Towards a System Theoretic Approach to Wireless Network Capacity in Finite Time and Space

Florin Ciucu
University of Warwick

Ramin Khalili
T-Labs / TU Berlin

Yuming Jiang
Norwegian University of Science and Technology

Liu Yang, Yong Cui
Tsinghua University

***Abstract*—In asymptotic regimes, both in time and space (network size), the derivation of network capacity results is grossly simplified by brushing aside queueing behavior in non-Jackson networks. This simplifying double-limit model, however, lends itself to conservative numerical results in finite regimes. To properly account for queueing behavior beyond a simple calculus based on average rates, we advocate a system theoretic methodology for the capacity problem in finite time and space regimes. This methodology also accounts for spatial correlations arising in networks with CSMA/CA scheduling and it delivers rigorous closed-form capacity results in terms of probability distributions. Unlike numerous existing asymptotic results, subject to anecdotal practical concerns, our transient results can be used in practical settings, e.g., to compute the time scales at which multi-hop routing is more advantageous than single-hop routing.**

I. Introduction

The fields of communication networks and information theory have been for long evolving in isolation of each other, in what is referred to as an unconsummated union (Ephremides and Hajek [12]). This is partly due to the fact that unlike communication networks which properly account for data burstiness and delay, information theory typically assumes saturated data sources and is practically oblivious to when data is received, say at the receiver of a point-to-point channel.

A groundbreaking work at the intersection of the two fields is a set of results obtained by Gupta and Kumar [15]. Under some simplifications at the network layers (e.g., no multi-user coding schemes, or ideal assumptions on power-control, routing, and scheduling), the authors derived network capacity results as asymptotic scaling laws on the maximal data rates which can be reliably sustained in multi-hop wireless networks. The elegance and importance of these results have been very inspirational, especially within the networking community.

The results from [15], and of most related work, rely on a *double-limit model*. The outer limit is explicitly taken in the number of nodes n— capturing an *infinite-space* model—in order to guarantee certain structural properties in random networks with high probability. The inner limit is implicitly taken in time—capturing an *infinite-time* model—and which enables a simple calculus based on average rates to derive upper and lower bounds on network capacity. The double-limit model can be regarded as being reminiscent of information theory and relating itself to the *infinite-space* model employed in the analysis of the multiaccess channel (i.e., infinitely many sources are assumed to coexist, Gallager [14]).

The key advantage of the technical arguments from [15] is that all nodes appear as smoothed-out at the data link layer and the network capacity analysis is drastically simplified; indeed, by solely reasoning in terms of average rates (first moments), the difficult problem of accounting for burstiness (e.g., higher moments) in non-Jackson queueing networks is avoided. While such a calculus is mathematically justified in asymptotic regimes, its implications in finite regimes have been largely evaded so far; by 'finite regime' we mean both finite time and finite number of nodes.

To shed light in the direction of computing the network capacity in finite regimes, this paper makes three contributions:

- C1. It scrutinizes the double-limit argument from [15]. Concretely, it is shown that the direct reproduction of asymptotic techniques in finite regimes does not capture a non-negligible factor for both the upper and lower capacity bounds, which (partially) justifies the anecdotal impracticality of numerous asymptotic results. These findings motivate the need for alternative analytical techniques to compute the network capacity in finite time and space regimes, beyond the conveniently simplistic average-based calculus.
- C2. It advocates a *system theoretic methodology* to the *transient* network capacity problem at the per-flow level. The crucial advantage of this approach is that it conveniently deals with inherent burstiness and queueing behavior at downstream nodes. Moreover, it also copes with spatial correlations arising in networks with CSMA/CA scheduling, in the sense that no artificial assumptions (e.g., statistical independence) are necessary.
- C3. It illustrates the applicability of finite time and space capacity results to decide when multi-hop routing is theoretically more advantageous than single-hop routing. In particular, our results lend themselves to the *time scale* at which the lower bound (on throughput capacity) in the case of multi-hop routing is greater than the upper bound in the case of single-hop routing.

From a technical point of view, the main idea of the advocated system theoretic methodology lies on a subtle analogy between single-hop links and linear time invariant (LTI) systems, by constructing impulse-responses to entirely characterize successful transmissions over single-hop links. The impulse-responses are closed under a convolution operator, which conveniently accounts for queueing behavior at downstream nodes. These ideas have been recently explored by Ciucu *et al.* [5], [4], [8] for the particular Aloha protocol. This

paper *generalizes* these prior works by formulating a *unified* system-theoretic framework which additionally captures two more MAC protocols: centralized scheduling and especially the challenging CSMA/CA.

An advantage of the proposed framework is that it yields capacity results in terms of probability distributions, and thus all the moments, including average rates or variances, are readily available. Moreover, the capacity results are directly obtained at the *per-flow* level. Such a per-flow analysis can provide information about the fairness of routing and scheduling algorithms and hence could be useful in protocol design. As multiple paths are available between source-destination pair, one can use this information to provide route optimization and load balancing in the network. The concrete practical application addressed in the paper was described in Item C3.

The rest of the paper is organized as follows. In Section II we discuss the limitations of the technical arguments from [15] based on a double-limit model in finite time and space regimes. In Section III we introduce the advocated system theoretic methodology to derive capacity results in finite regimes. In Section IV we show how to fit three MAC protocols in this methodology. Section V presents the multi-hop vs. single-hop practical application and Section VI concludes the paper.

II. On the Limitations of The Double-Limit Model

Consider the random network model from Gupta and Kumar [15] in which n nodes are uniformly placed on a disk of area one. For each node in the network, a random destination is chosen such that there are n source-destination pairs. We consider the Protocol Model from [15], which defines successful transmission in terms of Euclidean distances.

The capacity problem concerns the maximum value of $\lambda(n)$, i.e., the rate of each transmission, guaranteeing network stability in terms of bounded buffers. Computing upper and lower bounds on $\lambda(n)$ is based on a simple calculus involving the end-to-end (e2e) transmissions' average rates at the relay nodes, which are implicitly subject to a time limit. For an e2e transmission i, let $\tilde{\lambda}_{i,j}(n)$ denote the incoming average rate at node j, i.e.,

$$\tilde{\lambda}_{i,j}(n) = \limsup_{t\to\infty} \frac{A_{i,j}(t)}{t} \ ,$$

where $A_{i,j}(t)$ denotes the cumulative arrival process and t denotes time. In general it holds that $\tilde{\lambda}_{i,j}(n) \leq \lambda(n)$, whereas an exact relationship depends on many factors such as routing, scheduling, or the network stability; such factors may also lend themselves to conceivable scenarios in which the '$\lim$' does not exist, whence the '$\limsup$' definition above.

With the new notation, one can rephrase the network capacity problem as finding the maximal rate $\lambda(n)$ such that

$$\lambda(n) = \tilde{\lambda}_{i,j}(n) \ \forall i,j \ .$$

The main result from [15] is that

$$\lambda(n) = \Theta\left(\frac{1}{\sqrt{n\log n}}\right) \ . \tag{1}$$

Here, the underlying space limit in n guarantees useful structural properties in the considered random network with high probability (e.g., e2e connectivity or bounds on the number of transit transmissions at some node).

