)}\}$ and solving a complex nonlinear

[3]Note that the final commodity (d, M) staisfies flow conservation at all nodes except at its destination d, where it is immediately consumed upon arrival.

program. However, its existence is essential for proving the performance of our proposed algorithm.

IV. Dynamic Wireless Computing Network Control Algorithm

In this section, we propose a dynamic wireless computing network control (DWCNC) strategy that accounts for both transmission and processing flow scheduling and resource allocation decisions in a fully distributed manner. The proposed DWCNC algorithm online minimizes a linear metric extracted from an upper bound of the LDP function derived for the wireless computing network.

Let $\mathbf{Q}(t)$ represent the vector of queue backlog values of all the commodities at all the network nodes. The network's Lyapunov drift is defined as

$$\Delta\left(\mathcal{H}\left(t\right)\right) \triangleq \frac{1}{2}\mathbb{E}\left\{ \left\|\mathbf{Q}\left(t+1\right)\right\|^2 - \left\|\mathbf{Q}\left(t\right)\right\|^2 \middle| \mathcal{H}\left(t\right)\right\}, \quad (13)$$

where $\mathcal{H}(t) \triangleq \{\mathbf{Q}(t), \mathbf{S}(t-1)\}$ is the ensemble of queue backlog observations at time t and the CSI feedbacks at time $t-1$; $\|\mathbf{x}\|$ indicates Euclidean norm of a vector $\mathbf{x}$, and the expectation is taken over the ensemble of all the exogenous source commodity arrival realizations and channel realizations at time t. Let $\boldsymbol{\lambda}$ denote the vector form of $\{\lambda_i^{(d,m)}\}$ and have $\boldsymbol{\lambda}^T$ as its transpose. After standard LDP algebraic manipulations on (7) (see Ref. [14]), we obtain

$$\begin{aligned}\Delta(\mathcal{H}(t))+V\mathbb{E}\{h(t)|\mathcal{H}(t)\} \leq NB+\boldsymbol{\lambda}^T\mathbf{Q}(t)+ \\ \sum\nolimits_i \mathbb{E}\left\{Vh_i^{\text{pr}}(t)-Z_i^{\text{pr}}(t)+Vh_i^{\text{tr}}(t)-Z_i^{\text{tr}}(t)\middle|\mathcal{H}(t)\right\} \\ \triangleq NB+\boldsymbol{\lambda}^T\mathbf{Q}(t)+\Psi(t), \quad (14)\end{aligned}$$

where V is a non-negative control parameter introduced to tune the degree to which cost is emphasized with respect to congestion reduction; and $Z_i^{\text{pr}}(t)$, h_i^{pr}, $Z_i^{\text{tr}}(t)$, $h_i^{\text{tr}}(t)$, and the constant B (with $r_{\min} \triangleq \min_m\{r^{(m)}\}$ and $\xi_{\max} \triangleq \max_m\{\xi^{(m)}\}$) are given by

$$\begin{aligned}&Z_i^{\text{pr}}(t) \triangleq \sum\nolimits_{(d,m)} \mu_{i,\text{pr}}^{(d,m)}(t)\left[Q_i^{(d,m)}(t) - \xi^{(m+1)}Q_i^{(d,m+1)}(t)\right],\\
&Z_i^{\text{tr}}(t) \triangleq \sum\nolimits_{u=1}^{N-1}\sum\nolimits_{(d,m)} \mu_{iq_{i,u}}^{(d,m)}(t)\left[Q_i^{(d,m)}(t) - Q_{q_{i,u}}^{(d,m)}(t)\right],\\
&h_i^{\text{pr}}(t) \triangleq \sum\nolimits_{k=0}^{K_i^{\text{pr}}} w_{i,k}^{\text{pr}} y_{i,k}^{\text{pr}}(\tau), \quad h_i^{\text{tr}}(t) \triangleq \sum\nolimits_{k=0}^{K_i^{\text{tr}}} w_{i,k}^{\text{tr}} y_{i,k}^{\text{tr}}(\tau),\\
&B \triangleq \frac{1}{2}\max_i\left\{\left(\max_{j,\mathbf{s}:j\neq i,\mathbf{s}\in S}\left\{R_{ij,K_i^{\text{tr}}}(\mathbf{s})\right\}+\frac{R_{i,K_i^{\text{pr}}}}{r_{\min}}\right)^2+\right.\\
&\left.\left(\max_{\mathbf{s}\in S}\left\{\sum_{j:j\neq i} R_{ji,K_j^{\text{tr}}}(\mathbf{s})\right\}+\frac{\xi_{\max}R_{i,K_i^{\text{pr}}}}{r_{\min}}+A_{\max}\right)^2\right\},\end{aligned}$$

At each timeslot t, DWCNC is designed to make decisions that minimize $\Psi(t)$ based on the observation of $\mathcal{H}(t)$ and subject to the constraints $y_{i,k}^{\text{pr}}(t) \in \{0,1\}$, $y_{i,k}^{\text{tr}}(t) \in \{0,1\}$, $\sum_{k=0}^{K_i^{\text{pr}}} y_{i,k}^{\text{pr}}(t) \leq 1$, $\sum_{k=0}^{K_i^{\text{tr}}} y_{i,k}^{\text{tr}}(t) \leq 1$ and (1), (2), (5), (6).

A. Algorithm Description

Note that the minimization of $\Psi(t)$ is decomposable among network nodes and between processing and transmission, and therefore enables a distributed implementation of DWCNC, where each node i can make local decisions to minimize $\mathbb{E}\{Vh_i^{\text{pr}}(t) - Z_i^{\text{pr}}(t)|\mathcal{H}(t)\}$ (for processing) and $\mathbb{E}\{Vh_i^{\text{tr}}(t) - Z_i^{\text{tr}}(t)|\mathcal{H}(t)\}$ (for transmission). The minimization is summarized by Lemma B.1 in Appendix B. The resulting distributed DWCNC algorithm works as follows:

Dynamic Wireless Computing Network Control (DWCNC):

Local processing decisions: At the beginning of timeslot t, each node i observes its local queue backlogs and performs the following operations:

1) Compute the *processing utility weight* of each commodity (d,m):
$$W_i^{(d,m)}(t) \triangleq \frac{1}{r^{(m+1)}}\left[Q_i^{(d,m)}(t) - \xi^{(m+1)}Q_i^{(d,m+1)}(t)\right]^+.$$
Note that $W_i^{(d,m)}(t)$ indicates the "potential benefit" of executing function $(m+1)$ to process commodity (d,m) into commodity $(d,m+1)$ at time t, in terms of local congestion reduction per processing operation.
2) Compute the optimal commodity $(d,m)_{\text{pr}}^{\dagger}$ to process:
$$(d,m)_{\text{pr}}^{\dagger} = \arg\max_{(d,m)}\left\{W_i^{(d,m)}(t)\right\}.$$
3) Make resource allocation decision for processing:
$$k_{\text{pr}}^{\dagger} = \arg\max_{k\in\mathcal{K}_i^{\text{pr}}}\left\{R_{i,k}W_i^{(d,m)_{\text{pr}}^{\dagger}}(t) - Vw_{i,k}^{\text{pr}}\right\}.$$
4) Make the following flow rate assignment decisions:
$$\mu_{i,\text{pr}}^{(d,m)_{\text{pr}}^{\dagger}}(t) = R_{i,k_{\text{pr}}^{\dagger}}\Big/ r^{(m_{\text{pr}}^{\dagger}+1)};$$
$$\mu_{i,\text{pr}}^{(d,m)}(t) = 0, \quad \forall (d,m) \neq (d,m)_{\text{pr}}^{\dagger}.$$

Remarks: the computational complexity of the local processing decisions is $O(K_i^{\text{pr}} + MD)$.

Local wireless transmission decisions: At the beginning of timeslot t, each node i observes its local queue backlogs, the queue backlogs of its potential RXs and the associated statistical CSI, and performs the following operations:

1) For each outgoing link (i,j) and commodity (d,m), compute the *differential backlog weight*:
$$W_{ij}^{(d,m)}(t) \triangleq \left[Q_i^{(d,m)}(t) - Q_j^{(d,m)}(t)\right]^+.$$
2) For each transmission resource allocation choice $k \in \{0,\ldots,K_i^{\text{tr}}\}$, compute the *transmission utility weight* of each commodity (d,m):
$$\begin{aligned}&W_{i,k,\text{tr}}^{(d,m)}(t) \triangleq \sum_{\mathbf{s}\in S}\Pr\left(\mathbf{S}(t)=\mathbf{s}|\,\mathbf{S}(t-1)=\tilde{\mathbf{s}}\right)\times\\
&\sum_{n=1}^{N-1}\left[R_{iq_{i,n},k}(\mathbf{s}) - R_{iq_{i,n-1},k}(\mathbf{s})\right]\max_{j\in\Omega_{i,n}(\mathbf{s})}\left\{W_{ij}^{(d,m)}(t)\right\}, \quad (15)\end{aligned}$$
where $\tilde{\mathbf{s}}$ denotes the CSI feedbacks at time $t-1$, and, with an abuse of notation, $\Omega_{i,n}(\mathbf{s})$ is used to indicate the dependence of $\Omega_{i,n}$ on the network state $\mathbf{s}$.
3) Compute the optimal number of transmission resource units $k_{\text{tr}}^{\dagger}$ to allocate and the optimal commodity $(d,m)_{\text{tr}}^{\dagger}$ to transmit: $\left[k_{\text{tr}}^{\dagger}, (d,m)_{\text{tr}}^{\dagger}\right] =$

$\arg\max_{k,(d,m)} \left\{ W_{i,k,\text{tr}}^{(d,m)}(t) - V w_{i,k}^{\text{tr}} \right\}$. If $k_{\text{tr}}^{\dagger} = 0$, node i keeps silent in timeslot t.

4) After receiving the CSI feedbacks, node i identifies the information decoded by all the RXs and the experienced $\mathbf{S}(t)$, and assigns the forwarding responsibility for the n-th partition of the transmitted information to the RX in $\Omega_{i,n}(\mathbf{S}(t))$ with the largest positive $W_{ij}^{(d,m)}(t)$, while all other RXs in $\Omega_{i,n}(\mathbf{S}(t))$ and node i discard the information. If no receiver in $\Omega_{i,n}(\mathbf{S}(t))$ has positive $W_{ij}^{(d,m)}(t)$, node i retains the information of partition n, while all the receivers in $\Omega_{i,n}(\mathbf{S}(t))$ discard it.

