Future Generation Computer Systems 152 (2024) 179–192

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Dimensioning the pending interest table in content-centric networks

Amuda James Abu[a], Brahim Bensaou[a], Ahmed M. Abdelmoniem[b,c,*]

[a] Computer Science and Engineering Department, HKUST, Hong Kong
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom
[c] Faculty of Computers and Information, Assiut University, Egypt

ARTICLE INFO

Keywords:
Content-centric networking
Congestion control
Modelling

ABSTRACT

In chunk-based interest-driven content-centric networks, each interest packet forwarded upstream by a node on a face implies the return of at most one data chunk to the node from that face shortly after. As a consequence, the congestion of the downstream transmission buffer in the data path of the node is highly correlated with the occupancy of the pending interest table (PIT). Therefore a systematic study and analysis of the PIT occupancy are of paramount importance to understanding congestion in CCN. In particular, in this paper, we propose an analytical model to estimate the PIT occupancy distribution via a continuous-time Markov chain (CTMC) model that considers the effects of interest blocking, interest timeout and retries. To validate our model and assumptions, we invoke simulation with realistic traffic streams and show how the filtering effects of caching and interest aggregation make the Markov assumption reasonable in the nodes that are the most susceptible to experiencing congestion. To solve the model numerically, we use two alternative approximations, state space truncation and state aggregation, and then give some numerical results to demonstrate the accuracy of our approximations.

1. Introduction

Content-Centric Networking (CCN) [6] and other similar network models[1] were recently proposed to bridge the gap between the host-centric Internet architectural model inherited from the early service model of the Internet and the information-centric service model that the Internet is de facto supporting today. CCN is a new receiver-driven information-centric networking (ICN) paradigm designed to embrace information dissemination while curbing the traffic redundancy that plagues the Internet today. In CCN, an interested receiver requests content explicitly by issuing "interest" packets that identify the requested content by name. The efficiency of resource utilization in this paradigm is enforced via two defining features: (*i*) universal in-network caching where each data chunk that traverses a CCN node in response to an interest packet, is cached in the node's content store (CS) to serve future interests if needed; (*ii*) interest aggregation in the so-called pending interest table (PIT), to avoid forwarding duplicate interest packets for content that is already pending.

Fig. 1 illustrates the operation of a CCN router node. In the current CCNx code [7], CCN faces[2] are formed via TCP/UDP/IP sockets below the CCN layer. Typically Interest packets that arrive at the CCNx forwarder from upstream nodes leave breadcrumbs in the PIT and are forwarded downstream (or aggregated if already in the PIT). Incoming data packets from downstream nodes consume the corresponding PIT entries and are forwarded upstream as indicated by this PIT entry.

The predominance of non-reusable (or one-timer) content on the Internet, the Zipfian-distributed nature of content popularity and the huge ratio of the universal-content body[3] available on the Internet to the cache sizes in the nodes imply that, despite the benefits of content caching and interest aggregation, congestion can still take place in such networks. In particular, the foreseen continued existence of legacy protocols and hardware beneath CCN (e.g., IP or Ethernet, and so on), suggests that the CS and the PIT be both elevated from the existing router buffers. As a result, the PIT will play a central role in congestion control in the data buffer. On the downside, the PIT itself

* Corresponding author at: School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom.
E-mail address: ahmed.sayed@qmul.ac.uk (A.M. Abdelmoniem).

[1] For example, NDN [1], Content-Centric Inter-Networking Architecture (CONET) [2], Publish/Subscribe Internet Routing Paradigm (PSIRP) [3], Network of Information (NetInf) [4] and Data-oriented (and beyond) Network Architecture (DONA) [5].

[2] In CCN a face is a piece of information in/outlet from/to another CCN node or from/to an application running on top of CCN.

[3] Estimated in 2016 to be of order of 10^{15} unique content of different sizes.

https://doi.org/10.1016/j.future.2023.10.009
Received 6 December 2022; Received in revised form 26 April 2023; Accepted 20 October 2023
Available online 29 October 2023

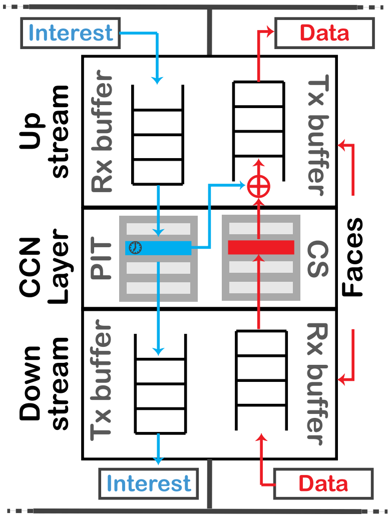
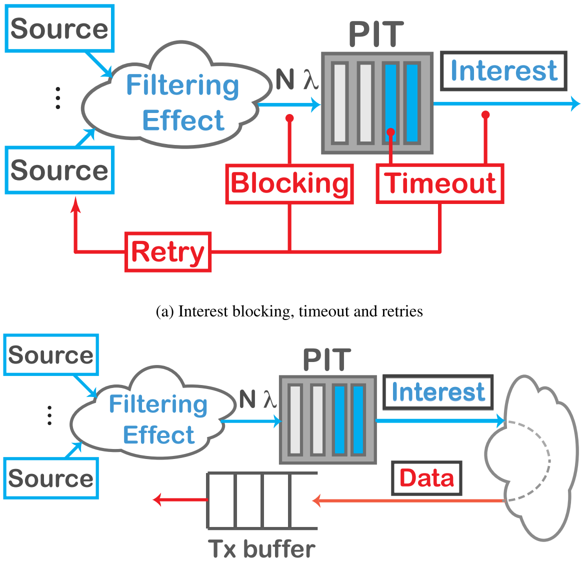


Fig. 1. Traffic flow in a CCN router: typically each CCN face is mapped to a network interface (or an application). For example in CCNx code, a face consists of two bonded BSD sockets (e.g. UDP) That ultimately map to a network interface. Incoming Interest packets leave breadcrumbs in the PIT to enable correct data forwarding when the data chunk is returned.



(a) Interest blocking, timeout and retries

(b) PIT occupancy impact on the data buffer congestion

Fig. 2. Importance of the PIT occupancy in CCN.

can become a bottleneck when the downstream network is congested while interests continue to arrive at a node; the PIT would fill up quickly, old interests may timeout, and new interests can be blocked, leading to their retransmission and further congestion (see Fig. 2(a)). On the upside, however, as pending interests in the PIT are matched one-to-one with incoming data packets, the PIT occupancy constitutes a good estimator of the incoming data traffic load on the reverse path buffer (see Fig. 2(b)). To deal with the former and take advantage of the latter, a careful dimensioning of the PIT and study of its occupancy is of paramount importance.

In this paper,[4] we derive an analytical model of the PIT occupancy where we take into account the effects of interest aggregation, interest blocking, and interest timeout and retransmissions.

More precisely, our contributions are as follows:

- First, we conduct a simulation to study the filtering effects of in-network caching and interest aggregation on the traffic (see Fig. 2(a), by investigating the inter-arrival time of interests at different points in the network. We then invoke the main findings to substantiate our use of a Poisson process to model interest arrivals, which is central to the tractability of the model;
- Second, we propose an analytical model to estimate the interest blocking probability at a given node in the network, in the presence of interest aggregation, interest blocking and interest timeout and retries, using a two-dimensional CTMC. To solve the model numerically, we consider two approaches: *(i)* state-space truncations and *(ii)* state-space aggregation and approximation.
- Third, we build an accurate simulation model using the NDNSim to validate our analytical results via simulation experiments that draw interest in arrivals from non-Poisson distributions.

The rest of this paper is organized as follows. We provide a brief description of the components of CCN and the basic operation of the PIT in Section 2, then proceed with a simulation study of the effects of in-network content filtering in Section 3. Based on the results and their discussion, we tackle the model of the PIT at a CCN node in isolation in Section 4. We present some numerical results in Section 5, then after reviewing related work in Section 6 draw our conclusions in Section 7.