Capacity results such as the one from Eq. (1) are based on a double-limit model, in which the outer (space) limit is in n whereas the inner (time) limit is in t. En passant, it is interesting to observe that the limits are not interchangeable; indeed, note that by letting the outer limit in t, the rates at downstream relay nodes tend to zero (e.g., when $n > t$). More interestingly, a single-limit model can be considered by suitably letting t as a function of n. Depending on structural network properties, the rate at which t should increase could be as large as

$$t = \omega\left(n^2\right) \ .$$

This is necessary, for instance, in the following scenario: n nodes numbered as $\{1, 2, \ldots, n\}$ are placed around a circle, every node i transmits to the counter-clockwise neighbor $(i+n-2)\%n+1$ along the clockwise path $i\%n+1, (i+1)\%n+1, \ldots, (i+n-2)\%n+1$, and all transmissions interfere with each other ('%' is the modulo operation). Under a perfect scheduling, we remark that at most k delivered packets from all n e2e transmissions could be guaranteed in kn^2 slots, for any k, whence the $\omega(n^2)$ lower bound.

Next we discuss the numerical implications of the double-limit model on existing bounds on $\lambda(n)$; in such a double-limit setting, we assume a single limit in n and a suitable (implicit) limit in t, e.g., $t = \omega(n^2)$.

A. A Calculus for $\lambda(n)$

We revisit the key ideas from [15] to compute upper and lower bounds on $\lambda(n)$. We argue in particular that both (asymptotic) bounds do not capture a non-negligible multiplicative/fractional factor, which means that the bounds can be quite loose in finite regimes; Subsection II-B provides related numerical results.

1) Upper bounds: The underlying idea is based on the condition

$$n\lambda(n)h \leq x \ , \tag{2}$$

where h is a lower bound on the number of average hops, whereas x is an upper bound on the number of simultaneous and successful active nodes (see p. 402, 2nd column, 1st equation from [15]). The left-hand side (LHS) is thus a lower bound on how much information *must* be transmitted, assuming a rate $\lambda(n)$ for each source, whereas the right-hand side (RHS) is an upper bound on how much information *can* be transmitted (note that both LHS and RHS are asymptotic *rates*, i.e., time averages of some stochastic processes). For the random network model from [15], $h = \Theta\left(\sqrt{\frac{n}{\log n}}\right)$ and $x = \Theta\left(\frac{n}{\log n}\right)$; the two asymptotic expressions are sufficient to guarantee structural properties in the random network model from [15] with high probability.

The upper-bound argument from Eq. (2) was extensively used in the network capacity literature: see, e.g., Eq. (2) in Li *et al.* [19] for unicast capacity in static ad hoc networks, Eqs. (18,20) in Mergen and Tong [21] for unicast capacity in networks with regular structure, Eq. (26) in Neely and Modiano [22] for unicast capacity in some mobile networks,

Eq. (2) in Shakkottai *et al.* [23] for multicast capacity, and even in several much earlier papers by Kleinrock and Silvester (see Eq. (18) in [18] for unicast capacity in uniform random networks and Eqs. (12,25,29) in [25] for unicast capacity in Aloha networks with regular structure).

Let us now discuss the validity of the upper-bound argument in more restrictive space/time models. In a finite-space (fixed n) infinite-time model, the argument also holds subject to further conditions: e2e paths must exist for all source-destination pairs and (non-asymptotic) expressions for h and x are known. Under the same structural conditions, the upper-bound continues to hold in a finite time/space model (fixed n and time span T) by properly interpreting rates over finite time intervals.

What is interesting to observe in the finite regime is that Eq. (2) can be (approximately) tightened as

$$n\lambda(n)h \leq g(n,T)x \ , \tag{3}$$

where $g(n,T)$ denotes the average fraction of the number of non-empty buffers. Indeed, over a finite time span T, the average number of simultaneous and successful transmissions decays due to transient burstiness effects, in particular due to the existence of empty buffers at the very nodes scheduled to transmit.

2) Lower bounds: By explicitly constructing a routing and scheduling scheme, the underlying idea to compute lower bounds on $\lambda(n)$ is based on the condition

$$\lambda(n)l \leq c \ , \tag{4}$$

where l is an *upper bound* on the number of e2e transmissions a node *needs* to act as a relay, whereas c is the maximal rate at which a node *can* transmit (see p. 400, 2nd column, 1st equation from [15]). For the network model from [15], $l = \Theta\left(\sqrt{n \log n}\right)$ and $c = \Theta(1)$.

Alike the upper-bound, the lower-bound continues to hold in a finite-space model under appropriate structural properties. In a finite-space/time model, however, the lower bound ceases to hold. For a counterexample (relative to current conditions), consider 3 nodes numbered as $1, 2, 3$, the (direct) source-destination pairs $(1,2), (2,3), (3,1)$, and assume that all transmissions interfere with each other. Fitting Eq. (4) yields $l = 1$, $c = \frac{1}{3}$, and thus $\lambda(3) = \frac{1}{3}$. Evidently, this rate can only be sustained for specific values of T (e.g., if exogenous arrivals occur at times $3s + 1$ at all nodes, and under a round-robin scheduling, then the lower bound only holds at times $T = 3s$ for $s \geq 0$).

In finite scenarios in which the lower-bound does hold, it can be further tightened using the same multiplicative term as for the upper bound, i.e.,

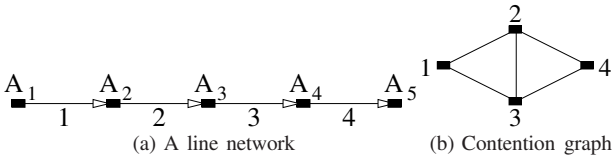
$$\lambda(n)lg(n,T) \leq c \ . \tag{5}$$

Note that the effective multiplicative factor for the lower bound is in fact $\frac{1}{g(n,T)}$. Alike Eq. (3), the improvement from Eq. (5) only holds approximately (without very strong non-asymptotic mixing conditions it is conceivably hard to exactly keep track of the nodes with non-empty buffers and which can successfully transmit).

In conclusion, the upper and lower bounds arguments from Eqs. (2) and (4) hold immediately in a finite-space model. In finite-space/time models, only the former holds in general, and both can be improved by a multiplicative and fractional, respectively, factor $g(n,T)$; next we will show that this factor can be quite small (and thus detrimental), including at large values of n. Finally, we remark that even by considering a tightening factor, as in Eqs. (3) and (5), capacity results remain restricted to first moments (time averages) only. These limitations demand thus for 'richer' capacity results in terms of probability distributions, which can readily render *all* moments, and more generally for alternative analytical techniques beyond the convenient but simplistic averages-based calculus.

B. Simulations for $g(n,T)$

Here we simulate the multiplicative/fractional factor $g(n,T)$ identified in Eqs. (3) and (5). For the clarity of the exposition, we consider both a simple setting, consisting of a single e2e transmission along a line network, and a more involved random network.