Remarks:

- In Step 2 of the local transmission decisions, $W_{i,k,\text{tr}}^{(d,m)}(t)$ is computed according to (15) using the transition probabilities $\Pr(\mathbf{S}(t) = \mathbf{s} | \mathbf{S}(t-1) = \tilde{\mathbf{s}})$, known as the statistical CSI, but the complexity can be high due to the possibly exponentially large network state space with respect to the number of links. However, the computation can be simplified when using discrete code layers for the broadcast approach, which is described in the next subsection.
- Discarding decoded information at the RXs that do not get the processing/forwarding responsibility, during Step 4 of the local transmission decisions, implies that DWCNC is a single-copy routing algorithm.

B. Transmission Utility Weight with Discrete Code Layers

Recall that, in practice, when using the broadcast approach, each node uses L_i discrete code layers, with $R_{ij,k}(t)$ taking values in $\{R_{i,k}^l : 0 \le l \le L_i\}$ as described in Sec. II-D).

Let $\bar{\Omega}_{i,l}(\mathbf{S}(t))$ denote the set of receivers that have channel gain no smaller than $\bar{g}_{i,l}$ at time t, *i.e.,* $g_{ij}(t) \ge \bar{g}_{i,l}$ for all $j \in \bar{\Omega}_{i,l}(\mathbf{S}(t))$, and $g_{ij}(t) < \bar{g}_{i,l}$ for all $j \notin \bar{\Omega}_{i,l}(\mathbf{S}(t))$.

Given $\mathbf{S}(t) = \mathbf{s}$ and $y_{i,k}^{\text{tr}}(t) = 1$, we have the following two possible cases for the maximum achievable transmission rate of the n-th partition: i) $R_{iq_{i,n},k}(\mathbf{s}) - R_{iq_{i,n-1},k}(\mathbf{s}) = 0$; ii) $R_{iq_{i,n},k}(\mathbf{s}) - R_{iq_{i,n-1},k}(\mathbf{s}) = \sum_{l=l_0}^{l_1} \left(\bar{R}_{i,k}^l - \bar{R}_{i,k}^{l-1} \right)$, for some l_0 and l_1 satisfying $1 \le l_0 \le l_1 \le L_i$, with $\bar{\Omega}_{i,l}(\mathbf{s}) = \Omega_{i,n}(\mathbf{s})$ for all $l_0 \le l \le l_1$. Then we have, for all i, $\mathbf{s}$, k, (d,m), t,

$$\sum_{n=1}^{N-1} \left[R_{iq_{i,n},k}(\mathbf{s}) - R_{iq_{i,n-1},k}(\mathbf{s}) \right] \max_{j \in \Omega_{i,n}(\mathbf{s})} \left\{ W_{ij}^{(d,m)}(t) \right\} = \sum_{l=1}^{L_i} \left(\bar{R}_{i,k}^l - \bar{R}_{i,k}^{l-1} \right) \max_{j \in \bar{\Omega}_{i,l}(\mathbf{s})} \left\{ W_{ij}^{(d,m)}(t) \right\},$$

based on which we can rewrite Eq. (15) as follows:

$$\begin{aligned} W_{i,k,\text{tr}}^{(d,m)}(t) &= \sum_{l=1}^{L_i} \left(\bar{R}_{i,k}^l - \bar{R}_{i,k}^{l-1} \right) \times \sum_{\mathbf{s} \in \mathcal{S}} \Pr(\mathbf{S}(t) = \mathbf{s} | \mathbf{S}(t-1)) \max_{j \in \bar{\Omega}_{i,l}(\mathbf{s})} \left\{ W_{ij}^{(d,m)}(t) \right\} \\ &= \sum_{l=1}^{L_i} \left(\bar{R}_{i,k}^l - \bar{R}_{i,k}^{l-1} \right) \mathbb{E} \left\{ \max_{j \in \bar{\Omega}_{i,l}(\mathbf{S}(t))} \left\{ W_{ij}^{(d,m)}(t) \right\} \middle| \mathcal{H}(t) \right\}. \end{aligned} \tag{16}$$

Let $1_{ij,l}^{(d,m)}(t)$ denote the indicator that takes value 1 if receiver j has the largest differential backlog $W_{ij}^{(d,m)}(t)$ among the receivers in $\bar{\Omega}_{i,l}(\mathbf{S}(t))$, and 0 otherwise. Then, we have

$$\begin{aligned} &\mathbb{E} \left\{ \max\nolimits_{j \in \bar{\Omega}_{i,l}(\mathbf{S}(t))} \left\{ W_{ij}^{(d,m)}(t) \right\} \middle| \mathcal{H}(t) \right\} \\ &= \mathbb{E} \left\{ \sum\nolimits_j W_{ij}^{(d,m)}(t) \, 1_{ij,l}^{(d,m)}(\mathbf{S}(t)) \middle| \mathcal{H}(t) \right\} \\ &= \sum\nolimits_j W_{ij}^{(d,m)}(t) \, \varphi_{ij,l}^{(d,m)}(\mathcal{H}(t)), \end{aligned} \tag{17}$$

where $\varphi_{ij,l}^{(d,m)}(\mathcal{H}(t))$ is the conditional probability that $1_{ij,l}^{(d,m)}(t)$ takes value 1 given $\mathcal{H}(t)$.

Plugging (17) into (16) to compute $W_{i,k,\text{tr}}^{(d,m)}(t)$, we replace Step 2 of the local transmission decisions of DWCNC in Sec. IV-A with the following two sub-steps:

2a) For each commodity (d,m), sort the receivers of node i according to their differential backlog weight $W_{ij}^{(d,m)}(t)$ in non-increasing order. Let $\Psi_{ij}^{(d,m)}(t)$ denote the set of receivers of node i with index smaller than the index of receiver j in the sorted list at time t. In this case, each receiver in $\Psi_{ij}^{(d,m)}(t)$ has no smaller differential backlog weight than receiver j.

2b) For each transmission resource allocation choice $k \in \{0, \ldots, K_i^{\text{tr}}\}$, compute the *transmission utility weight* of each commodity (d,m):

$$W_{i,k,\text{tr}}^{(d,m)}(t) = \sum_{l=1}^{L_i} \left(\bar{R}_{i,k}^l - \bar{R}_{i,k}^{l-1} \right) \sum_j W_{ij}^{(d,m)}(t) \, \varphi_{ij,l}^{(d,m)}(\mathcal{H}(t)), \tag{18}$$

where the conditional probability $\varphi_{ij,l}^{(d,m)}(\mathcal{H}(t))$ can be expressed as

$$\varphi_{ij,l}^{(d,m)}(\mathcal{H}(t)) = \Pr \left\{ g_{ij}(t) \ge \bar{g}_{i,l}, \max_{v \in \Psi_{ij}^{(d,m)}(t)} \{ g_{iv}(t) \} < \bar{g}_{i,l} \middle| \mathcal{H}(t) \right\}. \tag{19}$$

According to (19), the joint conditional probability $\varphi_{ij,l}^{(d,m)}(\mathcal{H}(t))$ for all the $j \in \mathcal{N} \setminus \{i\}$ can be estimated by a proper shadowing correlation model in the, possibly complicated, physical wireless scenario where the wireless computing network is located [22]. Computing $\varphi_{ij,l}^{(d,m)}(\mathcal{H}(t))$ can be significantly simplified if the channel gains of the links are mutually independent, which is the case when the mutual distances between the receiving nodes exceed the shadowing de-correlation distances [22]. In this scenario, it follows from (19) that

$$\begin{aligned} \varphi_{ij,l}^{(d,m)}(\mathcal{H}(t)) &= \Pr(g_{ij}(t) \ge \bar{g}_{i,l} | s_{ij}(t-1)) \times \prod\nolimits_{v \in \Psi_{ij}^{(d,m)}(t)} \Pr(g_{iv}(t) < \bar{g}_{i,l} | s_{iv}(t-1)) \\ &= \sum\nolimits_{l'=l}^{L_i} \Pr(s_{ij}(t) = \bar{s}_{i,l'} | s_{ij}(t-1)) \times \prod\nolimits_{v \in \Psi_{ij}^{(d,m)}(t)} \sum\nolimits_{l'=0}^{l-1} \Pr(s_{iv}(t) = \bar{s}_{i,l'} | s_{iv}(t-1)), \end{aligned} \tag{20}$$

where $\Pr(s_{ij}(t) = \bar{s}_{i,l'} | s_{ij}(t-1) = \bar{s}_{i,l''})$ is the statistical CSI of link (i,j).

In addition, when using L_i discrete code layers, the trans-

mitted information can be re-grouped into L_i partitions, each of which is decoded by the RX set $\bar{\Omega}_{i,l}(\mathbf{S}(t))$. Correspondingly, the step of making the forwarding decision for each partition is the same as Step 4 of the local transmission decisions of DWCNC in Sec. IV-A, except replacing $\Omega_{i,n}(\mathbf{S}(t))$ by $\bar{\Omega}_{i,l}(\mathbf{S}(t))$.

With discrete code layers and independent outgoing links, the computational complexity associated with the transmission decisions made by node i at each timeslot is $O(MDL_i(N+K_i^{\text{tr}}))$, which is dominated by computing the transmission utility weights for all commodities and resource allocation choices.

V. Performance Analysis

In this section, we analyze the throughput-optimality and average cost performance of DWCNC by extending conventional LDP analysis to wireless computing networks. The extension is in two fold: i) the analysis accounts for the MCC flow model with flow chaining and scaling instead of the traditional multi-commodity flow model; ii) the analysis for wireless transmission scheduling accounts for the effect of using the broadcast approach coding scheme instead of the traditional single-layer outage approach [15]. The resulting performance characterization is summarized by the following theorem:

Theorem 2. *For any exogenous input rate vetor $\boldsymbol{\lambda}$ strictly interior to the capacity region Λ, DWCNC stabilizes the wireless computing network, while achieving an average total resource cost arbitrarily close to the minimum average cost $\overline{h}^*(\boldsymbol{\lambda})$ with probability 1 (w.p.1); i.e.,*

$$\limsup_{t\to\infty}\frac{1}{t}\sum\nolimits_{\tau,i,d,m} Q_i^{(d,m)}(\tau) \leq \frac{1}{\epsilon}\left[NB+V\left(\overline{h}^*(\boldsymbol{\lambda}+\epsilon\mathbf{1})-\overline{h}^*(\boldsymbol{\lambda})\right)\right], \quad w.p.1, \tag{21}$$

$$\limsup_{t\to\infty}\frac{1}{t}\sum\nolimits_{\tau=0}^{t-1} h(\tau)\leq \overline{h}^*(\boldsymbol{\lambda})+\frac{NB}{V}, \quad w.p.1, \tag{22}$$

where ϵ is a positive constant satisfying $(\boldsymbol{\lambda}+\epsilon\mathbf{1})\in\Lambda$; and $\overline{h}^*(\boldsymbol{\lambda})$ denotes the average cost obtained by the optimal solution given input rates $\boldsymbol{\lambda}$. □

Proof. See Appendix B. □

In Theorem 2, the finite bound on the total queue backlog shown in Eq. (21) demonstrates that the wireless computing network is *strongly stable* with $\boldsymbol{\lambda}$ interior to Λ. Note that strong stability implies rate stability, i.e., that AgI service delivery rates asymptotically match the corresponding exogenous arrival rates (see Eq. (23) in Appendix A).