2. Effects of in-network content filtering

Poisson process has been widely invoked (e.g., [10–12]) to represent interest packet arrivals at a CCN node. In practice, this assumption is generally not valid, as the interest arrival process is a superposition of arrival processes that are not necessarily Poisson distributed. For example, consider a video-on-demand (VoD) system with multiple users. While it is reasonable to assume that the users of the VoD system arrive according to a Poisson process, the ensuing process of interest packets is driven by network protocols (including flow control [13], congestion control [14] and so on) and is not Poisson distributed in general. A simple model of such a process was studied in [15] where arrivals consist of primary events that arrive according to a Poisson process (in our case user arrivals) and secondary events that are generated at a constant rate after the corresponding primary event (in our case interest packet generated at a constant rate for example). Nevertheless, under certain conditions, the Poisson assumption is still reasonable, especially at the nodes that are most likely to suffer from PIT congestion such as core network nodes: (*i*) Thanks to content popularity being Zipf-distributed, there is a tremendous filtering effect that takes place at edge nodes close to the requester via content caching and interest aggregation. This results in content popularity and the arrival process of requests becoming more random in nature at the network core; (*ii*) the ratio of popular content to unpopular and one-timer content in the Internet is very small; this suggests that we can anticipate the remaining traffic load handled by the core network to be still non-negligible; (*iii*) finally, based on past clean slate redesigns (e.g., IPv6, IntServ, DiffServ, IPSec...) it is not expected that the whole Internet is changed overnight from its loosely hierarchical AS-based structure into a flat architecture, fully and exclusively based on CCN. Therefore even when it is fully under CCN there will still be stub ASes and a network core. This network would receive interests for non-popular content (i.e., that cannot be aggregated at the edge) and would handle data chunks that lead to little caching benefit in the core network if at all.

To verify these assertions, i.e., the random nature of arrivals in the core network and justify the Poisson-nature of arrivals at core nodes, we conducted a simulation study in realistic network topology, with non-Poisson arrivals and examined the interest inter-arrival times distribution at different nodes and depths in the network.

[4] This manuscript is an extended version of [8,9]

Table 1
Notations of variables as in Eq. (1).

Symbols	Meaning
L	Number of access routers
κ	Average size of a content (in chunks)
τ_n	Average RTT from node n to the content producer
λ_u	Average arrival rate of users per access router
λ_c	Average chunk request rate per user
t_u	Average time spent by a user before departure

3. Simulation study

We consider an access router where user arrivals follow a Poisson process with rate λ_u, and each user that arrives selects a particular content drawn from a Zipf content popularity distribution [16], with each content consisting of a randomly generated number of chunks (with mean κ). The chunks are requested by the user at a constant rate λ_c until all the chunks of this content are received. The superposition of all interest requests results in a process that comprises primary events (user arrivals) that follow a Poisson distribution triggering each a number of secondary events (Interest packet generation) that are deterministic relative to their primary event. Such arrival process and the corresponding queueing system have been studied in [15] and are known to be non-Markovian. In addition, in CCN, every user maintains a timer for each chunk request forwarded upstream, and upon timeout, the user retransmits the request with the same constant rate λ_c.

We estimate the average number of entries in the PIT, Γ_n, for a given node n in our network model as follows: Note that a user departs the network after receiving on average κ data chunks and the average time the user spends in the network before departure is $t_u = \kappa\lambda_c^{-1}$. The average number of users in the system, with L access routers within the time t_u is $L\lambda_u t_u$ each sending interest packets at a rate λ_c. If such requests were all served from the content producer, then the average number of requests in the PIT for node n would be $L\lambda_u t_u \lambda_c \tau_n$, τ_n being the Round-Trip Time (RTT) from node n to the content producer. Due to request aggregations (PIT hits, h_p) and caching (cache hits, h_c), not all requests will be served by the origin server, thus this is only a worst-case average number of PIT entries in the system. Given the variables as defined in Table 1, the actual average number of entries in the PIT for a node n is given by:

$$\Gamma_n = L\lambda_u t_u \lambda_c \tau_n (1 - h_c)(1 - h_p) \tag{1}$$

To avoid setting the size of the PIT to too small or too large in the case of finite PIT size, we use the worst-case average PIT size to guide the setting of the size of the PIT used in our simulation experiments.

3.0.1. *Methodology and simulation set-up*

We consider the network topology shown in Fig. 3. To generate this topology we extracted the graph for ISP Exodus, consisting of 157 nodes and 341 links, from Rocketfuel [17]; the nodes fall into one of three categories, access routers, gateways and core routers. Then to avoid consuming computational resources unnecessarily, we picked three nodes from among the access routers to play the role of content producers. Using Dijkstra's shortest path algorithm from each producer to the remaining access routers we obtained the paths used to forward all possible requests from the access network to the producers. We obtained the graph shown in the figure by superimposing the resulting paths and pruning the remaining links and nodes that have no impact on the simulation. (NB: since under light traffic the Poisson assumption is often valid, the three producers are chosen to result in a connected graph where all the paths to the content producers share a link, to subject it to the moderate-to-high traffic loads that are of interest).

Bottlenecked links in CCN can contribute to the congestion of the PIT and affect the caching performance. As our goal is to analyse the effect of in-network content filtering, we set the link capacities and delays as follows. The backbone-to-backbone links are set to 1 Gb/s with a 20 ms propagation delay; the backbone-to-gateway (including gateway to gateway) links are set to 0.5 Gb/s and 10 ms delay, and the gateway-to-access router links are set to 0.1 Gb/s and 5 ms delay. Each returned data chunk is 1500 bytes long. We use proWGen [18] to generate workloads for our simulation experiments with content popularity following Zipf-distribution with parameter α [16]. All content requests at each access router are from the same content popularity distribution. Other simulation parameters are shown in Table 2. The resulting network model was simulated in ns-3 network simulator using the ndn/ccn ns-3 modules [19] augmented with our custom application module that can use our traffic workload generated by ProwGen.

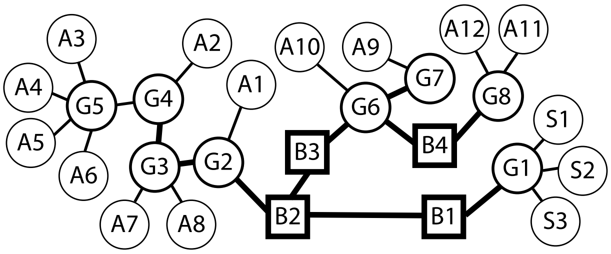


Fig. 3. Network topology (Extracted from Rocketfuel, Exodus ISP): S1-S3 are content servers, A1-A12 are access routers (acting on behalf of the requesters), G1-G8 are gateway routers and B1-B4 are backbone routers.

Table 2
Simulation parameters.

Parameter	Values
#requests per access router	10,000
Zipf skewness, α	0.75
#distinct content	4,200
Mean κ (and standard deviation) of content size, in chunks	193 (160)
Size of PIT (entries: one entry denotes one distinct interest)	infinite, finite (1000)
Cache size (chunks)	5% and 1% of universe content
Arrival rate of users, λ_u (users/sec/node)	10, 20
Chunk request sending rate, λ_c (chunks/user/sec)	10

Requested data are cached at every traversed router with caching capability. Given the limited router memory and cache memory may become full as such data are evicted from the cache to accommodate new data. The Least Recently Used (LRU) replacement policy has been used extensively in the context of CCN and other ICN contexts. Although there are other replacement policies such as Most Recently Used (MRU) and Most Frequently Used (MFU) in ICN, [20] shows that LRU shows no distinguishable performance from MRU and MFU. To this end, caching happens at every intermediate node and we use LRU as the replacement policy, similar to [21,22].

Note that there are no publicly available real CCN traffic traces at the time this work was carried out. As such we adopt in this paper the traffic characteristics described earlier in this section.

To study the effects of in-network chunk request filtering on the requests arrival process and the content popularity, we consider the cumulative distribution of the request inter-arrival times at each node. For the inter-arrival times, we fit the exponential distribution to the observed values. In addition, we present results showing the hit rates at the cache and PIT in the network. We also examine the squared coefficient of variation c_v^2 of the sampled data to reflect the nature of the inter-arrival time distribution at these nodes.

3.0.2. *Impact of cache size*

We show in Tables 3 and 5 the sample mean–variance, sample squared coefficient of variation, cache and PIT hit rates for a subset of

Table 3
Statistics for an infinite PIT size and cache size of 1% of universe contents with 10 users per second.

ID	Mean	Var	c_v	h_{cs} (%)	h_p (%)
B1	5.29E−05	3.35E−09	1.20E+00	0	0
B2	5.17E−05	3.20E−09	1.20E+00	1.83	0.35
G2	7.70E−05	6.87E−09	1.16E+00	1.15	0.16
G3	8.65E−05	8.68E−09	1.16E+00	2.1	0.38
G5	1.46E−04	2.43E−08	1.14E+00	4.13	0.39
G4	1.21E−04	1.68E−08	1.15E+00	1.57	0.17
A3	5.35E−04	3.27E−07	1.15E+00	8.33	0.07
A4	5.34E−04	3.30E−07	1.16E+00	8.21	0.07
A5	5.319e−04	3.280e−07	1.160e+00	8.37	0.10
A6	5.32E−04	3.14E−07	1.11E+00	8.34	0.09

Table 4
Statistics for an infinite PIT size and cache size of 5% of universe contents with 10 users per second.