(a) A line network (b) Contention graph

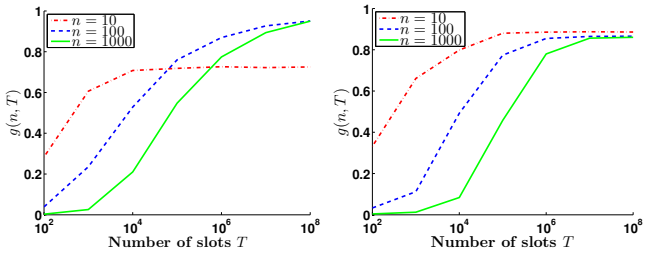
Fig. 1. A multi-hop network and its contention graph

1) Example 1: A Single E2E Transmission: Consider a multi-hop network with n nodes, of which a source node A_1 transmits (packets) to a destination node A_n using the relay nodes A_i, $i = 2, 3, \ldots, n-1$. The end-to-end (e2e) transmission is denoted by $[A_1 \rightarrow A_n]$. We denote by c_r the contention range of link i, i.e., link i interferes with any link j satisfying $|i - j| < c_r$. An example for $n = 5$ is shown in Figure 1.(a). A corresponding contention graph for $c_r = 3$ is shown in Figure 1.(b). Here, each vertex i stands for the uni-directional transmission $[A_i \rightarrow A_{i+1}]$, $i = 1, 2, 3, 4$, and there is a link between nodes i and j if the corresponding transmissions interfere with each other; according to this contention graph, links 1 and 4 can simultaneously and successfully transmit.

Next we illustrate the (time) average fraction of non-empty buffers $g(n,T)$ for the network setting from Figure 1 over a time span T, and for two MAC protocols: (slotted) Aloha and CSMA/CA, to be described in Section IV.

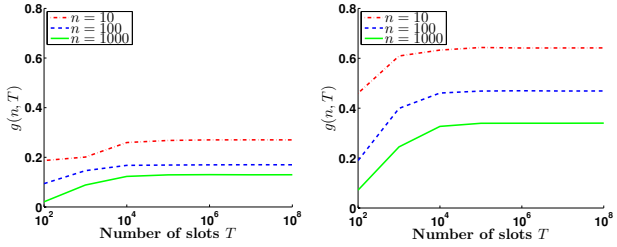
Figure 2.(a) illustrates the Aloha case. The network capacity is $\lambda(n) = p(1-p)^4$, where p is the transmission probability; by optimizing, $p = 0.2$ and $\lambda(n) \approx 0.08$, which is the rate injected at the first node (recall that there is a single e2e transmission). We observe that for small number of nodes (e.g., $n = 10$), the (average) fraction of empty buffers remain significant, even when taking the limit in T. This effect is due to insufficient amount of spatial reuse. For larger number of nodes, however, there is sufficient amount of spatial reuse and the fraction of empty buffers goes to zero. This convergence, however, is surprisingly slow (at $T = 10^8$ there are still $\approx 5\%$ empty buffers).

Figure 2.(b) illustrates the CSMA/CA case with average backoff and transmission times $\nu^{-1} = 10$ and $\mu^{-1} = 10$, respectively. Because a formula for $\lambda(n)$ is difficult to obtain,



(a) Aloha (b) CSMA/CA

Fig. 2. The average fraction of non-empty buffers $g(n,T)$ as a function of the time span T in a line network



(a) Aloha (b) CSMA/CA

Fig. 3. The average fraction of non-empty buffers $g(n,T)$ as a function of the time span T in a random network

unlike in the Aloha case, we numerically searched for the maximum value of $\lambda(n)$ such that the total amount of packets in all buffers, except destinations, at any time, is smaller than 10^6 over a maximum time span $T = 10^8$. We obtained $\lambda(n) \approx 0.17$ for $n = 10, 100, 1000$. As for Aloha, the homogeneity is due to the dominant effect of a bottleneck; unlike Aloha, however, CSMA/CA is subject to transmission correlations spanning the entire network. Relative to the Aloha case, we also remark a sharper rate of increase of $g(n,T)$; however, this rate slows down earlier (e.g., for $n = 100$ there are still $\approx 13\%$ empty buffers in contrast to only $\approx 5\%$ for Aloha).

2) Example 2: A Random Network: We now consider a closer network setting to the one from [15]. We first randomly place n nodes on a square and randomly choose a destination for all sources $1, 2, \ldots, n$, according to uniform distributions. Then we determine the minimum transmission range such that e2e paths exist for all source-destination pairs; these paths are constructed using a shortest path algorithm with equal weights for all links. Each node stores the locally generated and incoming packets in a FIFO buffer.

As in Example 1, we illustrate $g(n,T)$ as a function of T for both Aloha and CSMA/CA. For Aloha we set the nodes' transmission probability as the inverse of the maximum node-degree amongst all nodes. Moreover, we use the same numerical search form Example 1 to determine the capacity $\lambda(n)$ for both Aloha ($\lambda(10) \approx 0.01$, $\lambda(100) \approx 0.002$, and $\lambda(1000) \approx 0.0004$) and CSMA/CA ($\lambda(10) \approx 0.03$, $\lambda(100) \approx 0.007$, and $\lambda(1000) \approx 0.0009$).

From Figure 3.(a) we observe a clear convergence of $g(n,T)$; the perhaps surprisingly low limits are conceivably due to the homogeneous transmission probabilities accounting for the bottleneck region. Unlike in the line network, there is a consistent monotonic behavior in the number of nodes n, which is likely due to the more uniform structure of the random network setting. In the CSMA/CA case, Figure 3.(b) illustrates that much fewer buffers (by a factor of roughly three) are empty than in the Aloha case, which suggests a less burstier behavior in CSMA/CA.

Clearly, Figures 2 and 3 open several fundamental questions on network queueing behavior for Aloha and CSMA/CA, which may help improving the two. Their main purpose, however, is to convincingly show that the average fraction of non-empty buffers $g(n,T)$ is quite small especially in random networks, and in general at small time scales. The key observation is the monotonic (decreasing) behavior (excepting the special Aloha line with $n = 10$) in the number of nodes n. This behavior is 'somewhat expected' in large networks, by invoking laws of large numbers arguments, i.e., the overall incoming and outgoing flows tend to stabilize and thus buffers tend to decrease. This indicates that both the original upper and lower bounds from Eqs. (2) and (4) become conservative in asymptotic regimes (in n).

In conclusion, the results from this section motivate the need for an analytical approach to network capacity in finite regimes. At this point, we ought to be rigorous in defining capacity in finite time. Concretely, given a time t and an arrival process $D(t)$ at the destination of an e2e path, we are interested in bounds of the form

$$\mathbb{P}\left(D(t) \leq \lambda_t t\right) \leq \varepsilon \ ,$$

for some violation probability ε. Here, λ_t is a *lower* bound on the throughput (capacity) rate of the e2e transmission; a corresponding upper bound can be defined similarly.