According to Eq. (21) and (22), decreasing V results in a lower average total network backlog bound, but no average cost efficiency guarantee. On the other hand, the parameter V can be increased to push the average resource cost arbitrarily close to the minimum cost required for network stability, $\overline{h}^*(\boldsymbol{\lambda})$, with a linear increase in average total network backlog bound (and hence, by Little's Theorem, average delay bound). Thus, Theorem 2 demonstrates a $[O(1/V), O(V)]$ cost-delay tradeoff.

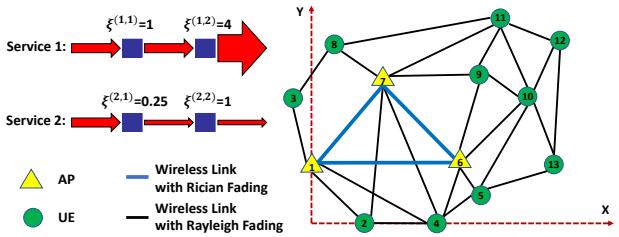


Fig. 5. A wireless computing network with three access points and eight user equipments, providing two AgI services.

TABLE I
Computing Nodes' Locations

Node Index	1	2	3	4	5	6	7
Location (X,Y)	(0,10)	(10,0)	(-5, 20)	(22, 0)	(27, 5)	(24, 10)	(13, 22)
Node Index	8	9	10	11	12	13	
Location (X,Y)	(5, 30)	(27, 23)	(35, 21)	(30, 33)	(40, 31)	(39, 9)	

VI. Numerical Experiments

In this section, we present numerical results obtained from simulating the performance of the DWCNC algorithm for the delivery of two AgI services over a wireless computing network with 13 nodes during 10^6 timeslots. The numerical values presented in this section for resource allocation costs, communication flow rates, processing flow rates, and queue backlogs are all measured in normalized units.

A. Network Structure

We consider a wireless computing network with 13 computing nodes, as illustrated in Fig. 5. Nodes 1, 6, and 7 represent access points (APs), while all other nodes are user equipments (UEs). We list the (X,Y) coordinates of all nodes' locations in Table I in normalized distance units. In terms of processing resources, each AP has five resource allocation choices, with associated cost and processing rate $w_{i,k}^{\text{pr}}=k$ and $R_{i,k}=20k$, $i\in\{1,6,7\}$ and $k\in\{0,1,\ldots,4\}$, while each UE has two resource allocation choices, with associated cost and rate $w_{i,k}^{\text{pr}}=2k$ and $R_{i,k}=20k$, $i\in\mathcal{N}\setminus\{1,6,7\}$ and $k\in\{0,1\}$. Note that the processing is cheaper at the APs than at the UEs. In terms of transmission resources, each node has two resource allocation choices of cost $w_{i,0}^{\text{tr}}=0$ and $w_{i,1}^{\text{tr}}=1$. The associated transmission rates are given in Section VI-B.

The edges in Fig. 5 represent the active wireless links, whose channel realizations are assumed mutually independent, i.e., we assume that the distances between nodes exceed the shadowing de-correlation distance. Each link suffers both small scale fading and large scale fading. The realizations of small scale fading of each link are independently and identically distributed (i.i.d.) across timeslots, which is approximately fullfilled when the timeslot length is the coherence time of the wireless medium. We assume that links between APs have Rician fading (see Ref. [23]) with Rice factor equal to 15 dB, while the rest of the links exhibit Rayleigh fading (see Ref. [23]). The path loss coefficient of each Rician and Rayleigh fading link is 2 and 3, respectively. Regarding the large scale fading, we assume time-variant shadowing components existing in the environment, *e.g.,* moving cars, which cause time-variant log-normal shadowing on each link. The shadowing's evolution over timeslots can be characterized

by a first order autoregressive model [25]: $F_{ij}(t+1) = \rho F_{ij}(t) + \sqrt{1-\rho^2} n_{ij}(t)$, where $F_{ij}(t)$ is the logarithm of the large scale fading variable for link (i,j) at time t, that satisfies Gaussian distribution with zero mean and variance σ_P equal to 4 dB for Rician fading links and 6 dB for Rayleigh fading links; the $n_{ij}(t)$ are Gaussian random variables for link (i,j) with zero mean and variance equal to σ_P and are i.i.d. across timeslots; ρ is a correlation coefficient for shadowing over one timeslot interval. In this simulation, we compute $\rho = \exp(-\delta/d_c)$, where δ is the maximum moving distance in one timeslot (with length of coherence time), and d_c is the shadowing de-correlation distance. Note that δ is approximately equal to one wavelength,[4] and therefore, by assuming that the carrier frequency is 1 GHz, we obtain $\delta = 0.3$ meters. By further assuming a shadowing de-correlation distance of 15 meters, we then obtain $\rho = 0.98$, which is very close to 1, implying that shadowing evolves slowly across timeslots. Setting $\rho = 0.98$ can also be validated by the experimental result in [26].

B. Communication Setup

To demonstrate the efficiency of adopting the broadcast approach, we first simulate the case of adopting the traditional *outage approach* coding scheme, under which each node i only uses a single coding layer to which all the power $P_{i,1}^{\text{tot}}$ is allocated. The positive transmission rate, denoted by $\bar{R}_{i,1}^{\text{out}}$, is fixed under the outage approach, and the information is reliably decoded when the instantaneous channel gain exceeds a threshold $\bar{g}_i^{\text{out}}$. Otherwise, no information is decoded. We set $P_{i,1}^{\text{tot}}$ to be the same among APs and UEs, respectively. The value of $P_{i,1}^{\text{tot}}$ is chosen such that, if having transmitted the signal using power $P_{i,1}^{\text{tot}}$ through the link with the largest path loss, the average SNR at the receiver would be 5 dB and 0 dB for node i being AP and UE, respectively. The value of $\bar{g}_i^{\text{out}}$ is heuristically chosen as -40.80 dB and -38.37 dB for node i being AP and UE, respectively. Then, with bandwidth $B_W = 10$ (measured in normalized units), we generate the maximum achievable rate as $\bar{R}_{i,1}^{\text{out}} = B_W \log_2(1 + P_{i,1}\bar{g}_i^{\text{out}})$ equal to 23.90 and 16.02 for node i being AP and UE, respectively.

When simulating the broadcast approach coding scheme, we assume that each node i uses three code layers, where the RX decoding the 2nd code layer gets the same rate as the outage approach, *i.e.,* $\bar{R}_{i,1}^2 = \bar{R}_{i,1}^{\text{out}}$, while decoding the 1st and 3rd layers requires worse and better channel condition than the 2nd layer, respectively, and therefore $\bar{R}_{i,1}^1 < \bar{R}_{i,1}^{\text{out}}$ and $\bar{R}_{i,1}^3 > \bar{R}_{i,1}^{\text{out}}$. Node i allocates the total power $P_{i,1}^{\text{tot}}$ among the code layers for transmission with fractions $[\frac{2}{3} : \frac{1}{6} : \frac{1}{6}]$ and $[\frac{5}{8} : \frac{1}{4} : \frac{1}{8}]$ for node i being AP and UE, respectively. The channel gain thresholds corresponding to the three layers are set to $[\bar{g}_{i,1} = -42.28, \bar{g}_{i,2} = -36.26, \bar{g}_{i,3} = -34.50]$ dB and $[\bar{g}_{i,1} = -38.48, \bar{g}_{i,2} = -33.83, \bar{g}_{i,3} = -31.87]$ dB for node i being AP and UE, respectively.[5] By applying Eq. (4), we generate the maximum achievable transmission rates $[\bar{R}_{i,1}^0 = 0, \bar{R}_{i,1}^1 = 14.45, \bar{R}_{i,1}^2 = 23.90, \bar{R}_{i,1}^3 = 66.81]$ and $[\bar{R}_{i,1}^0 = 0, \bar{R}_{i,1}^1 = 9.74, \bar{R}_{i,1}^2 = 16.02, \bar{R}_{i,1}^3 = 40.68]$ for node i being AP and UE, respectively.

C. Service Setup

The wireless computing network offers two services (see Fig. 5), each of which consists of two functions. To indicate multiple services, we let (ϕ, m), $\phi = 1, 2$, $m = 1, 2$, denote the m-th function of service ϕ; and (d, ϕ, m), $d \in \mathcal{N} \backslash \{1, 6, 7\}$, $\phi = 1, 2$, $m = 0, 1$, denote the commodity generated by function (ϕ, m) for destination d.

All four functions have the same complexity factor equal to 1, *i.e.,* $r^{(\phi,m)} = 1$, for $\phi = 1, 2$, $m = 0, 1$. In terms of flow scaling, as shown in Fig. 5, functions $(1, 1)$ and $(1, 2)$ have scaling factors 1 and 4, respectively, *i.e.,* $\xi^{(1,1)} = 1$ and $\xi^{(1,2)} = 4$; and functions $(2, 1)$ and $(2, 2)$ have scaling factors 0.25 and 1, respectively, *i.e.,* $\xi^{(2,1)} = 0.25$ and $\xi^{(1,2)} = 1$. Note that function $(1, 2)$ is an expansion function, while function $(2, 1)$ is a compression function.

We consider a scenario in which each service is requested by 90 clients corresponding to all UE source-destination pairs.