ID	Mean	Var	c_v	h_{cs} (%)	h_p (%)
B1	6.87E−05	5.49E−09	1.17E+00	0.02	0
B2	6.59E−05	5.04E−09	1.16E+00	3.92	0.1
G2	9.76E−05	1.08E−08	1.13E+00	2.26	0.04
G3	1.08E−04	1.32E−08	1.13E+00	4.55	0.14
G5	1.77E−04	3.54E−08	1.13E+00	7.83	0.13
G4	1.51E−04	2.55E−08	1.12E+00	3.17	0.08
A3	5.34E−04	3.25E−07	1.14E+00	24.44	0.02
A4	5.33E−04	3.22E−07	1.13E+00	24.64	0.06
A5	5.325e−04	3.220e−07	1.136e+00	24.43	0.02
A6	5.33E−04	3.09E−07	1.09E+00	24.42	0.05

nodes in the network, for a cache size of 1% of the content universe. In Tables 4 and 6 we show the same metrics for a cache size of 5% of the content universe. From the results with infinite PIT size, a large cache size reduces the traffic load on core routers including the variance. This is due to the increase in cache hit rate with a large cache size. Due to interest packet retransmission in the case of finite PIT size, the traffic load is a bit higher with a larger cache size. This implies that an improper setting of the PIT size can make the benefits of caching unrealistic or insignificant. The impact on c_v^2 is not significant. In addition, core routers experience a lower PIT hit rate with increased cache size for the case of infinite PIT size, as more content requests are served by routers closer to the requesters. Similar results are observed if the PIT size is finite, as smaller cache sizes induce the retransmitted interest packets to travel farther into the core of the network than with larger cache sizes. For 1% content universe cache size, B1 has no cache hit. However, an increase in the cache size results in B1 having 0.02% cache hit. This is counter-intuitive as one would expect more cache at the edge would absorb even more requests. This is due to the difference in the PIT occupancy of B1 and B2 caused by increasing the cache size and the Zipf distributed content popularity.

Finite PIT size causes interest packets to be dropped when the PIT is saturated. This leads to retransmission of the lost interest packets. The nature of the original traffic and the retransmissions cause the c_v^2 of the access routers to be much greater than 1, making the arrival process of requests to access routers much burstier than Poisson. However, due to the filtering effect at the aggregation router G5, c_v^2 is observed to become closer to 1, which implies that the arriving process of requests at G5 is nearly Poisson distributed. This is evident in the CDF shown in Fig. 4 (focusing on (b) and (d)) and then in Fig. 5 as the number of hops away from the source increases.

3.0.3. Impact of traffic load

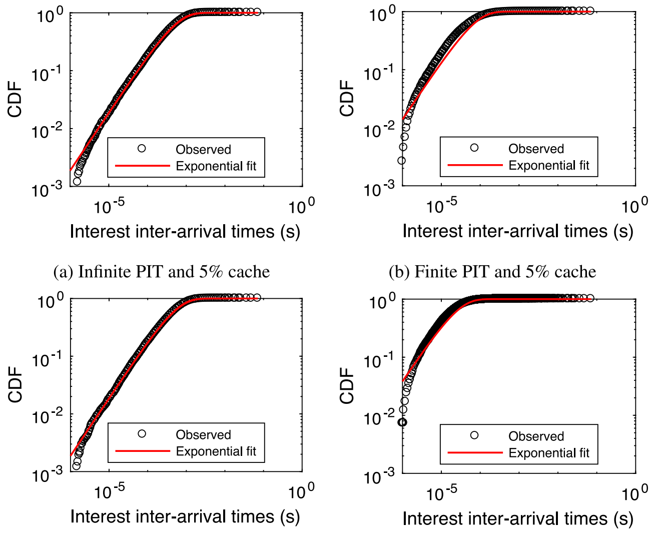
Next, we study the impact of the traffic load on the PIT hit rate and cache hit rate. We observe similar performance on cache hit rate for non-access routers when the arrival rate of users at the access routers is increased to 20 users per second. The results are shown in Table 7 for a finite PIT size of 1000 entries and cache size of 5% of the universe contents. We still ensure that the link capacities do not represent

Table 5
Statistics for a finite PIT size of 1000 entries and cache size of 1% of universe contents with 10 users per second.

ID	Mean	Var	c_v	h_{cs} (%)	h_p (%)
B1	7.62E−05	6.60E−09	1.14E+00	0	0
B2	7.45E−05	6.30E−09	1.14E+00	1.36	0.17
G2	1.33E−04	1.97E−08	1.12E+00	1.07	0.18
G3	1.63E−04	2.96E−08	1.12E+00	1.91	0.94
G5	6.28E−04	6.91E−07	1.76E+00	2.64	0.32
G4	3.30E−04	1.32E−07	1.21E+00	1.16	0.1
A3	2.61E−05	5.42E−09	7.93E+00	0.05	9.89
A4	2.44E−05	4.94E−09	8.31E+00	0.05	8.26
A5	2.643e−05	5.137e−09	7.355e+00	0.06	10.12
A6	2.44E−05	4.35E−09	7.32E+00	0.05	9.2

Table 6
Statistics for a finite PIT size of 1000 entries and cache size of 5% of universe contents with 10 users per second.

ID	Mean	Var	c_v	h_{cs} (%)	h_p (%)
B1	7.58E−05	6.66E−09	1.16E+00	0.03	0
B2	7.22E−05	6.04E−09	1.16E+00	4.01	0.09
G2	1.12E−04	1.48E−08	1.17E+00	2.47	0.12
G3	1.26E−04	1.87E−08	1.19E+00	4.56	0.36
G5	2.24E−04	7.72E−08	1.54E+00	6.46	0.19
G4	1.88E−04	4.44E−08	1.26E+00	3.47	0.1
A3	4.58E−05	3.19E−07	1.52E+02	0.71	7.46
A4	7.44E−05	2.30E−08	4.16E+00	1.48	13.54
A5	6.372e−05	1.801e−08	4.436e+00	1.05	11.03
A6	7.00E−05	1.95E−08	3.98E+00	1.29	12.31



(a) Infinite PIT and 5% cache

(b) Finite PIT and 5% cache

(c) Infinite PIT and 1% cache

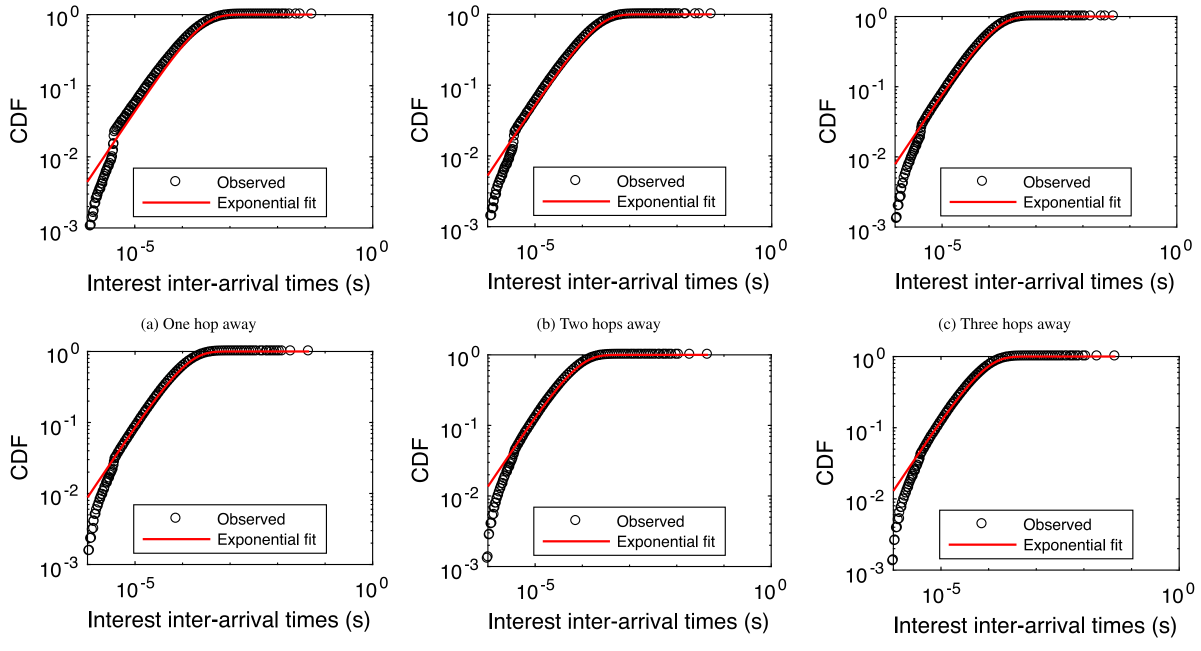
(d) Finite PIT and 1% cache

Fig. 4. CDF of the interest packet inter-arrival times at access routers.