III. A System Theoretic Approach to Finite Time Capacity

We have just shown that the main challenge to derive finite-time capacity results in terms of distributions is accounting for queueing behavior in a conceivably non-Jackson queueing network. In particular, it is especially hard to analytically keep track of buffer occupancies at relay nodes.

To address this problem, we next describe a general solution to circumvent the characterization of buffer occupancies at the relay nodes, by making an analogy with LTI systems. The idea is to view a single-hop transmission as follows: the data at the source and destination stands for the input and output signals, respectively, whereas the transmission and its characteristics, accounting for both data unavailability due to burstiness or noise due to interference, are modelled by '*the system*' transforming the input signal. Although this *system* is not linear, there is a subtle analogy with LTI systems which drives its analytical tractability.

To present the main idea in an approachable manner, from the point of view of notational complexity, we focus on the simplified simple line network from Figure 1.

Figure 4 illustrates a system view for the e2e transmission from Figure 1. With abuse of notation the A_i's stand for the input/output signals, and the S_i's stand for the impulse-responses of the systems. The key property of the impulse-responses is to relate the input and output signals through a

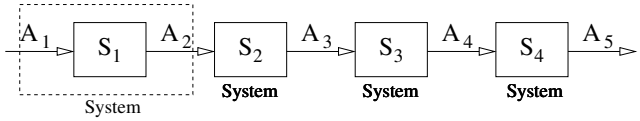


Fig. 4. A system interpretation of the multi-hop network from Figure 1

convolution operation, i.e.,

$$A_{i+1} = A_i * S_i \ . \tag{6}$$

As it will become more clear in Section IV, the convolution operation denoted here by the symbol '$*$' operates in a $(\min, +)$ algebra. To be more specific, let A_i stand for a stochastic process $A_i(t)$, which counts the number of packets in the time interval $[0, t]$ at node i; also, let S_i stand for (some) bivariate stochastic processes $S_i(s, t)$. Then the $(\min, +)$ convolution operation expands as

$$A_{i+1}(t) = \min_{0 \le s \le t} \{A_i(s) + S_i(s, t)\} \ \ \forall t \ge 0 \ .$$

The relationship from Eq. (6) has two key properties. One is that it holds for *any* input signal A_i, which is hard to derive at relay nodes (when $i \ge 2$). In other words, the impulse-response S_i *entirely* characterizes the *system*, i.e., the single-hop transmission i, which is a key feature of LTI systems since it enables their analytical tractability. The convolution operation has also the useful algebraic property of *associativity*. The two properties circumvent keeping track of $A_i(t)$ at the relay nodes (i.e., for $i = 2, 3, 4$). Indeed, by applying associativity, and using the physical property that the output signal in a system is the input signal at the downstream system, the composition of the four systems from Figure 4 yields the reduced system from Figure 5.

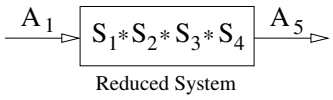


Fig. 5. Composition of the four systems from Figure 4 into a single system

The reduced system dispenses with the intermediary signals A_2, A_3, and A_4, and instead it retains the impulse-responses in a composition (or e2e) form, i.e.,

$$A_5 = A_1 * (S_1 * S_2 * S_3 * S_4) \ . \tag{7}$$

What has yet to be shown concerns the existence of (analytical) expressions for the impulse-responses S_i's, satisfying the key property from Eq. (6). The other open issue is whether the convenient reduction from Figure 5 and Eq. (7) is analytically tractable. We will next show that impulse-responses can be constructed for single-hop links in an analogous manner as in LTI systems, depending on the underlying MAC protocol, whereas analytical tractability follows the steps of large deviations or stochastic network calculus theories.

IV. Markov Modulated Transmission Processes (MMTPs)

Wireless networks must deal with the fundamental interference problem: two simultaneous transmissions may jointly fail if they interfere with each other. MAC protocols partially resolve this problem by reducing the number of collisions and consequently increasing the network capacity. Obviously, different MAC protocols can lead to different capacities.

To capture the influence of MAC protocols on the throughput capacity, we introduce the concept of Markov Modulated Transmission Process (MMTP). An MMTP is defined for each link, and models the link's activities (successful/unsuccessful transmissions and idle periods) as a time process and according to the workings of the underlying MAC protocol. The model consists of a Markov chain/process $X(t)$ (depending on the underlying discrete/continuous time model), where t is a time parameter, which *modulates* the transmission rate of a link $[i \to j]$, if the source i has data to send at time t. In discrete time, the transmission rate in a slot t is

$$S(t-1, t) = C_{X(t)} \ , \tag{8}$$

where $C_{X(t)}$ is the *Markov Modulated Transmission Process* defined on the state space $\mathcal{T}$ of the Markov chain $X(t)$. It is described as

$$C_{X(t)} = \begin{cases} C & , \quad X(t) \in \mathcal{T}_{[i \to j]} \\ 0 & , \quad \text{otherwise} \ , \end{cases} \tag{9}$$

where $\mathcal{T}_{[i \to j]} \subseteq \mathcal{T}$ denotes the set of *favorable* states of $X(t)$ for the link $[i \to j]$, which would guarantee a successful transmission if the link has data to send at time t; whenever a transmission is successful we assume a constant throughput capacity C. The rest of the states $X(t) \in \mathcal{T} \backslash \mathcal{T}_{[i \to j]}$ model the times when the link attempts an unsuccessful transmission or it is idle in accordance to the MAC protocol.

The MMTP process $C_{X(t)}$ defined in Eq. (9) is modulated by the Markov chain $X(t)$, and it is conceptually similar to Markov Modulated Poisson Processes. $X(t)$ can be defined either for the *whole* network (when it modulates the transmission opportunities of *all* the links) or for *each* link separately. In turn, $C_{X(t)}$ is always separately defined for *each* link.

We point out that the process $S(s, t)$, which we loosely introduced in Eq. (8) through its increments $S(t-1, t)$, directly corresponds to the *impulse-response* process introduced in Section III to *entirely characterize* the behavior of a single-hop transmission in system theoretic terms (see Figure 4 and Eq. (6)). The impulse-response defined in Eq. (8) corresponds to the *effective capacity* concept proposed by Wu and Negi in [27] to model the instantaneous channel capacity. This concept was used by Tang and Zhang [26] to analyze the impact of physical layer characteristics (e.g., MIMO) on delay at the data-link layer. The MMTP idea was also used explicitly by Fidler [13], Mahmood *et al.* [20], Al-Zubaidy *et al.* [29], Zheng *et al.* [28] and implicitly by Ciucu *et al.* [5], [4], [8], for the Aloha protocol.

Relative to these previous works, our contribution is to fit the effective capacity concept for *three* MAC protocols, and in a *unified* manner. In the following we explicitly construct the corresponding impulse-response processes $S(s, t)$ and outline the key steps to compute lower-bounds on the per-flow capacity (for the complete results, including upper bounds, see [6]).