D. Broadcast Approach v.s. Outage Approach

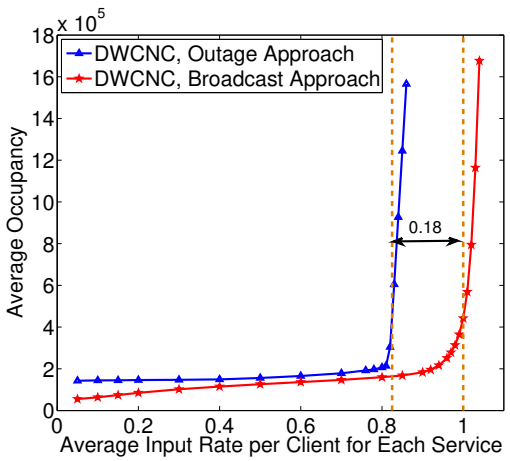
The throughput performance of DWCNC with the broadcast approach and with the outage approach is shown in Fig. 6(a), where we plot the time average occupancy (total queue backlog) as a function of the average (exogenous) input rate per client for each service, assuming the input rate to be the same for all clients and all services, while setting the control parameter V equal to 1500. Observe how the average occupancy exhibits a sharp increase when the exogenous input rate reaches approximately 0.82 for the outage approach and 1.00 for the broadcast approach. According to Theorem 2, and considering $\epsilon \to 0$, this sharp increase indicates that the average input rate has reached the boundary of the computing network capacity region, and hence it is indicative of the maximum achievable throughput. It can be seen from Fig. 6(a) that the maximum throughput using the broadcast approach is larger than that of using the the outage approach. This significant throughput difference is a clear indication of the enhanced transmission ability of the broadcast approach.

In the following, we assume an average input rate of 0.7, which is interior to the capacity region of DWCNC with both the outage approach and the broadcast approach (see Fig. 6(a)).

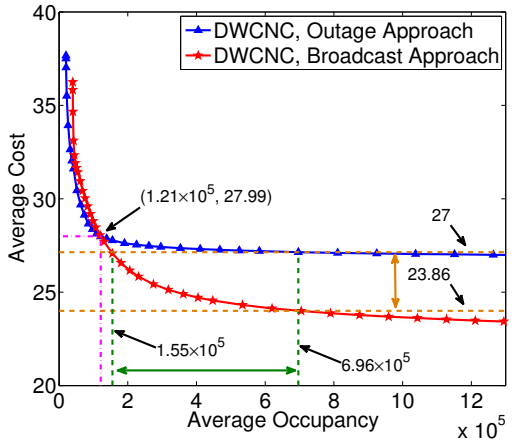
Fig. 6(b) shows the tradeoff between the average cost and the average occupancy as the control parameter V varies between 0 and 10^4, when running DWCNC with the broadcast approach and the outage approach. It can be seen from Fig. 6(b) that, with either coding scheme, the average cost decreases with the increase of the average occupancy. In

[4]With uniform direction of arrival (DOA) spectrum, the de-correlation distance for small scale fading is typically half wavelength and can be larger for limited angular spread [23]. But the moving distance of one wavelength is typically large enough to support the assumption of i.i.d. small scale fading across timeslots.

[5]The optimization of the power allocation among different code layers and the associated channel gain thresholds at each transmitting node is beyond the scope of this paper. In this simulation, power allocation and the channel gain threshold values that satisfy $\bar{R}_{i,1}^2 = \bar{R}_{i,1}^{\text{out}}$ may be suboptimal for throughput maximization, but we treat them as given parameters.



(a)



(b)

Fig. 6. Performance of DWCNC with the broadcast approach and the outage approach in the large scale scenario: a) Average occupancies evolving with varying average exogenous input rate: throughput optimality b) Average Cost vs. Average Occupancy

general, both evolutions follow the $[O(1/V), O(V)]$ cost-delay tradeoff of Theorem 2. However, the corresponding trade-off ratios are different. Note that the broadcast approach exhibits a significantly improved cost-delay tradeoff, in the sense that for a given target cost it can achieve a lower occupancy, and viceversa. For example, if we fix the average cost to 27, the outage approach requires an average occupancy of 6.96×10^5, while the broadcast approach can achieve the same average cost with an average occupancy of 1.55×10^5, leading to a factor of $4.49\times$ reduction in average delay. On the other hand, fixing the average occupancy to be *e.g.,* 6.96×10^5 requires an average cost of 27 with the outage approach, while the broadcast approach can reduce the average cost to 23.86 for the same average occupancy.

E. Processing Flow Distribution

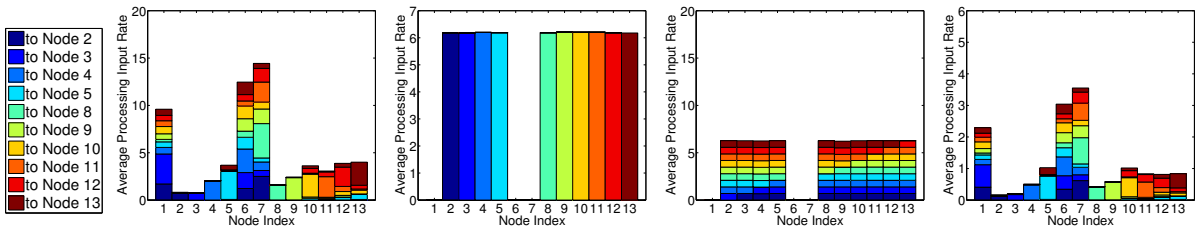
In this section, we simulate the average processing input rate distribution for the 4 functions and 90 clients across the computing network nodes under DWCNC with the outage approach and with the broadcast approach, respectively shown in Fig. 7 and Fig. 8. The average input rate for each client and each service is again equal to 0.7 and we set the control parameter V to 10^4.

Observe from Figs. 7(a) and 8(a) that the implementation of function $(1, 1)$ mostly concentrates at the APs (nodes $1, 6, 7$), motivated by the fact that the APs have cheaper processing resources than the UEs. Note, however, that part of the processing of function $(1, 1)$ still takes place at the UEs, even though the APs still have available processing capacity. This results from the fact that, for certain $s \to d$ pairs, there exist short paths connecting node s and d not passing through any AP, such that commodity $(d, 1, 0)$ steadily gets routed along these paths and gets processed at the corresponding UEs, instead of getting routed along longer paths that pass through APs. Comparing Fig. 7(a) and Fig. 8(a), it can be seen that the implementation of function $(1, 1)$ concentrates even more at the APs when using the broadcast approach. This is due to the enhanced transmission ability of the broadcast approach, which lowers the cost of taking longer paths passing through APs.

Figs. 7(b) and 8(b) show the average processing input rate distribution of function $(1, 2)$, which is an expansion function. As shown in Fig. 7(b), the processing of commodity $(d, 1, 1)$ concentrates at its destination node d when using the outage approach, which results from DWCNC trying to minimize the transmission cost impact of the expanded-size commodities generated by function $(1, 2)$. In contrast, as shown in Fig. 8(b), the processing of commodity $(d, 1, 1)$ with the broadcast approach for certain destinations, *e.g.,* $d = 2, 9, 10, 11$, is partly implemented at the APs, which is again motivated by the enhanced transmission ability of the broadcast approach to route commodities for cheaper processing at the APs.

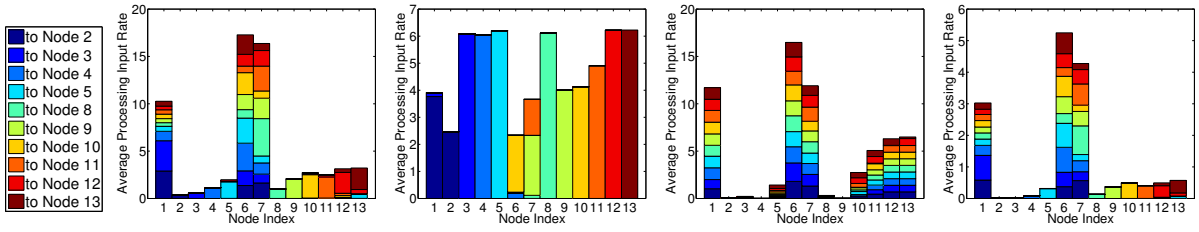
For Service 2, observe that the average processing input rate distribution of function $(2, 1)$ is quite different depending on the coding scheme used, as illustrated in Figs. 7(c) and 8(c). With the outage approach, Fig. 7(c) shows that function $(2, 1)$, a compression function, is implemented at all the UEs except the destination node d, and at the APs. This is because, for each client $s \to d$, implementing function $(2, 1)$ at the source node s reduces the transmission cost of service 2 by compressing the source commodity $(d, 2, 0)$ before entering the network. In contrast, as shown in Fig. 8(c), the implementation of function $(2, 1)$ using the broadcast approach mostly concentrates at the APs. This is once more due to the increased transmission efficiency of the broadcast approach, which allows to push the processing of commodity $(d, 2, 0)$ to the cheaper APs with a smaller penalty in the transmission cost required to route the uncompressed commodity. On the other hand, note that a portion of commodity $(d, 2, 0)$ is processed at nodes $10, 11, 12, 13$ due to the fact that the these UEs either have long distance wireless links to APs (node 10 and 11) or have no active wireless links to APs (node 12 and 13), such that local processing is more effficient.

The processing distribution of function $(2, 2)$, shown in Figs. 7(d) and 8(d), display a similar behavior as that of function $(1, 1)$. Note how the processing distribution concentrates more on the APs when adopting the broadcast approach, illustrating, once more, how its enhanced transmission efficiency allows a better utilization of the cheaper processing nodes.



(a) (b) (c) (d)

Fig. 7. Average processing input rate distribution of DWCNC with the outage approach. a) Service 1, Function 1; b) Service 1, Function 2; c) Service 2, Function 1; d) Service 2, Function 2.



(a) (b) (c) (d)

Fig. 8. Average processing input rate distribution of DWCNC with the broadcast approach. a) Service 1, Function 1; b) Service 1, Function 2; c) Service 2, Function 1; d) Service 2, Function 2.

VII. Conclusion

We considered the problem of optimal distribution of augmented information services over wireless computing networks. We characterized the capacity region of a wireless computing network and designed a dynamic wireless computing network control (DWCNC) algorithm that drives local transmissions-plus-processing flow scheduling and resource allocation decisions without knowledge of service demands or precise channel state information. DWCNC is shown to achieve arbitrarily close to minimum average network cost while subject to network delay increase with general trade off order $[O(1/V), O(V)]$. Our solution captures the unique chaining and flow scaling aspects of AgI services, while exploiting the use of the broadcast approach coding scheme for enhanced wireless transmission efficiency. Simulation results demonstrate the efficiency of DWCNC to route and process source information flows through the appropriate sequence of service functions hosted at wireless computing nodes, and the significant throughput and cost-delay tradeoff improvements obtained when using the broadcast approach coding scheme as opposed to the conventional outage approach.

Appendix A
Proof of Theorem 1:Necessity

A. Proof of Necessity

We prove that (10a)-(10h) are necessary for the stability of the wireless computing network, and that the minimum average cost can be achieved according to (11) and (12).