bottlenecks in the network. Increasing the user arrival rate decreases the PIT hit rate in the case under consideration especially for the access routers, thus increasing the traffic load at the core routers. In addition, high caching dynamics are observed at the access routers, evident by the decrease in the observed cache hit rates when we increase the traffic from 10 users per second (see Table 6) to 20 users per second (see Table 7). The finite PIT size considered in this case does not allow us to fully reap the benefits of interest aggregation in the presence of a high traffic load at the access router.

3.1. Discussion

It is well known that due to the skewness of the content popularity distribution and the caching that takes place in CCN as well as interest



(a) One hop away

(b) Two hops away

(c) Three hops away

(d) Four hops away

(e) Five hops away

(f) Six hops away

Fig. 5. CDF of the interest packet inter-arrival times at "aggregation" routers: G5, G4, G3, G2, B2, B1, respectively, in Fig. 3.

Table 7
Statistics for a finite PIT size of 1000 entries and cache size of 5% of universe contents with 20 users per second.

ID	Mean	Var	c_v	h_{cs} (%)	h_p (%)
B1	7.632E−05	6.863E−09	1.178E+00	0.04	0.00
B2	7.216E−05	6.095E−09	1.171E+00	4.07	0.10
G2	1.202E−04	1.708E−08	1.183E+00	3.75	0.09
G3	1.621E−04	3.516E−08	1.338E+00	5.74	0.13
G5	4.038E−04	2.367E−07	1.452E+00	9.42	0.26
G4	3.362E−04	1.488E−07	1.317E+00	4.74	0.09
A3	2.440E−05	4.116E−09	6.911E+00	0.45	3.96
A4	2.352E−05	4.393E−09	7.939E+00	0.41	3.77
A5	2.356E−05	3.664E−09	6.598E+00	0.41	4.26
A6	2.363E−05	3.504E−09	6.278E+00	0.41	3.89

aggregation, heavy interest packet filtering takes place at access nodes as shown in Tables 4 to 7, and only those interests that are not cached in the access nodes or aggregated in the PIT would reach the rest of the network. In fact, many opponents of systematic caching in CCN use this argument to justify the need for one to two levels of caching in the network instead. The takeout of our simulation results in Section 3 is that the deeper we go into the network towards a content origin server the closer the interest inter-arrival times get to an exponential distribution.

Assuming that downstream and upstream nodes do not alter the arrival process of requests at a given node, we can approximate the performance of the system by analysing the PIT of a single node in isolation using a Markovian arrival process. We represent the PIT of a single node by the system shown in Fig. 2(a). We do not include the content store and the forwarding information base in the system as they are out of the scope of this work. An entry in the PIT is created for every new interest that arrives and experiences cache and PIT misses at a CCN node, provided that the PIT is not full. Conversely, no entry is created in the PIT when the PIT is saturated, in which case the interest is blocked/dropped. Each entry that is created in the PIT has an associated timer and the entry is removed from the PIT when the timer expires. The entry can also be removed if the requested data chunk is received before the entry timer expires. Blocked interest and timeout entries (one entry per unique interest) can be retransmitted by the source.

We model the system presented in Fig. 2(a) by a 2-dimensional Markov process (P_t, D_t), where P_t represents the content of the PIT at time t and D_t represents the number of Interests pending retry at time t. This system can be represented by a queueing model where; *(i)* the PIT is modelled as a blocking system with M servers (PIT places), denoted by P in Fig. 6; *(ii)* interests blocked or timeout will (eventually) be retransmitted by their respective sources later. To model this, each blocked or timeout interest enters a delay system (an infinite number of servers), where it waits for a certain time before it is redirected to the input of the PIT. Note that queue D does not exist in reality, it only represents the blocked interests that are pending retransmission by their respective sources. The details of the analytical model are given below in Section 4.

4. Analytical model of the PIT occupancy

In this section, we discuss the analytical model for PIT occupancy in detail.

4.1. Model description and assumptions

Let a Pending Interest Table with a total number of M entries. The PIT receives from N consumers interest packets. We assume that, for each consumer, the new arrivals of interests at CCN nodes happen after the content store is filtered following a Poisson process with the rate λ. That is the new interest arrival rate at the CNN node is $N\lambda$. We show in Fig. 6 a simple diagram modelling a queue system for M-servers with no waiting room. And, the arrival process in the system follows a Poisson process with rate $N\lambda$. Let h_p be the hit rate of the PIT which indicates the rate at which interests with the same name are matching interest arrivals. These matching hits are filtered out and only new interests are used for creating a new entry with probability P_b where P_b is the probability that an interest arrives at a full PIT. This case occurs with a rate of $\lambda' = (1 - h_p)N\lambda$, which is represented by the action of the aggregator element A in Fig. 6.

Different nodes in the network may be serving the same interest at different times, depending on caching status, content popularity, and incoming requests for the same block of data. So it seems that the time

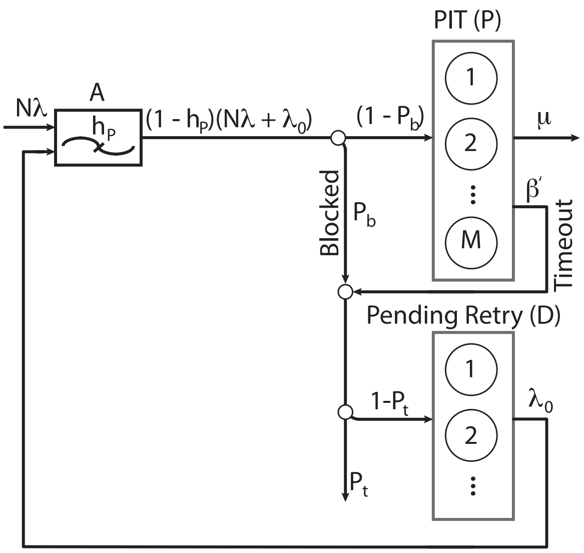


Fig. 6. System model.

between transferring interest upstream and receiving the requested data should in practice be modelled by a phase distribution of rate μ.[5]

New or resubmitted Interests blocked in the PIT with probability P_b and pending Interests that expired and timed out are placed in an infinite queue D so that they can be retried. We wait for the resend timer with probability $1 - P_t$. In other words, the consumer chooses not to resend interest with probability P_t, and the re-sent interest is paid at rate $\alpha = (1 - h_p)\lambda_D$ reaches P.

Pending interest (entries) whose timer expires leave queue P at rate $\beta' = 1/T$ where T is the PIT entry timeout. Therefore, we assume that the time t between the creation and deletion of PIT entries follows a truncated exponential distribution where the percentage of requests that time out is $\eta = 1 - e^{-\mu T}$ and wait for $t > T$.

4.2. Continuous time Markov chain model of pending interest table

The state transition diagram of the system model assuming an infinite retry queue D and M servers (PIT places) is shown in Fig. 7. We define μ' and β as

$$\mu' = (1 - \eta)\mu + \eta\beta' P_t \tag{2}$$

$$\beta = (1 - P_t)\eta\beta' \tag{3}$$

Each state in the model can be represented by a 2-tuple (i, j) where $i = 0, \ldots, M$ is the number of pending interests in the PIT and $j = 0, \ldots$ the number of interests waiting for the expiry of the retry timeout. Events that trigger transitions fall into one of the following categories as presented in Fig. 7:

- **Event 1:** Arrival of a new interest.
- **Event 2:** A returned data packet consumes the pending interest before its entry's timeout expires in the PIT.
- **Event 3:** The PIT entry is queued in the retry queue D due to the expiry of its timeout.
- **Event 4:** A re-transmission of a timeout or blocked interest from the retry queue D.