A. Centralized scheduling

Assuming a time-slotted model and the nodes' perfect synchronization, the idea of centralized scheduling is to pre-allocate the transmission slots to the nodes in order to avoid

collisions. In unsaturated scenarios, an optimal solution (i.e., attaining maximal throughput) would require significant overhead as the centralized scheduler would require keeping track of the arrival processes at each of the nodes. Even in saturation scenarios, the optimality problem is in fact NP-complete in general networks (see, e.g., Sharma *et al.* [24]).

For the network model from Figure 1.(a-b), the optimal scheduling allocation starting from slot 1, in terms of links, is: $\{1, 2, 3, (1, 4), 2, 3, (1, 4), \dots\}$; for instance, link 2 is allocated the slots $2, 5, 8, \dots$ That means that link 2 is given full transmission capacity (say C) during these slots, which suggests that the bivariate function

$$S_2(s,t) = C \sum_{u=s+1}^{t} I_{(u-2)\%3=0} \ \forall 2 \leq s \leq t \ , \qquad (10)$$

and 0 everywhere else, characterizes the capacity of link 2 in terms of the $(\min, +)$ convolution from Eq. (6) (I_E denotes the indicator function taking the values 0 and 1, depending on whether the event E is false or true). The intuition is that $S_2(s,t)$ counts the number of packets transmitted over link 2 in the time interval $(s,t]$, if node 2 is saturated.

This saturation condition translates into system theoretic terms as follows: the input signal to the second system in Figure 4 is the infinite signal $A_2(t) = \infty$ for $\forall t$ (also called the *impulse*), whereas the corresponding output, i.e., the *impulse-response*, is the signal $S_2(t)$, or $S_2(0,t)$ in the notation from Eq. (10). Therefore, the construction of $S_2(s,t)$ is analogous to the construction of impulse-response functions in LTI systems, which are the output from an LTI system with input given by the Kronecker signal. Although the system representing the link's transmission is *not* linear, even under the $(\min, +)$ algebra, the constructed process $S_2(s,t)$ *entirely characterizes* link 2, i.e.,

$$A_3(t) = \min_{0 \leq s \leq t} \{A_2(s) + S_2(s,t)\} \ \ \forall t \geq 0 \ , \qquad (11)$$

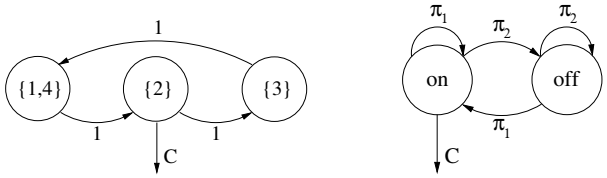
for *all* $A_2(t)$ at the input of the second system.

So far we directly constructed $S_2(s,t)$ without resorting on an MMTP $C_{X(t)}$. The underlying MMTP, and also the modulating Markov chain $X(t)$, are depicted in Figure 6.(a). The states of $X(t)$ denote the set of transmitting links (according to the centralized schedule). The transition probabilities between the states are all equal to 1, thus reflecting the deterministic nature of centralized scheduling. The MMTP process for link $[A_2 \to A_3]$ is

$$C_{X(t)} = \begin{cases} C & , \ \text{if } X(t) = \{2\} \\ 0 & , \ \text{otherwise} \ . \end{cases}$$

The MMTPs for the other links are defined similarly; for instance, for links 1 and 4, the only change is that $C_{X(t)} = C$ when $X(t) = \{1, 4\}$. Note that all MMTPs share the same Markov chain modulating the transmission opportunities at the network level. Moreover, $X(t)$ and $C_{X(t)}$ jointly reproduce the expressions of the impulse responses (e.g., of $S_2(s,t)$ from Eq. (10)) according to the definition from Eq. (8).

Concerning analytical tractability, we remark that the reduced system from Figure 5 is implicitly tractable since the constructed impulse-responses $S_i(s,t)$ are deterministic



(a) Centralized scheduling (b) Aloha

Fig. 6. Markov Modulated Transmission Processes (MMTPs) for link $[A_2 \to A_3]$

functions. The overall impulse-response of the e2e path $S = S_1 * S_2 * S_3 * S_4$ from Eq. (7) can be computed directly, i.e.,

$$S(s,t) = C \sum_{u=s+1}^{t} I_{(u-1)\%3=0} \ \forall 1 \leq s \leq t \ , \qquad (12)$$

and 0 everywhere else.

Although the constructions of the MMTP's above is not technically necessary, as the impulse-responses were directly constructed, and the impulse-response $S(s,t)$ of the e2e path could be in principle determined by other means than computing an e2e convolution, we regard this detour to be insightful for the construction of impulse-responses for the more challenging cases of Aloha and CSMA/CA protocols.

B. Aloha

The (slotted) Aloha MAC protocol is an elegant solution to circumvent centralized scheduling (see Ambramson [1]). The key idea is that each node attempts to transmit with some probability in each time slot and when data is available; a transmission $[i \to j]$ is successful in a time slot t if i is the only node in the interference range of node j attempting to transmit in that slot. While the protocol is entirely distributed, it may experience significant performance decay, e.g., the achieved capacity can be as small as 36% of the theoretical limit.

To construct the impulse-response processes for the line network from Figure 1.(a-b), we first construct the MMTP processes. We focus again on link $[A_2 \to A_3]$. The underlying MMTP, and also the modulating Markov chain $X(t)$, are depicted in Figure 6.(b). The meaning of state 'on' is that, while $X(t)$ delves in it, the relay node A_2 successfully transmits (if there is data to send). In turn, while $X(t)$ delves in state 'off', A_2 is either idle or it is involved in a collision. Assume for convenience that all nodes transmit with the same probability p. The transition probabilities are $\pi_1 = p(1-p)^3$ and $\pi_2 = 1 - \pi_1$ (the power of 3 is the degree of node 2 in the contention graph from Figure 1.(b)). For this Markov chain, the steady-state probabilities are $\pi_{\text{on}} = \pi_1$ and $\pi_{\text{off}} = \pi_2$, and $X(t)$ has the convenient property of statistically independent increments, e.g., $\forall t$

$$\mathbb{P}\left(X(t+1) = \text{'on'} | X(t) = \text{'off'}\right) = \mathbb{P}\left(X(t+1) = \text{'on'}\right) \ . \qquad (13)$$

The definition of the associated MMTP should be intuitive at this point, i.e.,

$$C_{X(t)} = \begin{cases} C & , \ \text{if } X(t) = \text{'on'} \\ 0 & , \ \text{otherwise} \ , \end{cases} \qquad (14)$$

as also illustrated in Figure 6.(b). In other words, A_2 can successfully transmit (assuming it has data) at full rate C while the modulating process $X(t)$ delves in the favorable state 'on'.

The construction of the other links' MMTP processes is almost identical, except for the transmission probabilities of the modulating process. For instance, for the links $[A_1 \to A_2]$ and $[A_4 \to A_5]$, the new transmission probabilities are $\pi_1 = p(1-p)^2$ and $\pi_2 = 1-\pi_1$ (the power of 2 is the common degree of nodes 1 and 4 in the contention graph from Figure 1.(b)).