Recall that our policy space includes policies that use multi-copy routing, which allow multiple copies of the same information unit to travel through the network. We say that two information units are *equivalent* if the successful delivery of one of them to its destination does not require the delivery of the other to satisfy the service demand. Note that equivalent information units may be exact copies of each other, but may also be distinct units that have evolved via service processing from the same information unit.

Let us assume that when an information unit of final commodity (d, M) gets delivered to destination d, all other *equivalent* information units are immediately discarded from the network – an ideal assumption for traffic reduction of algorithms with multi-copy routing. We define $\mathcal{I}^{(d,m)}(t)$ as the set of information units of commodity (d, m) that, after going through the sequence of service functions $\{m+1, m+2, \ldots, M\}$, are delivered to destination d within the first t timeslots. Suppose there exists an algorithm that stabilizes the wireless computing network, possibly allowing multi-copy routing. Under this algorithm, define

- $I_i^{(d,m)}(t)$: the number of information units within $\mathcal{I}^{(d,m)}(t)$ that exogenously enter node i;
- $I_{i,\text{pr}}^{(d,m)}(t)$ and $I_{\text{pr},i}^{(d,m)}(t)$: the number of information units within $\mathcal{I}^{(d,m)}(t)$ that enter/exit the processing unit of node i;
- $I_{ij}^{(d,m)}(t)$: the number of times the information units within $\mathcal{I}^{(d,m)}(t)$ flow over link (i, j).

Since the algorithm stabilizes the network, we have, with probability 1,

$$\lim_{t\to\infty} \frac{\sum_{\tau=0}^{t} a_i^{(d,m)}(t)}{t} = \lim_{t\to\infty} \frac{I_i^{(d,m)}(t)}{t} = \lambda_i^{(d,m)}, \quad \forall i, (d, m). \tag{23}$$

Moreover, the total number of arrivals (both exogenous and endogenous) to node i of information units within $\mathcal{I}^{(d,m)}(t)$ must be equal to the number of departures from node i of information units within $\mathcal{I}^{(d,m)}(t)$. Therefore, we have, for $i \neq d$ or $m < M$,

$$\sum_{j:j\neq i} I_{ji}^{(d,m)}(t) + I_{\text{pr},i}^{(d,m)}(t) + I_i^{(d,m)}(t) = \sum_{j:j\neq i} I_{ij}^{(d,m)}(t) + I_{i,\text{pr}}^{(d,m)}(t), \tag{24}$$

and, for $m < M$ and for all i and d,

$$I_{\text{pr},i}^{(d,m+1)}(t) = \xi^{(d,m+1)} I_{i,\text{pr}}^{(d,m)}(t). \tag{25}$$

Define the following variables for transmission:

- $T(\mathbf{s}, t)$: the number of timeslots within the first t timeslots in which the network state is $\mathbf{s}$;
- $\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)$: the number of timeslots in the first t timeslots in which k transmission resource units are allocated at node i, while the previous network state is $\tilde{\mathbf{s}}$;
- $\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)$: the accumulated time (in possibly fractional amount of timeslots) during the first t timeslots used by node i to transmit information units within $\mathcal{I}^{(d,m)}(t)$, while k resource units are allocated for transmission, and the previous network state is $\tilde{\mathbf{s}}$;
- $\rho_{i,\mathbf{s}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)$: the accumulated time during the first t timeslots used by node i to transmit information units within $\mathcal{I}^{(d,m)}(t)$ when the network state is $\mathbf{s}$, while the previous network state is $\tilde{\mathbf{s}}$, and k resource units are allocated for transmission;
- $\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)$: the number of times during the first t timeslots that an information unit within $\mathcal{I}^{(d,m)}(t)$ is transmitted by node i with k transmission resource units allocated, and fall into the n-th partition, while the network state is $\mathbf{s}$ and the previous network state is $\tilde{\mathbf{s}}$;
- $\tilde{\eta}_{ij}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, n, t)$: the number of times during the first t timeslots that an information unit within $\mathcal{I}^{(d,m)}(t)$ transmitted by node i with k transmission resource units, is obtained by node j while belonging to the n-th partition, when the network state is $\mathbf{s}$, and the previous network state is $\tilde{\mathbf{s}}$.

The above definitions and the transmission constraints yield the following relations:

$$\frac{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)}{T_{\tilde{\mathbf{s}}}(t)} \geq 0, \quad \sum\nolimits_{k=0}^{K_i^{\text{tr}}} \frac{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)}{T_{\tilde{\mathbf{s}}}(t)} = 1, \qquad \forall i, \tilde{\mathbf{s}}, t, \tag{26}$$

$$\frac{\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)}{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)} \geq 0, \ \sum_{(d,m)} \frac{\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)}{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)} \leq 1, \ \forall i, k, \tilde{\mathbf{s}}, t, \tag{27}$$

$$\frac{\tilde{\eta}_{ij}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, n, t)}{\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)} \geq 0, \ \sum_{j \in \Omega_{i,n}} \frac{\tilde{\eta}_{ij}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, n, t)}{\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)} \leq 1, \quad \forall i, d, m, \tilde{\mathbf{s}}, \mathbf{s}, t, \tag{28}$$

where we define $0/0 = 1$ for any term on the denominator happen to be zero. For each link (i, j), each commodity (d, m), and all t, we then have

$$\frac{I_{ij}^{(d,m)}(t)}{t} = \sum_{\tilde{\mathbf{s}} \in \mathcal{S}} \frac{T_{\tilde{\mathbf{s}}}(t)}{t} \sum_{k=0}^{K_i^{\text{tr}}} \frac{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)}{T_{\tilde{\mathbf{s}}}(t)} \frac{\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)}{\tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}, t)} \times \sum_{\mathbf{s} \in \mathcal{S}} \frac{\rho_{i,\mathbf{s}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)}{\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)} \sum_{n=1}^{g_{i,\mathbf{s}}^{-1}(j)} \frac{\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)}{\rho_{i,\mathbf{s}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)} \frac{\tilde{\eta}_{ij}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, n, t)}{\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)}. \tag{29}$$

The network state process $\mathbf{S}(t)$ yields, for all $\mathbf{s} \in \mathcal{S}$,

$$\lim_{t \to \infty} \frac{T_{\mathbf{s}}(t)}{t} = \pi_{\mathbf{s}}, \text{ with prob. 1}, \tag{30}$$

and due to fact that $y_{i,k}^{\text{tr}}(\tau)$ is independent of $\mathbf{S}(\tau)$ given $\mathbf{S}(\tau - 1) = \tilde{\mathbf{s}}$, we also have, for all i, d, m,

$$\lim_{t \to \infty} \frac{\rho_{i,\mathbf{s}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)}{\tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)} = P_{\tilde{\mathbf{s}}\mathbf{s}}, \tag{31}$$

where $P_{\tilde{\mathbf{s}}\mathbf{s}} \triangleq \Pr(\mathbf{S}(t) = \mathbf{s} | \mathbf{S}(t-1) = \tilde{\mathbf{s}})$. In addition, the average rate of the n-th partition satisfies, for all $i, k, d, m, \tilde{\mathbf{s}}, \mathbf{s}, t$,

$$0 \leq \frac{\gamma_{i,n,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, t)}{\rho_{i,\mathbf{s}}^{(d,m)}(\tilde{\mathbf{s}}, k, t)} \leq R_{i,g_{i,n},k}(\mathbf{s}) - R_{i,g_{i,n-1},k}(\mathbf{s}), \tag{32}$$

Define the following variables for processing:

- $\alpha_{i,k}^{\text{pr}}(t)$: the number of timeslots during the first t timeslots in which node i allocates k processing resource units;
- $\beta_{i,\text{pr}}^{(d,m)}(k, t)$: the accumulated time used by node i to process information units within $\mathcal{I}^{(d,m)}(t)$, while k resource units are allocated for processing;
- $\gamma_{i,\text{pr}}^{(d,m)}(k, t)$: the number of information units within $\mathcal{I}^{(d,m)}(t)$ that are processed by node i with k processing resource units allocated during the first t timeslots.

Based on the above definitions and the processing constraints, we have the following relations:

$$\frac{\alpha_{i,k}^{\text{pr}}(t)}{t} \geq 0, \ \sum\nolimits_{k=0}^{K_i^{\text{pr}}} \frac{\alpha_{i,k}^{\text{pr}}(t)}{t} = 1, \quad \forall i, t, \tag{33}$$

$$\frac{\beta_{i,\text{pr}}^{(d,m)}(k, t)}{\alpha_{i,k}^{\text{pr}}(t)} \geq 0, \ \sum\nolimits_{(d,m)} \frac{\beta_{i,\text{pr}}^{(d,m)}(k, t)}{\alpha_{i,k}^{\text{pr}}(t)} \leq 1, \quad \forall i, k, t, \tag{34}$$

$$0 \leq \frac{\gamma_{i,\text{pr}}^{(d,m)}(k, t)}{\beta_{i,\text{pr}}^{(d,m)}(k, t)} \leq \frac{R_{i,k}}{r^{(d,m+1)}}, \qquad \forall i, t, k, d, m < M. \tag{35}$$

For each node i, we then have, for all i, (d, m) and t,

$$\frac{I_{i,\text{pr}}^{(d,m)}(t)}{t} = \sum\nolimits_{k=0}^{K_i^{\text{pr}}} \frac{\alpha_{i,k}^{\text{pr}}(t)}{t} \frac{\beta_{i,\text{pr}}^{(d,m)}(k, t)}{\alpha_{i,k}^{\text{pr}}(t)} \frac{\gamma_{i,\text{pr}}^{(d,m)}(k, t)}{\beta_{i,\text{pr}}^{(d,m)}(k, t)}. \tag{36}$$

Because the constraints in (26)-(28), (32), and (33)-(35) define bounded ratio sequences with finite dimensions, there exists an infinitely long subsequence of timeslots $\{t_u\}$ over which the time average cost achieves its $\liminf$ value $\underline{h}$, while the ratio terms converge, which are shown in (37) (see the bottom of the next page).