The rates and the events that trigger the transitions in Fig. 7 are as follows:

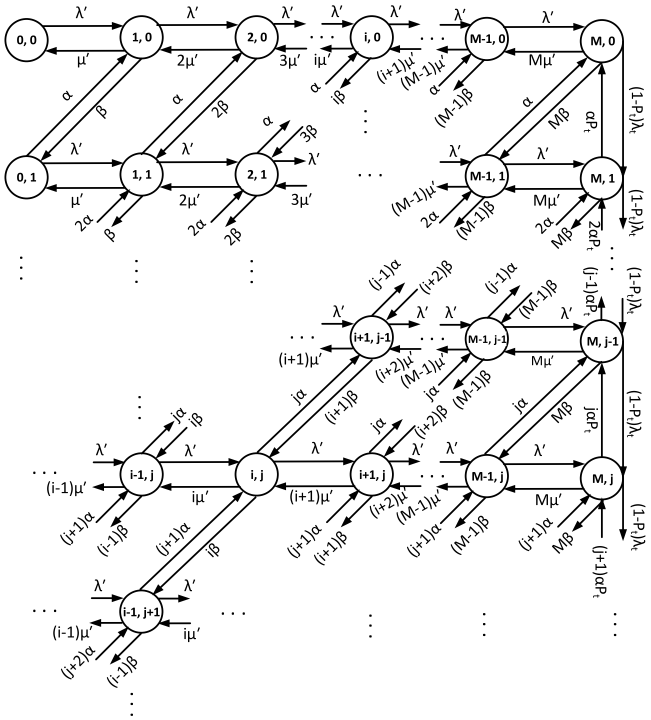


Fig. 7. Transition diagram of CTMC model for an infinite number of blocked and timeout interests.

- Event 1 takes place and
  - the PIT is non-full: Corresponds to $(i, j) \to (i + 1, j), 0 \le i < M, 0 \le j < \infty$ and happens at a rate λ'
  - the PIT is full: Corresponds to $(M, j) \to (M, j + 1), 0 \le j < \infty$ and happens at a rate $(1 - P_t)\lambda'$
- Event 2 takes place: Corresponds to $(i, j) \to (i-1, j), 1 \le i \le M, 0 \le j < \infty$ and happens at a rate $i\mu'$
- Event 3 takes place: Corresponds to $(i, j) \to (i - 1, j + 1), 1 \le i \le M, 0 \le j < \infty$ and happens at a rate $i\beta$
- Event 4 takes place and
  - the PIT is non-full: Corresponds to $(i, j) \to (i + 1, j - 1), 0 \le i < M, 1 \le j < \infty$ and happens at a rate $j\alpha$
  - the PIT is full: Corresponds to $(M, j) \to (M, j - 1), 1 \le j < \infty$ and happens at a rate $j\alpha P_t$

Let $\Pi_{i,j}$ be the probability that the system is in state (i, j). By writing the balance equation based on the model shown in Fig. 7 and the normalization condition $\sum_i \sum_j \Pi_{i,j} = 1$, the system of equations can be numerically solved according to the equilibrium state probability. From $\Pi_{i,j}$, we can compute the interest rate blocking probability P_b^t using a system of equations in an iterative manner.

The interest blocking probability is therefore obtained as:

$$P_b^t = \sum_{i=M, j\ge 0} \Pi_{i,j} \tag{4}$$

where P_b^t is P_b after truncating the state space for queue D at a state with a reasonably negligible probability.

4.3. Approximate continuous time Markov chain model of the PIT

To solve the Markov chain of in Fig. 7 and described in Section 4.2, we reduce the state space by truncation, an approach that is associated with several shortcomings such as state space explosion (with a size of $(M+1)(N+1)$ where M is the PIT size and N is the maximum number of

[5] Estimating the parameters of the distribution requires a thorough investigation of real networks (which do not yet exist).

blocked and timeout interests) and the model solution being sensitive to the value of N. To overcome these shortcomings, we propose to use an approximate CTMC, an approach invoked by Marsan et al. in [23] cellular networks.

To form a solution, we represent the state of the system as a 2-tuple (i, d) where i is the same as previously defined (i.e. the number of PIT entries with outstanding interest). However, now d is introduced as a boolean variable representing whether the queue D has a blocked interest awaiting retry. $d = 1$ if the retry queue D is not empty, otherwise $d = 0$. In this case, the state space is greatly reduced to $2(M + 1)$ states.

We introduce $\overline{N}_D$ to represent the average number of blocked/timed out interests in the system. Then, the rate at which blocked/timed out interests are resent is $\overline{N}_D \lambda_D$. Similar to [23], we define the probability P as the probability that a blocked or timed-out interest does not empty the queue D. Therefore, P can be used to estimate $\overline{N}_D$. We estimate $\overline{N}_D$ and P following the approach used in [23]. However, we differ from their approach since we are accommodating timed-out interest which goes into queue D.

Now, the state transition diagram of the approximate CTMC model with M entries and $d \in \{0, 1\}$ is shown in Fig. 8. We define μ' and β as in (2) and (3), respectively. We also define two rates of retry from queue D: 1. ω is defined as the average conditional rate of retry from queue D given that queue D is empty (i.e., that leaves queue D empty, i.e., $\omega = (1 - P)\overline{N}_D \lambda_D$); and 2. γ is the average conditional rate of retry from queue D given that queue D is non-empty (i.e., that leaves queue D non-empty, i.e. $\gamma = P\overline{N}_D \lambda_D$).

We detail, as shown in Fig. 8, the rates and events that trigger the transitions as follows:

- A new interest arrives and queue D is empty: Corresponds to $(i, 0) \rightarrow (i + 1, 0), 0 \leq i < M$ and happens at a rate λ'
- An interest arrives and queue D is non-empty: Corresponds to $(i, 1) \rightarrow (i + 1, 1), 0 \leq i < M$ and happens at a rate $\lambda' + \gamma$
- A pending interest is consumed by a returned data packet before the corresponding PIT entry's timeout expires and queue D is empty: Corresponds to $(i, 0) \rightarrow (i - 1, 0), 0 < i \leq M$ and happens at a rate of $i\mu'$
- A pending interest is consumed either by a returned data packet or the PIT entry's timeout expires and queue D is non-empty.: Corresponds to $(i, 1) \rightarrow (i - 1, 1), 0 < i \leq M$ and happens at a rate of $i(\mu' + \beta)$
- A PIT entry's timeout has expired while no blocked/timeout interest in queue D. The timeout interest is queued in the retry queue D: Corresponds to $(i, 0) \rightarrow (i - 1, 1), 0 < i \leq M$ and happens at a rate of $i\beta$
- A new interest arrives when the PIT is full and queue D is empty: Corresponds to $(M, 0) \rightarrow (M, 1)$ and happens at a rate of $(1 - P_t)\lambda'$
- The last interest in queue D is retransmitted causing queue D to become empty: Corresponds to $(i, 1) \rightarrow (i + 1, 0), 0 \leq i < M$ and happens at a rate of ω
- The last interest in queue D is retransmitted when the PIT is full and the blocked/timeout interest decides not to enter retry queue D. The timeout or blocked interest is retransmitted from the retry queue D: Corresponds to $(M, 1) \rightarrow (M, 0)$ and happens at a rate of $P_t\omega$

Let $\Pi_{i,d}$ denote the steady-state probability that the system is in state (i, d). To obtain the model solution, we express P and $\overline{N}_D$ as in (5):

$$\overline{N}_D = \frac{P}{1 - P} \tag{5}$$

where P is given as

$$P = \frac{\lambda' \sum_{i=0}^{M} \Pi(i, 1)}{\lambda' (\sum_{i=1}^{M} \Pi(i, 0) + \sum_{i=0}^{M} \Pi(i, 1))} \tag{6}$$

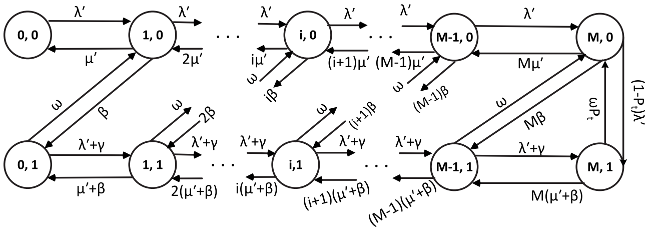


Fig. 8. Transition diagram of CTMC model for the existence of a blocked/timeout interest.

$$= \frac{\sum_{i=0}^{M} \Pi(i, 1)}{\sum_{i=1}^{M} \Pi(i, 0) + \sum_{i=0}^{M} \Pi(i, 1)} \tag{7}$$

We use a similar iterative method used in [23] to compute the values of P and $\overline{N}_D$. The method stops when a level of the relative accuracy of 10^{-6} is reached.