These MMTP processes directly determine the impulse-response functions $S_i(s,t)$, corresponding to the single-hop links $i = 1, 2, 3, 4$, according to the definition from Eq. (8). Furthermore, the composition property of the $S_i(s,t)$'s in the underlying $(\min, +)$ algebra lends itself to the *entire characterization* of the throughput capacity over the e2e path as in Eq. (7), i.e., $A_5 = A_1 * S$, where $S = S_1 * S_2 * S_3 * S_4$ is the impulse-response of the e2e path.

Unlike centralized scheduling which may lend itself to an explicit expression for S (e.g., as in Eq. (12)), Aloha is more challenging with respect to the analytical tractability of the reduced system. One immediate issue lies in the probabilistic structure of the local impulse-responses S_i's. A more subtle issue lies in the fact that the S_i's are statistically correlated random processes, even in the simplified line network.

To deal with these challenges, the key idea is to *trade analytical exactness for tractability*. More concretely, instead of *exactly* deriving the e2e transient capacity in closed-form (an open problem in itself), we compute bounds by relying on large deviation techniques (e.g., as in [5]). Let us illustrate such computations for the first two hops only, and a saturation assumption at node A_1 (the relay nodes are however *not* assumed to be saturated). The probability of violating a lower bound λ_t, on the *transient throughput rate* over the time scale $[0,t]$, can be computed as follows for some $\theta > 0$

$$\begin{aligned} \mathbb{P}\left(A_3(t) \le \lambda_t t\right) &= \mathbb{P}\left(S_1 * S_2(t) \le \lambda_t t\right) \\ &= \mathbb{P}\left(\sup_{0\le s\le t}\left\{\lambda_t t - S_1(s) - S_2(s,t)\right\} \ge 0\right) \\ &\le \sum_{0\le s\le t} e^{\theta\lambda_t t} E\left[e^{-\theta S_1(s)}\right] E\left[e^{-\theta S_2(s,t)}\right] , \end{aligned} \tag{15}$$

by using Boole's inequality and the Chernoff bound; in the first line we also used that A_1 is saturated. The last step is based on the statistical independence between the impulse-responses $S_1(u,s)$ and $S_2(s,t)$ (in order to apply $E[XY] = E[X]E[Y]$ for some independent r.v.'s X and Y); this holds because $(u,s]$ and $(s,t]$ are non-overlapping intervals, whereas the corresponding Markov modulated processes of $S_1(u,s)$ and $S_2(s,t)$ have statistically independent increments (recall Aloha's useful property from Eq. (13)). Therefore, although $S_1(u,s)$ and $S_2(u,s)$ are correlated over overlapping intervals, the expansion of the $(\min, +)$ convolution (in terms of non-overlapping intervals) and the independent increments property from Eq. (13) justify the last step.

Finally, the Laplace transforms in the last equation can be computed explicitly for the impulse-responses. Concretely, $E\left[e^{-\theta S_i(s,t)}\right] \le e^{-\theta r_s(t-s)}$, where $r_s = \frac{\log\left(qe^{-\theta C}+1-q\right)^{-1}}{\theta}$ and $q = p(1-p)^3$. Evaluating the nested sums from Eq. (15) finally yields the closed-form bound

$$\mathbb{P}\left(A_3(t) \le \lambda_t t\right) \le \inf_{\theta>0}(t+1)e^{-\theta(r_s-\lambda_t)t} . \tag{16}$$

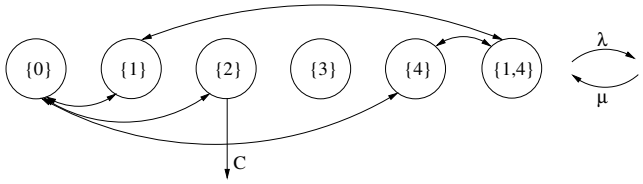


Fig. 7. A Markov Modulated Trans. Proc. (MMTP) for link $[A_2 \to A_3]$

Next we give a general result for computing the lower bounds on the end-to-end throughput capacity for a flow crossing k hops (for the proof see [6]).

Theorem 1: (CAPACITY BOUNDS (LOWER BOUND) - ALOHA) Consider a flow crossing k hops. Assume that the impulse-response process $S_j(s,t)$, at each hop j, satisfies the following bounds on the Laplace transform: $E\left[e^{-\theta S_j(s,t)}\right] \le e^{-\theta r_j(-\theta)(t-s)}$ for some $r_j(-\theta)$ and all $\theta > 0$. Let $r(-\theta) := \min_j r_j(-\theta)$. Assume also that $S_j(s,t)$ are statistically independent over non-overlapping intervals. Then, for some $\varepsilon > 0$, a probabilistic lower bound on the capacity rate is for all $t \ge 0$

$$\lambda_t = \sup_{\theta>0}\left\{r(-\theta) + \frac{\log\varepsilon - \log\binom{t+k-1}{k-1}}{\theta t}\right\} . \tag{17}$$

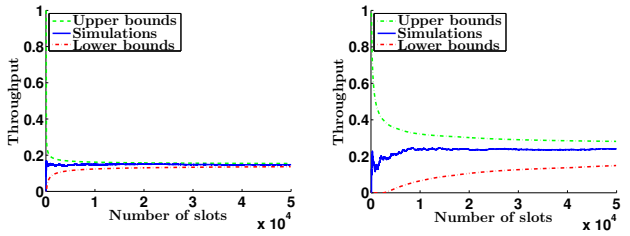
C. CSMA/CA

The CSMA/CA protocol was motivated by the need to increase the (very) low capacity of Aloha, while preserving the distributive aspect of the protocol. One key idea is to prevent collisions from happening by enabling nodes to 'listen to the channel' before transmitting. The other key idea is that once a node perceives the channel as being busy, it enters in an exponentially distributed backoff mode.

We use a simplified CSMA/CA protocol, developed by Durvy et al. [10], which retains the key features of CSMA/CA. For the network from Figure 1.(a-b), the construction of the MMTP processes, and also of the impulse-response processes $S_i(s,t)$, follows similarly as for centralized scheduling and Aloha. For the link $[A_2 \to A_3]$, the underlying MMTP, and also the modulating (now continuous-time) Markov process $X(t)$, are depicted in Figure 7. $X(t)$ is constructed exactly as in [10], where ν^{-1} and μ^{-1} denote the average backoff and transmission times. The interpretation of the states is identical as for centralized scheduling (see Figure 6.(a)); while $X(t)$ delves in the new state $\{0\}$, all nodes are in a backoff mode. Ignoring the details of switching from discrete to continuous time, the MMTP is defined as for Aloha (see Eq. (14)), i.e.,

$$C_{X(t)} = \begin{cases} C & , \text{ if } X(t) = \{2\} \\ 0 & , \text{ otherwise} , \end{cases}$$

and the impulse-response $S_2(s,t)$ is defined as in Eq. (8). The MMTPs for the other links are defined similarly (e.g., for $[A_1 \to A_2]$, the only change is that $C_{X(t)} = C$ when $X(t) \in \{\{1\}, \{1,4\}\}$).