Define $f_{ij}^{(d,m)} \triangleq \lim_{t_u \to \infty} I_{ij}^{(d,m)}(t_u) \big/ t_u$. Then, it follows from (29) that

$$\begin{aligned} f_{ij}^{(d,m)} &\overset{(a)}{\leq} \sum\nolimits_{\tilde{\mathbf{s}} \in \mathcal{S}} \pi_{\tilde{\mathbf{s}}} \sum\nolimits_{k=0}^{K_i^{\text{tr}}} \tilde{\alpha}_{i,k}^{\text{tr}}(\tilde{\mathbf{s}}) \tilde{\beta}_{i,\text{tr}}^{(d,m)}(\tilde{\mathbf{s}}, k) \sum\nolimits_{\mathbf{s} \in S} P_{\tilde{\mathbf{s}}\mathbf{s}} \times \\ &\quad \sum\nolimits_{n=1}^{g_{i,\mathbf{s}}^{-1}(j)} \left[R_{i,g_{i,n},k}(\mathbf{s}) - R_{i,g_{i,n-1},k}(\mathbf{s}) \right] \tilde{\eta}_{ij}^{(d,m)}(\tilde{\mathbf{s}}, k, \mathbf{s}, n) \\ &\overset{(b)}{=} \sum\nolimits_{\mathbf{s} \in S} \pi_{\mathbf{s}} \sum\nolimits_{k=0}^{K_i^{\text{tr}}} \alpha_{i,k}^{\text{tr}}(\mathbf{s}) \beta_{i,\text{tr}}^{(d,m)}(\mathbf{s}, k) \times \\ &\quad \sum\nolimits_{n=1}^{g_{i,\mathbf{s}}^{-1}(j)} \left[R_{i,g_{i,n}}(\mathbf{s}) - R_{i,g_{i,n-1}}(\mathbf{s}) \right] \eta_{ij}^{(d,m)}(\mathbf{s}, k, n), \end{aligned}$$

where inequality (a) holds true due to the converging terms in (30) and (37) for transmission and the fact that $F_{ij}^{(d,m)}(k, \mathbf{s}) \leq R_{i,g_{i,n},k}(\mathbf{s}) - R_{i,g_{i,n-1},k}(\mathbf{s})$; equality (b) holds true due to the

fact that $\pi_{\mathbf{s}} = \sum_{\tilde{\mathbf{s}}\in\mathcal{S}} \pi_{\tilde{\mathbf{s}}} P_{\tilde{\mathbf{s}}\mathbf{s}}$ and the following definitions:

$$\alpha^{\text{tr}}_{i,k}(\mathbf{s}) \triangleq \sum\nolimits_{\tilde{\mathbf{s}}\in S} \frac{\pi_{\tilde{\mathbf{s}}} P_{\tilde{\mathbf{s}}\mathbf{s}}}{\pi_{\mathbf{s}}} \tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}}),$$

$$\beta^{(d,m)}_{i,\text{tr}}(\mathbf{s},k) \triangleq \sum\nolimits_{\tilde{\mathbf{s}}\in S} \frac{\pi_{\tilde{\mathbf{s}}} P_{\tilde{\mathbf{s}}\mathbf{s}} \tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}})}{\pi_{\mathbf{s}} \alpha^{\text{tr}}_{i,k}(\mathbf{s})} \tilde{\beta}^{(d,m)}_{i,\text{tr}}(\tilde{\mathbf{s}},k),$$

$$\eta^{(d,m)}_{ij}(\mathbf{s},k,n) \triangleq \sum_{\tilde{\mathbf{s}}\in S} \frac{\pi_{\tilde{\mathbf{s}}} P_{\tilde{\mathbf{s}}\mathbf{s}} \tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}}) \tilde{\beta}^{(d,m)}_{i,\text{tr}}(\tilde{\mathbf{s}},k)}{\pi_{\mathbf{s}} \alpha^{\text{tr}}_{i,k}(\mathbf{s}) \beta^{(d,m)}_{i,\text{tr}}(\mathbf{s},k)} \tilde{\eta}^{(d,m)}_{ij}(\tilde{\mathbf{s}},k,\mathbf{s},n).$$

In addition, define $f^{(d,m)}_{i,\text{pr}} \triangleq \lim_{t_u\to\infty} I^{(d,m)}_{i,\text{pr}}(t_u) \Big/ t_u$. With the converging terms for processing and the fact that $F^{(d,m)}_{i,\text{pr}}(k) \le R_{i,k}/\, r^{(m+1)}$, for $m < M$, it follows from (36) that

$$f^{(d,m)}_{i,\text{pr}} \le \sum\nolimits_{k=0}^{K^{\text{pr}}_i} \alpha^{\text{pr}}_{i,k} \beta^{(d,m)}_{i,tr}(k) \frac{R_{i,k}}{r^{(m+1)}}.$$

Moreover, the flow efficiency and non-negativity constraints follow: $f^{(d,M)}_{i,\text{pr}} = 0$, $f^{(d,0)}_{\text{pr},i} = 0$, $f^{(d,M)}_{dj} = 0$, $f^{(d,m)}_{i,\text{pr}} \ge 0$, $f^{(d,m)}_{ij} \ge 0$. Furthermore, dividing by t_u on both sides of (24) and (25), and letting $t_u \to \infty$, we have, for $i \ne d$ or $m < M$, with the result of (23),

$$\sum\nolimits_j f^{(d,m)}_{ji} + f^{(d,m)}_{\text{pr},i} + \lambda^{(d,m)}_i = \sum\nolimits_j f^{(d,m)}_{ij} + f^{(d,m)}_{i,\text{pr}},$$

and, for $m < M$ and all i and d, $f^{(m+1)}_{\text{pr},i} = \xi^{(m+1)} f^{(m)}_{i,\text{pr}}$.

Finally, the time average cost satisfies

$$\begin{aligned}\underline{h} &= \lim_{t_u\to\infty} \sum\nolimits_{i\in N} \Bigg[\sum\nolimits_{k=0}^{K^{\text{pr}}_i} \frac{\alpha^{\text{pr}}_{i,k}(t_u)}{t_u} w^{\text{pr}}_{i,k} + \\ &\qquad \sum\nolimits_{\tilde{\mathbf{s}}\in S} \frac{T_{\tilde{\mathbf{s}}}(t_u)}{t_u} \sum\nolimits_{k=0}^{K^{\text{tr}}_i} \frac{\tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}},t_u)}{T_{\tilde{\mathbf{s}}}(t_u)} w^{\text{tr}}_{i,k}\Bigg] \\ &\overset{(a)}{=} \sum\nolimits_{i\in N} \Bigg(\sum\nolimits_{k=0}^{K^{\text{pr}}_i} \alpha^{\text{pr}}_{i,k} w^{\text{pr}}_{i,k} + \sum\nolimits_{k=0}^{K^{\text{tr}}_i} w^{\text{tr}}_{i,k} \sum\nolimits_{\mathbf{s}\in S} \pi_{\mathbf{s}} \alpha^{\text{tr}}_{i,k}(\mathbf{s})\Bigg),\end{aligned}$$

where, for (a), we used the fact that $\sum_{\tilde{\mathbf{s}}\in\mathcal{S}} \pi_{\tilde{\mathbf{s}}} \tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}}) = \sum_{\mathbf{s}\in\mathcal{S}} \pi_{\mathbf{s}} \alpha^{\text{tr}}_{i,k}(\mathbf{s})$.

In summary, given $\{\lambda^{(d,m)}_i\} \in \Lambda$, this proves that there exists a set of flow variables and probability values that satisfy the constraints in Theorem 1. The minimum average cost $\overline{h}^*$ follows from taking the minimum of $\underline{h}$ over all the variable sets that stabilize the network.

B. Proof of Sufficiency

Given exogenous input rate matrix $\{\lambda^{(d,m)}_i + \epsilon\}$, $\epsilon > 0$, probability values $\alpha^{\text{tr}}_{i,k}(\mathbf{s})$, $\beta^{(d,m)}_{i,\text{tr}}(\mathbf{s},k)$, $\eta^{(d,m)}_{ij}(\mathbf{s},k,n)$, $\alpha^{\text{pr}}_{i,k}$, $\beta^{(d,m)}_{i,\text{pr}}(k)$, and multi-commodity flow variables $f^{(d,m)}_{ij}$, $f^{(d,m)}_{i,\text{pr}}$, $f^{(d,m)}_{\text{pr},i}$ satisfying (10)-(12), we construct a stationary randomized policy using single-copy routing such that:

$$\mathbb{E}\left\{\mu^{(d,m)}_{ij}(t)\right\} = f^{(d,m)}_{ij}, \tag{38a}$$

$$\mathbb{E}\left\{\mu^{(d,m)}_{i,\text{pr}}(t)\right\} = f^{(d,m)}_{i,\text{pr}},\ \mathbb{E}\left\{\mu^{(d,m)}_{\text{pr},i}(t)\right\} = f^{(d,m)}_{\text{pr},i}, \tag{38b}$$

where $\mu^{(d,m)}_{ij}(t)$, $\mu^{(d,m)}_{i,\text{pr}}(t)$, and $\mu^{(d,m)}_{\text{pr},i}(t)$ respectively denote the flow rates assigned by the stationary randomized policy for transmission and processing. Plugging $\{\lambda^{(d,m)}_i + \epsilon\}$ and (38) into (10a), after algebraic manipulations, we have

$$\begin{aligned}\mathbb{E}\Big\{\sum\nolimits_j \mu^{(d,m)}_{ij}(t) + \mu^{(d,m)}_{i,\text{pr}}(t) - \\ \sum\nolimits_j \mu^{(d,m)}_{ji}(t) - \mu^{(d,m)}_{\text{pr},i}(t)\Big\} \ge \lambda^{(d,m)}_i + \epsilon.\end{aligned} \tag{39}$$

By applying the standard LDP analysis [14], $\{\lambda^{(d,m)}_i\}$ is proven to be interior to Λ.