We apply the normalization condition, $\sum_i \sum_d \Pi_{i,d} = 1$ and solve the system of equation numerically for the equilibrium state probabilities $\Pi_{i,d}$, from which we can calculate the interest blocking probability P_b^a defined as:

$$P_b^a = \sum_{d \in 0,1} \Pi_{M,d} \tag{8}$$

where P_b^a is P_b for the approximation method.

4.4. Model complexity

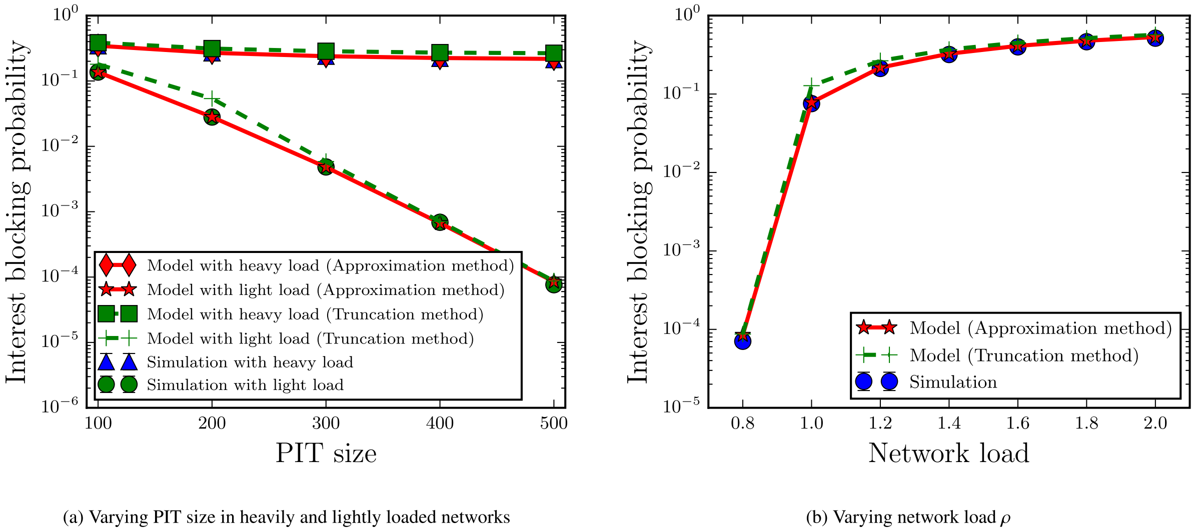
On the state space complexity, our proposed model with a retry queue D of infinite capacity (truncation model) has $(M+1)(N+1)$ states in the Markov chain where N is the size of truncated retry queue D and M is the size of the PIT. On the other hand, the approximate model has $2(M + 1)$ states in the Markov chain. The latter model requires a significantly reduced amount of memory to store the infinitesimal generator matrix. Furthermore, as discussed in [24], solving a system with S states using for example Gaussian elimination or LU factorization requires an $\mathcal{O}(S^3)$ operation. The state aggregation method is obviously much less complex. To strike a balance between accuracy and speed in our numerical examples, we set the accuracy level to 10^{-6} and observe the number of iterations to be in the range [500, 20000].

5. Numerical results

In this section, we will first validate our models via simulation using a realistic network topology with realistic traffic including a comparative analysis between the truncation method presented in Section 4.2 and the aggregation method in Section 4.3. For the methodology and simulation description and set-up, please refer to Section 3.

We consider two scenarios in our simulation experiments:

- **Scenario 1:** Users send requests for contents through an access router. The router forwards the requests to upstream routers. Each request is satisfied by an upstream router with a uniform probability. Interest packets arrive at the access router at a rate of 3000 interests per second and we set the interest lifetime to 1s. To avoid the impact of any hidden factors in our simulation, we set the link capacities such that they do not contribute to the PIT congestion. This scenario is used to verify and validate the accuracy of our model.
- **Scenario 2:** The network topology considered in this second scenario is shown in Fig. 3. Traffic and user characteristics are as described in Section 3. All nodes in the network, except one (default is node B2), have an infinite PIT size. Each node has a cache of size 1% of the content universe. Node B2 has a finite PIT size. We also use this scenario to validate our models.



(a) Varying PIT size in heavily and lightly loaded networks

(b) Varying network load ρ

Fig. 9. Scenario 1: Interest blocking probability for different PIT sizes and network loads.

5.1. Model validation and comparison

5.1.1. Using scenario 1

For the simulation results, we report the 95% confidence intervals of the average interest-blocking probability over several simulation runs. We collected the interest-blocking probabilities at the access router for different values of the PIT size and network load. We consider different values of the PIT size in the simulation from 100 to 500 for heavily and lightly loaded networks.

Fig. 9(a) shows how the interest blocking probability varies with the increasing size of the PIT in both heavy load and light load networks. As we expect, the blocking probability decreases with increasing PIT size for both the simulation and model. For Scenario 1 that we consider here, our models achieve over 90% accuracy compared to the simulation in heavily and lightly loaded networks.

In addition, we analyse the impact of the network load on the interest-blocking probability by increasing the network load ρ from 0.8, 2.0 for a fixed value of the PIT size, 500 in this case. As shown in Fig. 9(b), we observe that increasing the load in the network leads naturally to a higher probability of blocking. Interestingly, however, the results from our model match the simulation with over 90% accuracy for all loads. Furthermore, the approximate CTMC model via states aggregation seems to give slightly better accuracy than the truncated CTMC model, while being computationally less costly, which makes the case for the former approach. Notice that for mathematical tractability, the service time distribution in the PIT, i.e., the time from when an interest is forwarded until it is consumed by the returned data, was assumed to be exponential in the two models. In contrast in the simulation of Scenario 1, it is taken to be uniform. The accuracy of the results given in Figs. 9(a) and 9(b) hint that the blocking probability is not sensitive to the service times distribution but rather depends on its average.

5.1.2. Using scenario 2

To see how well the results from our model match simulation results in a more complex network topology with a packet-level simulation, we simulate Scenario 2. We consider a target node B2 in Fig. 3 and collect the interest-blocking probabilities for different sizes of the PIT in heavily and lightly loaded networks. We vary the PIT size in the range $[1000, 1400]$ and the rate at which users arrive on the network (10 and 20 users per second). In this scenario, we also report the 95% confidence intervals of the average interest-blocking probability over several simulation runs. Clearly, Fig. 10 shows that increasing the size of the PIT decreases the interest-blocking probability. The figure further shows the accuracy of our model including its independence on the service times distribution as contents can be fetched from any nodes beyond B2 up to the content producer. On the comparative performance of our CTMC models, similar behaviour to Figs. 9(a) and 9(b) is observed.

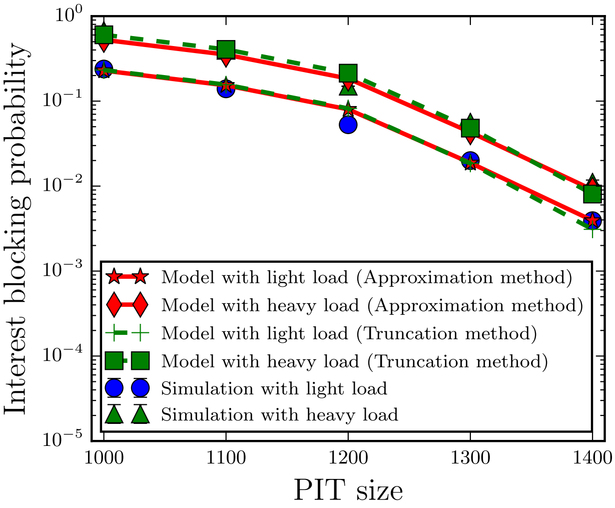


Fig. 10. Scenario 2: Interest blocking probability versus PIT size under heavily and lightly loaded networks.

5.2. Impact of PIT size on the cache and PIT hit rates

To see how the size of the PIT affects the hit rates at the cache and PIT, Fig. 11 shows the PIT size against the cache and PIT hit rates. Increasing the PIT size yields higher cache hit rates at the core router. This is due to a decrease in the interest packet blocking probability. It also shows that there is little or no impact on the cache hit rate as the PIT size reaches the maximum occupancy. The PIT hit rate also increases as we increase the PIT size.