(b) Aloha, $k = 3$ (d) CSMA/CA, $k = 3$

Fig. 8. Throughput rates as a function of the number of time slots for the network from Figure 9.(b) (per-hop rate $r_{\text{mh}} = 1$, $p = \frac{1}{k}$ for Aloha, $\frac{1}{\nu} = \frac{1}{\mu} = 10$ for CSMA/CA, and a violation probability $\varepsilon = 10^{-3}$).

Unlike in Aloha, deriving the e2e capacity in CSMA/CA is more challenging. While the first two lines from Eq. (15) still hold, the last line does not hold anymore since $S_1(u, s)$ and $S_2(s, t)$ are *not* independent, even over non-overlapping intervals. The reason is that the modulating process $X(t)$ does *not* have independent increments (as in Eq. (13)). The immediate work-around is Hölder's bound, i.e., $E\left[e^{-\theta S_1(s)} e^{-\theta S_2(s,t)}\right] \leq E\left[e^{-p\theta S_1(s)}\right]^{\frac{1}{p}} E\left[e^{-q\theta S_2(s,t)}\right]^{\frac{1}{q}}$ for $\frac{1}{p} + \frac{1}{q} = 1$.

The second challenge is to compute the Laplace transforms of the impulse-responses $S_i(s, t)$'s. Since these processes are Markov arrival processes (MAPs), their Laplace transforms can be computed using standard techniques. Let us compute in particular $L_t := E\left[e^{-\theta S_2(t)}\right]$, for some $\theta > 0$. Denote the six states of the MAP from Figure 7 by the numbers $0, 1, \ldots, 5$, and the elements of the generator matrix by $p_{i,j}$ (e.g., $p_{4,5} = \nu$). Denote also the conditional Laplace transforms $L_{i,t} := E\left[e^{-\theta S_2(t)} \mid X(0) = i\right]$, i.e., conditioned on the initial state of the Markov chain $X(t)$, which starts in steady-state. For any initial state (e.g., $i = 2$) we have the backward equation

$$\begin{aligned} L_{2,t+\Delta t} &= E\left[e^{-\theta S_2(\Delta t)} \mid X(0) = 2\right] \\ &\quad \sum_j E\left[e^{-\theta S_2(\Delta t, t+\Delta t)} \mid X(\Delta t) = j\right] p_{2,j} \\ &= e^{-\theta C \Delta t}\left(L_{0,t}\mu\Delta t + L_{2,t}\left(1 - \mu\Delta t\right) + o(\Delta t)\right) , \end{aligned}$$

where $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$. In the last line we used the stationarity of $S_2(t)$. Using the Taylor's expansion $e^{-\theta C \Delta t} = 1 - \theta C \Delta t + o(\Delta t)$, rearranging terms, and taking the limit $\Delta t \to 0$ it follows that

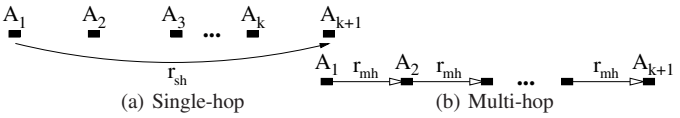
$$\frac{\partial L_{2,t}}{\partial t} = L_{0,t}\mu - L_{2,t}(\mu + \theta C) \ . \tag{18}$$

One can proceed similarly to derive the PDE's of the other $L_{i,t}$'s for $i \neq 2$, and arrive at the system of PDE's

$$\frac{\partial \boldsymbol{L_t}}{\partial \boldsymbol{t}} = B\boldsymbol{L_t} \ , \tag{19}$$

where $\boldsymbol{L_t} = (L_{0,t}, \ldots, L_{5,t})^T$ and some matrix B. The drawback of the obtained solution is that it depends on the eigenvalues/eigenvectors of the matrix B, and is thus not easily amenable to convex optimizations[1]. For a brief illustration of the bounds' tightness see Figure 8.

[1]More general lower and upper bounds on the e2e capacity are given in the technical report [6].



(a) Single-hop (b) Multi-hop

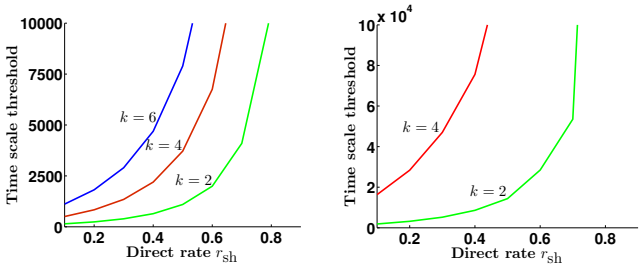
Fig. 9. Which strategy should node A_1 choose in order to transmit to A_{k+1}? (all nodes hear each other, all are saturated, and $r_{\text{sh}} < r_{\text{mh}}$)

V. Application: Single-Hop vs. Multi-Hop

In this section we demonstrate how to use the *finite time and space* key features of our capacity bounds for the following problem. Consider the network from Figure 9 with $k+1$ nodes, all within the interference range of each other, and all being saturated (i.e., being the source with infinite data for some e2e transmission) and attempting to access the channel using either the Aloha or CSMA/CA protocols. Given that node A_1 intends to transmit to node A_{k+1}, the problem concerns choosing between the following two routing strategies:

1) Single-hop: Node A_1 directly transmits to node A_{k+1} at rate r_{sh}.
2) Multi-hop: Node A_1 transmits using the nodes $A_2, A_3, \ldots, A_k$ as relays; the rate for each transmission $[A_j \to A_{j+1}]$ is r_{mh}.

We assume that at each node i the transmission $[A_1 \to A_{k+1}]$ has priority over all others. Moreover, to avoid a trivial answer, we assume that $r_{\text{sh}} < r_{\text{mh}}$.



(a) Aloha (b) CSMA/CA

Fig. 10. The time scale (threshold) after which the multi-hop is more advantageous than the single-hop strategy ($\varepsilon = 10^{-3}$, normalized per-hop rate $r_{\text{mh}} = 1$, variable direct rate r_{sh}, Aloha transmission probability $p = \frac{1}{k}$, and $\nu = \mu = 0.1$ for CSMA/CA)

Figure 10 illustrates the threshold at which the multi-hop strategy is more advantageous. More concretely, the values displayed (i.e., the 'Threshold') are the time scales at which the *lower-bound* for the multi-hop transmission is larger than the *upper-bound* for the single-hop transmission[2]. Both (a) and (b) indicate the intuitive facts that the 'Threshold' is exponential in the relative direct rate r_{sh} and also increasing in the number of hops k. In (b), for CSMA/CA, the benefits of multi-hop routing hold only for very low relative direct rate r_{sh} and quickly vanish by increasing k. We point out however that this quick blow-up may be due to the loose underlying upper bounds on the CSMA/CA per-flow capacity (see Figure 8).