APPENDIX B
PROOF OF THEOREM 2

Following from (14), we first prove the following lemma:

Lemma B.1. *Among the algorithms using single-copy routing, the DWCNC algorithm, in each timeslot t, maximizes $\mathbb{E}\{Z^{\text{tr}}_i(t) - V h^{\text{tr}}_i(t) \,|\, \mathcal{H}(t)\}$ subject to (5)-(6) and $\mathbb{E}\{Z^{\text{pr}}_i(t) - V h^{\text{pr}}_i(t) \,\big|\, \mathcal{H}(t)\}$ subject to (1)-(2).* □

Proof. See Appendix C. □

Lemma B.1 implies that the right hand side of (14) is minimized by DWCNC, and is therefore no larger than the corresponding expression under the optimal stationary randomized policy (characterized in Theorem 1) that supports $(\boldsymbol{\lambda} + \epsilon\mathbf{1}) \in \Lambda$ and achieves average cost $\overline{h}^*(\boldsymbol{\lambda} + \epsilon\mathbf{1})$:

$$\begin{aligned}&\Delta(\mathcal{H}(t)) + V\mathbb{E}\{h(t) \,|\, \mathcal{H}(t)\} \le NB + \boldsymbol{\lambda}'\mathbf{Q}(t) + \\ &\qquad \sum\nolimits_i \mathbb{E}\big\{Z^{*\text{pr}}_i(t) - V h^{*\text{pr}}_i(t) + Z^{*\text{tr}}_i(t) - V h^{*\text{tr}}_i(t)\big\} \\ &\le NB + V\overline{h}^*(\boldsymbol{\lambda} + \epsilon\mathbf{1}) - \epsilon \sum\nolimits_i \sum\nolimits_{(d,m)} Q^{(d,m)}_i(t).\end{aligned} \tag{40}$$

Finally, we can use the theoretical result in [24] for the proof of network stability and average cost convergence with probability 1. Note that the following bounding conditions are satisfied in the network system:

1) The second moment of $\mathbb{E}\{(h(t))^2\}$ is upper bounded by $[\sum_i (w^{\text{tr}}_{i,K^{\text{tr}}_i} + w^{\text{pr}}_{i,K^{\text{pr}}_i})]^2$ and therefore satisfies $\sum_{\tau=0}^{\infty} \mathbb{E}\{(h(t))^2\}/\,\tau^2 < \infty$.
2) We have $\mathbb{E}\{h(t) \,|\, \mathcal{H}(t)\}$ lower bounded for all $\mathcal{H}(t)$ and t: $\mathbb{E}\{h(t) \,|\, \mathcal{H}(t)\} \ge 0$.
3) The conditional fourth moment of backlog change is upper bounded for all t, i and (d,m):

$$\begin{aligned}&\mathbb{E}\left\{\left(Q^{(d,m)}_i(t+1) - Q^{(d,m)}_i(t)\right)^4 \,\middle|\, \mathcal{H}(t)\right\} \le \\ &\max_i \left\{\left[\max_{\mathbf{s}\in S}\left\{\sum_{j:j\ne i} R_{ji,K^{\text{u}}_j}(\mathbf{s})\right\} + \frac{\xi_{\max} R_{i,K^{\text{pr}}_i}}{r_{\min}} + A_{\max}\right]^4\right\}.\end{aligned}$$

$$\lim_{t_u\to\infty} \frac{1}{t_u} \sum_{\tau=0}^{t_u-1} h(\tau) = \underline{h},\ \lim_{t_u\to\infty} \frac{\tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}},t_u)}{T_{\tilde{\mathbf{s}}}(t_u)} = \tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}}),\ \lim_{t_u\to\infty} \frac{\tilde{\beta}^{(d,m)}_{i,\text{tr}}(\tilde{\mathbf{s}},k,t_u)}{\tilde{\alpha}^{\text{tr}}_{i,k}(\tilde{\mathbf{s}},t_u)} = \tilde{\beta}^{(d,m)}_{i,\text{tr}}(\tilde{\mathbf{s}},k),\ \lim_{t_u\to\infty} \frac{\tilde{\eta}^{(d,\phi,m)}_{ij}(\tilde{\mathbf{s}},k,\mathbf{s},n,t_u)}{\gamma^{(d,m)}_{i,n,\text{tr}}(\tilde{\mathbf{s}},k,\mathbf{s},t_u)} = \tilde{\eta}^{(d,\phi,m)}_{ij}(\tilde{\mathbf{s}},k,\mathbf{s},n),$$

$$\lim_{t_u\to\infty} \frac{\gamma^{(d,m)}_{i,n,\text{tr}}(\tilde{\mathbf{s}},k,\mathbf{s},t_u)}{\rho^{(d,m)}_{i,\mathbf{s}}(\tilde{\mathbf{s}},k,t_u)} = F^{(d,m)}_{ij}(k,\mathbf{s}),\ \lim_{t_u\to\infty} \frac{\alpha^{\text{pr}}_{i,k}(t_u)}{t} = \alpha^{\text{pr}}_{i,k},\ \lim_{t_u\to\infty} \frac{\beta^{(d,m)}_{i,\text{pr}}(k,t_u)}{\alpha^{\text{pr}}_{i,k}(t_u)} = \beta^{(d,m)}_{i,\text{pr}}(k),\ \lim_{t_u\to\infty} \frac{\gamma^{(d,m)}_{i,\text{pr}}(k,t_u)}{\beta^{(d,m)}_{i,\text{pr}}(k,t_0)} = F^{(d,m)}_{i,\text{pr}}(k). \tag{37}$$

With the above three conditions satisfied, based on the derivations in [24], Eq. (40) under DWCNC leads to the network stability (21) and average cost (22) bound, with probability 1.

APPENDIX C
PROOF OF LEMMA B.1

Regarding the processing decisions, since the *computing channel* is always known, maximizing $\mathbb{E}\{ Z_i^{\text{pr}}(t) - V h_i^{\text{pr}}(t) \big| \mathcal{H}(t)\}$ is equivalent to maximizing $Z_i^{\text{pr}}(t) - V h_i^{\text{pr}}(t)$. And the maximization of $Z_i^{\text{pr}}(t) - V h_i^{\text{pr}}(t)$ subject to (1)-(2) can be achieved by the choice of commodity (d,m), resource allocation k, and flow rate $\mu_{i,\text{pr}}^{(d,m)}(t)$ described by the local processing decisions of DWCNC in Sec. IV-A according to a straightforward *max-weight matching*.

For the transmission decisions, by plugging (5) in, we can express $Z_i^{\text{tr}}(t)$ (see (14)) as follows:

$$Z_i^{\text{tr}}(t) = \sum_{(d,m)} \sum_{n=1}^{N-1} \sum_{u=n}^{N-1} \mu_{iq_{i,u},n}^{(d,m)}(t) \left[Q_i^{(d,m)}(t) - Q_{q_{i,u}}^{(d,m)}(t) \right]. \tag{41}$$

Let $\chi_{i,\text{tr}}^{(d,m)}(t)$ be the fraction of the transmission time allocated to the transmission of commodity (d,m) in timeslot t, and let $\eta_{ij,n}^{(d,m)}(t)$ be the fraction of the transmitted commodity (d,m) in the n-th partition that is retained by node j, with $n \leq q_{i,\mathbf{S}(t)}^{-1}(j)$. Then, assuming single-copy routing, Eq. (6) yields

$$\mu_{iq_{i,u},n}^{(d,m)}(t) = \chi_{i,\text{tr}}^{(d,m)}(t) \eta_{iq_{i,u},n}^{(d,m)}(t) \times \left[R_{iq_{i,n},k}(\mathbf{S}(t)) - R_{iq_{i,n-1},k}(\mathbf{S}(t)) \right], \quad \forall i,t,(d,m) \tag{42}$$

$$\sum\nolimits_{(d,m)} \chi_{i,\text{tr}}^{(d,m)}(t) \leq 1, \quad \forall i,t, \tag{43}$$

$$\sum\nolimits_{j} \eta_{ij,n}^{(d,m)}(t) \leq 1, \quad \forall i,t,(d,m). \tag{44}$$

Plugging (42) into (41) and taking the expectation conditioned on $\mathcal{H}(t)$ and $\{y_{i,k}^{\text{tr}}(t) = 1\}$, it follows that

$$\begin{aligned}
&\mathbb{E}\left\{ Z_i^{\text{tr}}(t) \big| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \right\} \\
&\overset{(a)}{\leq} \sum_{(d,m)} \sum_{n=1}^{N-1} \mathbb{E}\Bigg\{ \chi_{i,\text{tr}}^{(d,m)}(t) \left[R_{iq_{i,n},k}(\mathbf{S}(t)) - R_{iq_{i,n-1},k}(\mathbf{S}(t)) \right] \\
&\qquad \times \sum_{u=n}^{N-1} \eta_{iq_{i,u},n}^{(d,m)}(t) W_{ig_{i,u}}^{(d,m)}(t) \Bigg| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \Bigg\} \\
&\overset{(b)}{\leq} \sum_{(d,m)} \sum_{n=1}^{N-1} \mathbb{E}\Bigg\{ \chi_{i,\text{tr}}^{(d,m)}(t) \left[R_{iq_{i,n},k}(\mathbf{S}(t)) - R_{iq_{i,n-1},k}(\mathbf{S}(t)) \right] \\
&\qquad \times \max_{j \in \Omega_{i,n}(\mathbf{S}(t))} \left\{ W_{ij}^{(d,m)}(t) \right\} \Bigg| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \Bigg\} \\
&\overset{(c)}{=} \sum_{(d,m)} \mathbb{E}\left\{ \chi_{i,\text{tr}}^{(d,m)}(t) \Big| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \right\} \times \\
&\qquad \sum_{n=1}^{N-1} \mathbb{E}\left\{ \max_{j \in \Omega_{i,n}(\mathbf{S}(t))} \left\{ W_{ij}^{(d,m)}(t) \right\} \times \right. \\
&\qquad \left. \left[R_{iq_{i,n},k}(\mathbf{S}(t)) - R_{iq_{i,n-1},k}(\mathbf{S}(t)) \right] \Big| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \right\} \\
&\overset{(d)}{\leq} \max_{(d,m)} \Bigg\{ \sum_{n=1}^{N-1} \mathbb{E}\Big\{ \left[R_{iq_{i,n},k}(\mathbf{S}(t)) - R_{iq_{i,n-1},k}(\mathbf{S}(t)) \right] \times \\
&\qquad \max_{j \in \Omega_{i,n}(\mathbf{S}(t))} \left\{ W_{ij}^{(d,m)}(t) \right\} \Big| \mathcal{H}(t), y_{i,k}^{\text{tr}}(t) = 1 \Big\} \Bigg\} \\
&\overset{(e)}{=} \max_{(d,m)} \left\{ W_{i,k,\text{tr}}^{(d,m)}(t) \right\}.
\end{aligned} \tag{45}$$

In (45), inequality (a) follows from the definition of $W_{ij}^{(d,\phi,m)}(t)$; inequality (b) follows from (44); equality (c) holds because, given $\mathcal{H}(t)$ and $\{y_{i,k}^{\text{tr}}(t) = 1\}$, the values of $R_{iq_{i,n},k}(\mathbf{S}(t))$ and $\max_{j \in \Omega_{i,n}(\mathbf{S}(t))}\{W_{ij}^{(d,m)}(t)\}$ are determined by $\mathbf{S}(t)$ and therefore are independent from $\chi_{i,\text{tr}}^{(d,m)}(t)$; inequality (d) follows from (43); equality (e) follows from the definition of $W_{i,k,\text{tr}}^{(d,m)}(t)$ in (15).