5.3. Impact of PIT entry timeout

Retransmitted interests can lead to more interest packets that are blocked especially if the requester(s) does(do) not slow down the interest sending rate(s). We use our proposed approximate model to

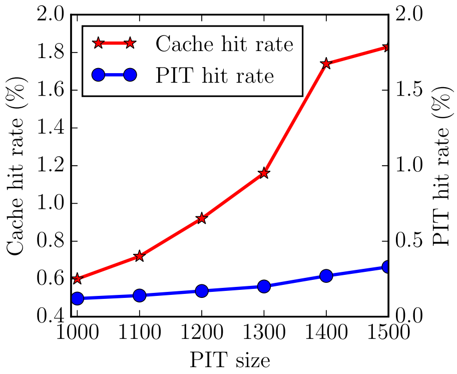


Fig. 11. Effect of PIT size on the cache hit rate and PIT hit rate from simulation using Scenario C.

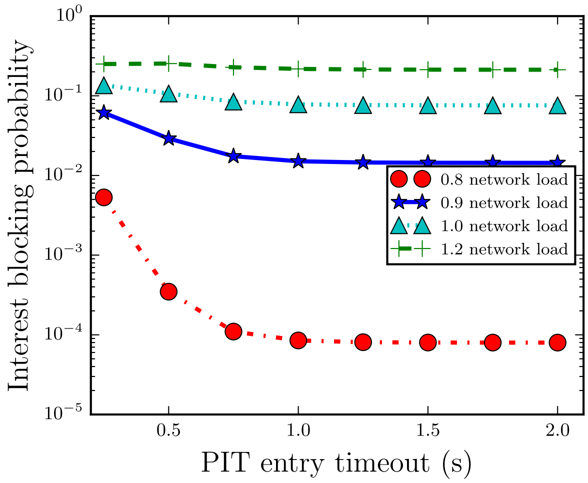


Fig. 12. Interest blocking probability versus PIT entry timeout under different loads in the network.

show the impact of the PIT entry timeout on the interest blocking probability in Fig. 12 under different loads in the network. In all the network loads considered, the interest blocking probability decreases with an increasing timeout value. The decrease factor depends on the traffic load in the network.

As shown in Fig. 12 high network loads result in little or no impact of the PIT entry timeout on the interest blocking probability while low network loads accentuate the effect of increasing the timeout on the blocking probability. In the case of high network loads, most of the interests that are blocked are due to the PIT being saturated, making the impact of timeout insignificant. On the other hand, most of the blocked interests are due to interests that time out and are later retransmitted in the case of low network loads. In Fig. 12, our model results suggest that setting the timeout too small in low network loads can lead to a significant number of interests that are blocked. Setting the timeout too high may not be desirable as the PIT may become full when interests take too long time to return data if at all [25].

Next, we analyse the validity of our models for different nodes at different locations in the network while considering the impact of interest retransmission, different distributions of the per-user interest inter-arrival times and the impact of larger cache size. For the remainder of our results, we consider the lightly loaded scenarios where the user arrival rate is 10 users per second.

5.4. Impact of interest retransmissions

Interest retransmission can cause the PIT to be more congested leading to an increased blocking probability. We use our model to substantiate this assertion including validation with simulation results.

In Fig. 13, we present results showing the interest-blocking probability as we increase the size of the PIT for different nodes in the network. It can be observed in Fig. 13a and Fig. 13b that the impact of interest retransmission on the interest blocking probability becomes insignificant at a PIT size greater than 1000 while in Fig. 13c and Fig. 13d the impact becomes less obvious at PIT sizes beyond 800 and 900, respectively. This is because increasing the PIT size reduces the blocking probability and consequently reduces the rate of interest retransmissions.

In addition, results from the approximate aggregation model can be observed to match the simulation results better than the ones from the truncation model as the PIT size increases. With all the PIT sizes considered, our models are more accurate for nodes, for example, B2 at 5 hops away from the consumer, that are several hops away from the access router than the nodes are just 1 or 2 hop(s) away. This is because of the in-network content filtering effects discussed in Section 2.

5.5. Impact of per user inter-arrival time distributions

In Section 3, we assume a deterministic per-user interest inter-arrival times. The goal of this section is to show how accurate are our models if we change this assumption. To this end, we consider 2 other distributions of the per-user interest inter-arrival times, uniform and exponential. Similarly, we consider different nodes at different numbers of hops away from the access routers in the network.

Fig. 14 shows results from simulation and our analytical models. For all the nodes and distributions considered our model results match well the results from the simulation for the relatively small size of the PIT. However, as the size of the PIT becomes large (≥ 1200 in Fig. 14d, Fig. 14e and Fig. 14f; ≥ 900 in Fig. 14g - Fig. 14l). We observe at only node B2 for all the 3 distributions that our model results match the simulation results to a reasonable level of accuracy for each of the sizes of the PIT considered. This substantiates again our claims on the effects of content filtering effect discussed in Section 2.

5.6. Impact of cache size

Caching in CCN is aimed at bringing popular content closer to the users, thus reducing the traffic load at core routers and the content origin servers. In this case, we expect that the interest blocking probability decreases with the increasing size of the cache.

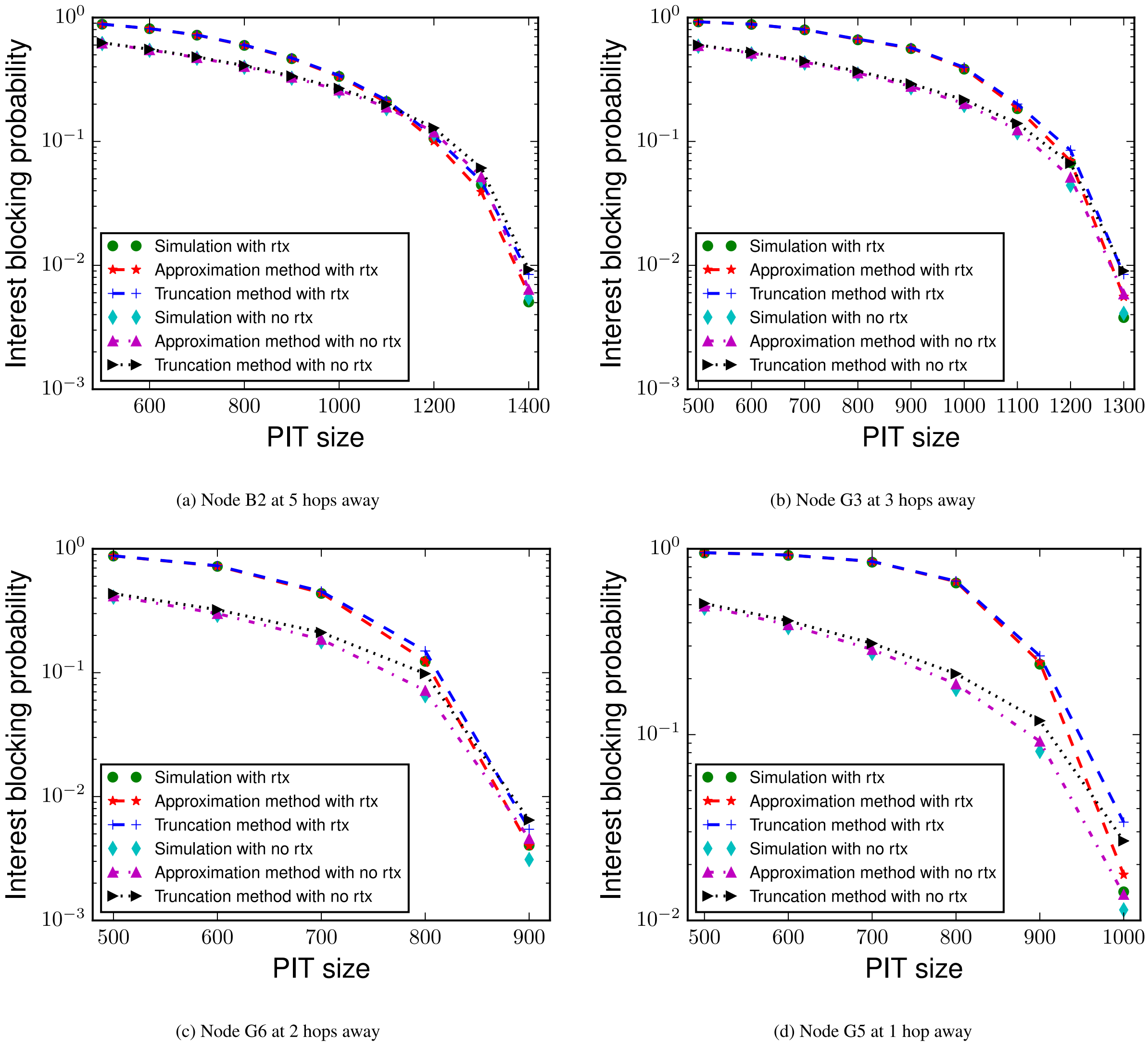
We increase the size of the cache from 1% to 5% of the universe content and ran simulation experiments for different sizes of the PIT. Fig. 15 shows results for interest blocking probability with increasing PIT size for different nodes at the different number of hops in the network. Note that, when the PIT size is greater than 1100 for node B2 results in no interest being blocked at the PIT as traffic load at B2 reduces for a cache size of 5% universe content. See Fig. 15a. Similar observations can be made for other nodes, G3, G6 and G5.