The above routing problem has been debated in different settings such as wireless mesh and sensor networks. Experimental results by De Couto *et al.* [9] showed that minimizing the hop count is not always the best option as long hops may incur a high packet error rate. Jain *et al.* [17] showed that, due

[2]The lower and upper bounds are applications of Theorems 1 and 2 from [6].

to interference, shortest paths with long hops may not provide the best performance. In contrast, there are several results supporting long-hop routing. Haenggi and Puccinelli [16] provided many reasons why short-hop routing is not as beneficial as it seems to be. Moreover, in energy limited networks such as sensor networks, long-hop routing may also be preferable (see Ephremides [11] and Björnemo *et al.* [3]).

Our contribution to this debate is to bring a new perspective on single vs. multi-hop routing by focusing on the underlying time scale. Concretely, we provided theoretical evidence that multi-hop routing is more advantageous in the long-run for Aloha. In turn, in the case of CSMA/CA, the advantage of multi-hop vanishes in most cases. We raise however the awareness that, for the purpose of analytical tractability, our results are restricted to a line network and no frequency or power management being accounted for.

VI. Conclusions

We have presented the key ingredients of a *unified* system-theoretic methodology to compute the per-flow capacity in finite time and space network scenarios, and for three MAC protocols: centralized scheduling, Aloha, and CSMA/CA. We have also confirmed the anecdotal practical conservative nature of alternative asymptotic results, by scrutinizing a widely used double-limit argument. Moreover, we have demonstrated that our finite time/space results can lend themselves to engineering insight, i.e., on the time scales at which multi-hop routing becomes more advantageous than single-hop routing. Overall, the advocated system-theoretic approach has the potential to contribute to the development of the long desirable *functional network information theory* (see Andrews *et al.* [2]). Immediate future work concerns fitting more realistic CSMA/CA protocols, and the improvement of the derived stochastic bounds, especially in the case of CSMA/CA, using advanced techniques as in [7].

VII. Acknowledgments

This work was partly supported by NSFC (no. 61120106008) and 863 project (no. 2013AA010401).

References

[1] N. Abramson. The Aloha system: another alternative for computer communications. In *Proceedings of AFIPS Joint Computer Conferences*, pages 281–285, 1970.

[2] J. G. Andrews, N. Jindal, M. Haenggi, R. Berry, S. Jafar, D. Guo, S. Shakkottai, R. Heath, M. Neely, S. Weber, and A. Yener. Rethinking Information Theory for mobile ad hoc networks. *IEEE Communications Magazine*, 46(12):94–101, Dec. 2008.

[3] E. Björnemo, M. Johansson, and A. Ahlén. Two hops is one too many in an energy-limited wireless sensor network. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[4] F. Ciucu. On the scaling of non-asymptotic capacity in multi-access networks with bursty traffic. In *IEEE International Symposium on Information Theory (ISIT)*, 2011.

[5] F. Ciucu, O. Hohlfeld, and P. Hui. Non-asymptotic throughput and delay distributions in multi-hop wireless networks. In *Allerton Conference on Communications, Control and Computing*, 2010.

[6] F. Ciucu, R. Khalili, Y. Jiang, L. Yang, and Y. Cui. Towards a system theoretic approach to wireless network capacity in finite time and space. *CoRR*, abs/1307.7584, 2013.

[7] F. Ciucu, F. Poloczek, and J. Schmitt. Sharp per-flow delay bounds for bursty arrivals: The case of FIFO, SP, and EDF scheduling. In *IEEE Infocom*, 2014.

[8] F. Ciucu and J. Schmitt. On the catalyzing effect of randomness on the per-flow throughput in wireless networks. In *IEEE Infocom*, 2014.

[9] D. S. J. De Couto, D. Aguayo, B. A. Chambers, and R. Morris. Performance of multihop wireless networks: shortest path is not enough. *SIGCOMM Computer Communications Review*, 33(1):83–88, Jan. 2003.

[10] M. Durvy, O. Dousse, and P. Thiran. Self-organization properties of CSMA/CA systems and their consequences on fairness. *IEEE Transactions on Information Theory*, 55(3):931–943, Mar. 2009.

[11] A. Ephremides. Energy concerns in wireless networks. *IEEE Wireless Communications*, 9(4):48–59, Aug. 2002.

[12] A. Ephremides and B. E. Hajek. Information theory and communication networks: An unconsummated union. *IEEE Transactions on Information Theory*, 44(6):2416–2434, Oct. 1998.

[13] M. Fidler. A network calculus approach to probabilistic quality of service analysis of fading channels. In *IEEE Globecom*, 2006.

[14] R. G. Gallager. A perspective on multiaccess channels. *IEEE Transactions on Information Theory*, 31(2):124–142, Mar. 1985.

[15] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, Mar. 2000.

[16] M. Haenggi and D. Puccinelli. Routing in ad hoc networks: a case for long hops. *IEEE Communications Magazine*, 43(10):93–101, Oct. 2005.

[17] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu. Impact of interference on multi-hop wireless network performance. In *ACM Mobicom*, pages 66–80, 2003.

[18] L. Kleinrock and J. Silvester. Optimum transmission radii for packet radio networks or why six is a magic number. In *Proceedings of IEEE National Telecommunication Conference*, pages 4.3.1–4.3.5, 1978.

[19] J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, and R. Morris. Capacity of ad hoc wireless networks. In *ACM Mobicom*, pages 61–69, 2001.

[20] K. Mahmood, M. Vehkapera, and Y. Jiang. Delay constrained throughput analysis of CDMA using stochastic network calculus. In *IEEE International Conference on Networks*, pages 83–88, 2011.

[21] G. Mergen and L. Tong. Stability and capacity of regular wireless networks. *IEEE Transactions on Information Theory*, 51(6):1938–1953, June 2005.

[22] M. J. Neely and E. Modiano. Capacity and delay tradeoffs for ad hoc mobile networks. *IEEE Transactions on Information Theory*, 51(6):1917–1937, June 2005.

[23] S. Shakkottai, X. Liu, and R. Srikant. The multicast capacity of large multihop wireless networks. *IEEE/ACM Transactions on Networking*, 18(6):1691–1700, Dec. 2010.

[24] G. Sharma, N. Shroff, and R. Mazumdar. Joint congestion control and distributed scheduling for throughput guarantees in wireless networks. In *IEEE Infocom*, pages 2072–2080, 2007.

[25] J. Silvester and L. Kleinrock. On the capacity of multihop slotted ALOHA networks with regular structure. *IEEE Transactions on Communications*, 31(8):974–982, Aug. 1983.

[26] J. Tang and X. Zhang. Cross-layer modeling for quality of service guarantees over wireless links. *IEEE Transactions on Wireless Communications*, 6(12):4504–4512, Dec. 2007.

[27] D. Wu and R. Negi. Effective capacity: A wireless link model for support of quality of service. *IEEE Transactions on Wireless Communication*, 2(4):630–643, July 2003.

[28] K. Zheng, F. Liu, L. Lei, C. Lin, and Y. Jiang. Stochastic performance analysis of a wireless finite-state Markov channel. *IEEE Transactions on Wireless Communications*, 12(2):782–793, Feb. 2013.

[29] H. Al-Zubaidy, J. Liebeherr, and A. Burchard. A (min, $\times$) network calculus for multi-hop fading channels. In *IEEE Infocom*, 2013.