Finally, taking expectation over $y_{i,k}^{\text{tr}}(t)$ on (45), we obtain

$$\begin{aligned}
&\mathbb{E}\left\{ Z_i^{\text{tr}}(t) - V h_i^{\text{tr}}(t) \big| \mathcal{H}(t) \right\} \\
&\leq \sum_{k=0}^{K_i^{\text{tr}}} \left[\max\nolimits_{(d,m)} \left\{ W_{i,k,\text{tr}}^{(d,m)}(t) \right\} - V w_{i,k}^{\text{tr}} \right] \Pr\left\{ y_{i,k}^{\text{tr}}(t) = 1 \right\} \\
&\overset{(f)}{\leq} \max_{k,(d,m)} \left\{ W_{i,k,\text{tr}}^{(d,m)}(t) - V w_{i,k}^{\text{tr}} \right\},
\end{aligned} \tag{46}$$

where (f) follows due to the fact $\sum_{k=0}^{K_i^{\text{tr}}} \Pr\left\{ y_{i,k}^{\text{tr}}(t) = 1 \right\} = 1$.

In (45) and (46), the upper bounds (a) and (b) can be achieved by step 4 of DWCNC local transmission decisions; the upper bound (d), (e), and (f) can be achieved by step 2 and 3 of DWCNC local transmission decisions. This concludes the proof of Lemma B.1.

REFERENCES

[1] H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch, "On the delivery of augmented information services over wireless computing networks," in *Proc. IEEE ICC*, Jun. 2017, pp. 1-7.

[2] M. Weldon, "The future X network," *CRC Press*, Oct. 2015.

[3] M. Barcelo, J. Llorca, A. M. Tulino, and N. Raman, "The cloud servide distribution problem in distributed cloud networks," in *Proc. IEEE ICC*, Sep. 2015, pp. 344-350.

[4] M. F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutaba, "On orchestrating network function chains in NFV," in *Proc. IEEE/ACM CNSM*, Nov. 2015, pp. 50-56.

[5] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Proc. IFIP/IEEE IM*, May 2015, pp. 98-106.

[6] M. Barcelo, A. Correa, J. Llorca, A. M. Tulino, J. Vicario, A. Morell, "IoT-Cloud Service Optimization in Next Generation Smart Environments," in *IEEE JSAC*, Dec. 2016, pp. 4077-4090.

[7] L. Lewin-Eytan, J. Naor, R. Cohen, and D. Raz, "Near optimal placement of virtual network functions," in *Proc. IEEE/ACM INFOCOM*, Apr. 2015, pp. 1346-1354.

[8] H. Feng, J. Llorca, A. M. Tulino, D. Raz, and A. F. Molisch, "Approximation algorithms for the network service distribution problem," in *Proc. IEEE/ACM INFOCOM*, May 2017, pp. 1-9.

[9] J. Kuo, S. Shen, H. Kang, D. Yang, M. Tsai, and W. Chen, "Service chain embedding with maximum flow in software defined network and application to the next-generation cellular network architecture", in *Proc. IEEE/ACM INFOCOM*, May 2017, pp. 1-9.

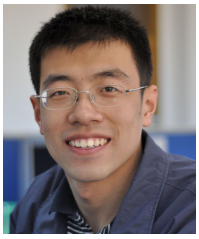
[10] H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch, "Dynamic Network service optimization in distributed cloud networks," in *Proc. IEEE/ACM INFOCOM SWFAN Workshop*, Sep. 2016, pp. 300-305.

[11] H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch, "Optimal dynamic cloud network control," in *Proc. IEEE ICC*, Sep. 2016, pp. 1-7.

[12] M. Chiang and T. Zhang, "Fog and IoT: an overview of research opportunities", *IEEE Internet of Things J.*, vol. 3, no. 6, pp. 854-864, Dec. 2016.

[13] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing", *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14-23, Oct.-Dec. 2009.

[14] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems", *Synthesis Lectures on Communication Networks*, Morgan & Claypool, 2010.
[15] M. J. Neely, "Optimal backpressure routing for wireless networks with multi-receiver diversity", *Ad Hoc Netw.*, vol. 7, pp. 862–881, Jul. 2009.
[16] H. Feng and A. F. Molisch, "Diversity backpressure scheduling and routing with mutual information accumulation in Wireless Ad-hoc Networks", *IEEE Trans. on Inf. Theory*, vol. 62, no. 12: pp. 7299-7323, Dec. 2016.
[17] S. Shamai and A. Steiner. "A broadcast approach for a single-user slowly fading MIMO channel." *IEEE Trans. on Inf. Theory*, vol. 49, no. 10, pp. 2617-2635, Oct. 2003.
[18] A. M. Tulino, G. Caire, and S. Shamai. "The broadcast approach for the sparse-input random-sampled MIMO gaussian channel," in *Proc. IEEE ISIT*, Jun. 2014, pp. 621-625.
[19] E. N. ISG, "Network functions virtulisation (NFV); architectural framwork," *ETSI GS NFV 002 v 2.1.1*, Mar. 2016.
[20] W. Haeffner, J. Napper, M. Stiemerling, D. Lopez, and J. Uttaro, "Service function chaining use cases in mobile networks," *Service Function Chaining*, Internet-Draft, Oct. 2016. [Online] Available: https://tools.ietf.org/html/draft-ietf-sfc-use-case-mobility-07.
[21] A. El Gamal and Y.-H. Kim, "Network information theory," *Cambridge University Press*, 2011.
[22] S. S. Szyszkowicz, H. Yanikomeroglu, and J. S. Thompson, "On the feasibility of wireless shadowing correlation models," *IEEE Trans. on Veh. Technol.*, vol. 59, no. 9, pp. 4222-4236, 2010.
[23] A. F. Molisch, "Wireless communications," 2nd ed., *John Wiley & Sons*, 2012.
[24] M. J. Neely, "Queue stability and probability 1 convergence via lyapunov optimization," *arXiv preprint:1008.3519*, 2010.
[25] V. Kristem, N. Mehta, A. F. Molisch, "Optimal receive antenna selection in time-varying fading channels with practical training constraints," *IEEE Trans. on Commun.*, vol. 58, no. 7, pp. 2023-2034, 2010.
[26] Z. Li, R. Wang, A. F. Molisch. "Shadowing in urban environments with microcellular or peer-to-peer links," in *Proc. IEEE EUCAP*, 2012, pp. 44-48.



Hao Feng (S'08–M'18) respectively received the B.S. Degree and M.S. Degree in 2006 and 2008 from Department of Information Science & Electronic Engineering at Zhejiang University, Hangzhou, P. R. China. In 2017, he received the Ph.D. degree from Department of Electrical Engineering at University of Southern California, Los Angeles, CA, USA. He joined Intel Labs, Hillsboro, OR, USA, in 2018, where he is currently a Research Scientist. His research interests include stochastic network optimization, cross-layer optimization, and cooperative communication with their applications in cloud networks, wireless computing networks, information centric networks, and wireless ad-hoc & V2X networks. He received the best paper award in IEEE ICC 2016 conference, and was awarded Annenberg Graduate Fellowship from 2009-2013 for his Ph.D. study.

Jaime Llorca (S'03-M'09) received the B.E. degree in Electrical Engineering from Universidad Politecnica de Catalunya, Barcelona, Spain, in 2001, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from University of Maryland, College Park, MD, USA, in 2003 and 2008, respectively. He held a post-doctoral position at the Center for Networking of Infrastructure Sensors (CNIS), College Park, MD, USA, from 2008 to 2010. He joined Bell Labs at Holmdel, NJ, USA, in 2010, where he is currently a Senior Research Scientist with the Mathematics and Algorithms Group. His research interests are in the field of network algorithms, network optimization, distributed control, and network information theory, with applications to next generation communication networks, distributed cloud networking, and content distribution. He currently serves as an Associate Editor for the IEEE Transactions on Networking. He is a recipient of the 2007 Best Paper Award at the IEEE International Conference on Sensors, Sensor Networks and Information Processing (ISSNIP), the 2016 Best Paper Award at the IEEE International Conference on Communications (ICC), and the 2015 Jimmy H.C. Lin Award for Innovation.

Antonia M. Tulino (F'13) received the Ph.D. degree in Electrical Engineering from Seconda Università degli Studi di Napoli, Italy, in 1999. She held research positions at Princeton University, at the Center for Wireless Communications, Oulu, Finland and at Università degli Studi del Sannio, Benevento, Italy. From 2002 until 2016, she has been Associate Processor at the Universityà degli Studi di Napoli "Federico II". Since 2017, she is Full Professor at the Università degli Studi di Napoli "Federico II" and since 2009 she collaborates with Bell Labs. From 2011 until 2013, Dr. Tulino has been Member of the Editorial Board of the IEEE Transactions on Information Theory and in 2013, she was elevated to IEEE Fellow. She has received several paper awards and among the others the 2009 Stephen O. Rice Prize in the Field of Communications Theory for the best paper published in the IEEE TRANSACTION ON COMMUNICATION in 2008. She has been principal investigator of several research projects sponsored by the European Union and the Italian National Council, and was selected by the National Academy of Engineering for the Frontiers of Engineering program in 2013. Prof. Tulino has been recipient of the 2018-2019 UC3M-Santander Chair of Excellence. Her research interests lay in the area of communication systems approached with the complementary tools provided by signal processing, information theory and random matrix theory.

Andreas F. Molisch (S'89-M'95-SM'00-F'05) received the Dipl. Ing., Ph.D., and habilitation degrees from the Technical University of Vienna, Vienna, Austria, in 1990, 1994, and 1999, respectively. He subsequently was with AT&T (Bell) Laboratories Research (USA); Lund University, Lund, Sweden, and Mitsubishi Electric Research Labs (USA). He is now a Professor and Solomon-Golomb – Andrew-and-Erna-Viterbi Chair at the University of Southern California, Los Angeles.

His current research interests are the measurement and modeling of mobile radio channels, multi-antenna systems, ultra-wideband communications and localization, novel modulation and multiple access systems, and wireless video distribution. He has authored, coauthored, or edited four books (among them the textbook Wireless Communications, Wiley-IEEE Press), 19 book chapters, more than 220 journal papers, 300 conference papers, as well as more than 80 patents and 70 standards contributions.

Dr. Molisch has been an Editor of a number of journals and special issues, General Chair, Technical Program Committee Chair, or Symposium Chair of multiple international conferences, as well as Chairman of various international standardization groups. He is a Fellow of the National Academy of Inventors, Fellow of the AAAS, Fellow of the IEEE, Fellow of the IET, an IEEE Distinguished Lecturer, and a member of the Austrian Academy of Sciences. He has received numerous awards, among them the Donald Fink Prize of the IEEE, the IET Achievement Medal, and the Eric Sumner Award of the IEEE.