Due to the effect of in-network content filtering, results from our analytical models match simulation results for B2 and G3 better than the results for G6 and G5 which are 2 and 1 hop(s) away from the access routers.

5.7. Complexity

The solution complexity of our approximate model is significantly lower than the complexity of obtaining the solution of the truncation model. On average over all the points on the curves presented in this paper, the number of iterations until convergence with the approximate CTMC model is smaller than the truncated CTMC model.

The mean CPU times taken by both the approximate and truncation CTMC models to reach convergence are shown in Fig. 16. We also show the standard error of the mean quantifying how accurate is the mean.



(a) Node B2 at 5 hops away

(b) Node G3 at 3 hops away

(c) Node G6 at 2 hops away

(d) Node G5 at 1 hop away

Fig. 13. Interest blocking probability with or without interest retransmissions for different nodes at different depths in the network.

6. Related work

Over the past few years, efforts have been directed towards a comprehensive performance evaluation of CCN caching, routing and forwarding, security and transport under different network conditions. More importantly, how to avoid the PIT from becoming a bottleneck in CCN networks still remains an interesting but challenging research problem. We group existing works on the performance analysis and occupancy management of the PIT into two categories: (1) PIT occupancy and entry lifetime management and (2) PIT analytical modelling. In this section, we review the existing works under these categories.

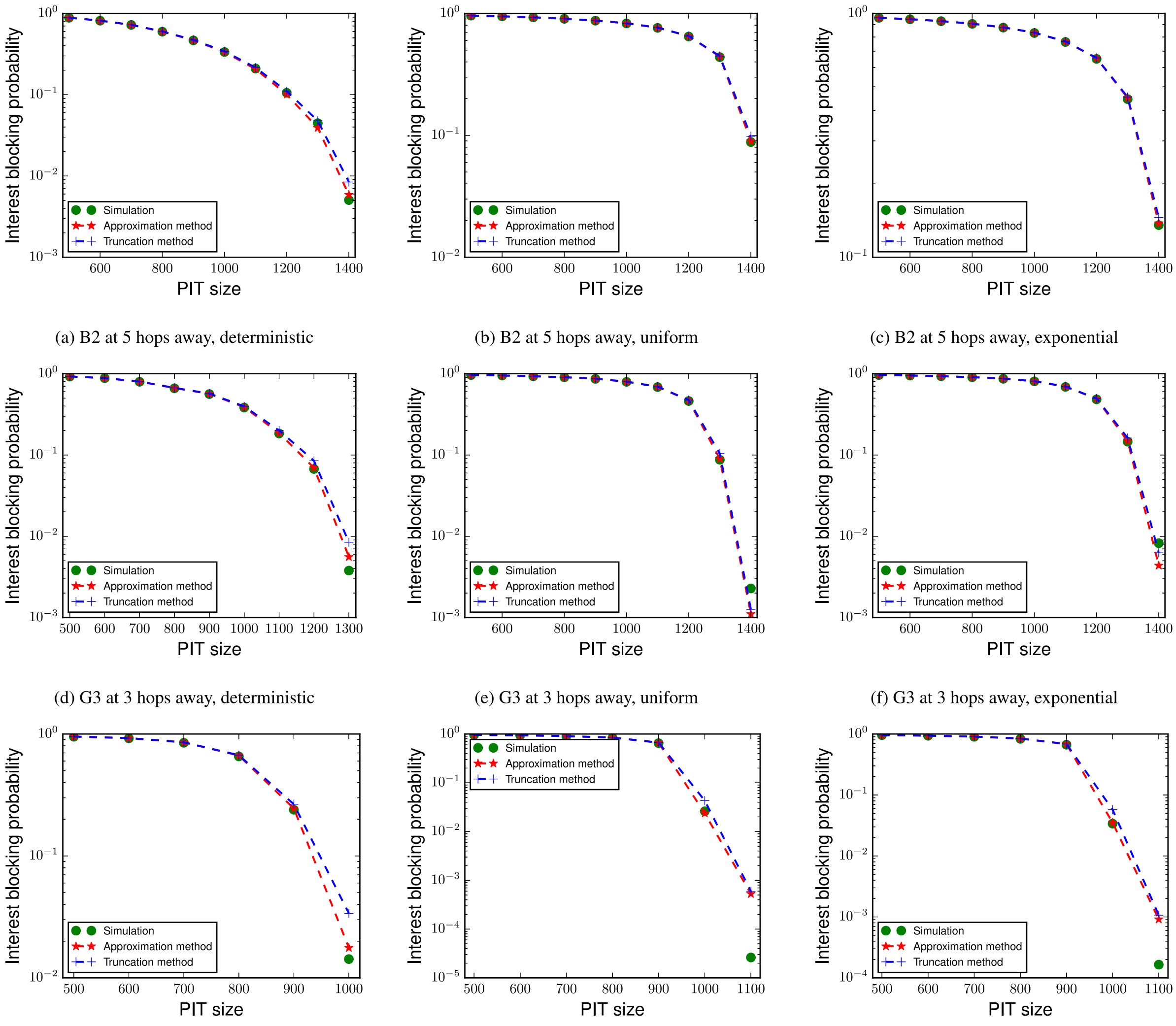
6.1. PIT occupancy and entry lifetime management:

A comparative analysis of the existing PIT architectures (namely; SimplePIT, HashedPIT and DiPIT) under heavy traffic load was carried out by Virgilio et al. [26]. Simulation results reveal that all three architectures are adversely affected, making the case for the need for better PIT occupancy and entry lifetime management. In view of this, several works have employed two different approaches: Fixed [26,27] and dynamic [13,25,28–31] PIT entry lifetime.

The fixed value approach is simple but oblivious of the network conditions such as delay and packet loss. For the dynamic value approach, Kazi and Badr propose a novel method for estimating the PIT entry lifetime at routers and the interest packet-timeout at receivers [28]. However, this approach assumes both interests and data chunks traverse the full diameter of the network. This assumption is indeed not realistic as in-network caching which enables interest packets to be satisfied by intermediate routers in the network is one of the selling points of CCN.

In [29], Safdar et al. propose a dynamic PIT entry lifetime (PEL) for NDN in vehicular networks using interest satisfaction rate and hop count to adjust the PEL. The authors derive a decay rate function for PEL with parameters, namely, the decay constant and the initial decay value. Another notable recent work looks into PIT management in the context of NDN-based VANETs [32]. A similar effort was recently dedicated to Named Data Networking of Things (NDNoT) [33]. Simulation results show that the proposed mechanism achieves its objectives but initial decay values greater than 0.5 are not considered. An initial decay value that is close to 1 may give a similar performance to the mechanism proposed in [25]. Our work uses the maximum data chunk response time observed within an interval of time. Simulation results show the efficiency of this approach as compared to using a fixed PIT entry lifetime.

Ravindran et al. employ a different strategy for managing the PIT occupancy by using the idea of traffic differentiation where non-shareable traffic (one-timer content) bypasses both the content store and the PIT while shareable traffic (always cached content) follows the



(a) B2 at 5 hops away, deterministic

(b) B2 at 5 hops away, uniform

(c) B2 at 5 hops away, exponential

(d) G3 at 3 hops away, deterministic

(e) G3 at 3 hops away, uniform

(f) G3 at 3 hops away, exponential

(g) G5 at 1 hop away, deterministic

(h) G5 at 1 hop away, uniform

(i) G5 at 1 hop away, exponential

Fig. 14. Interest blocking probability with different distributions of the per-user interest inter-arrival times for nodes: B2, G3, G6 and G5 in Fig. 3.

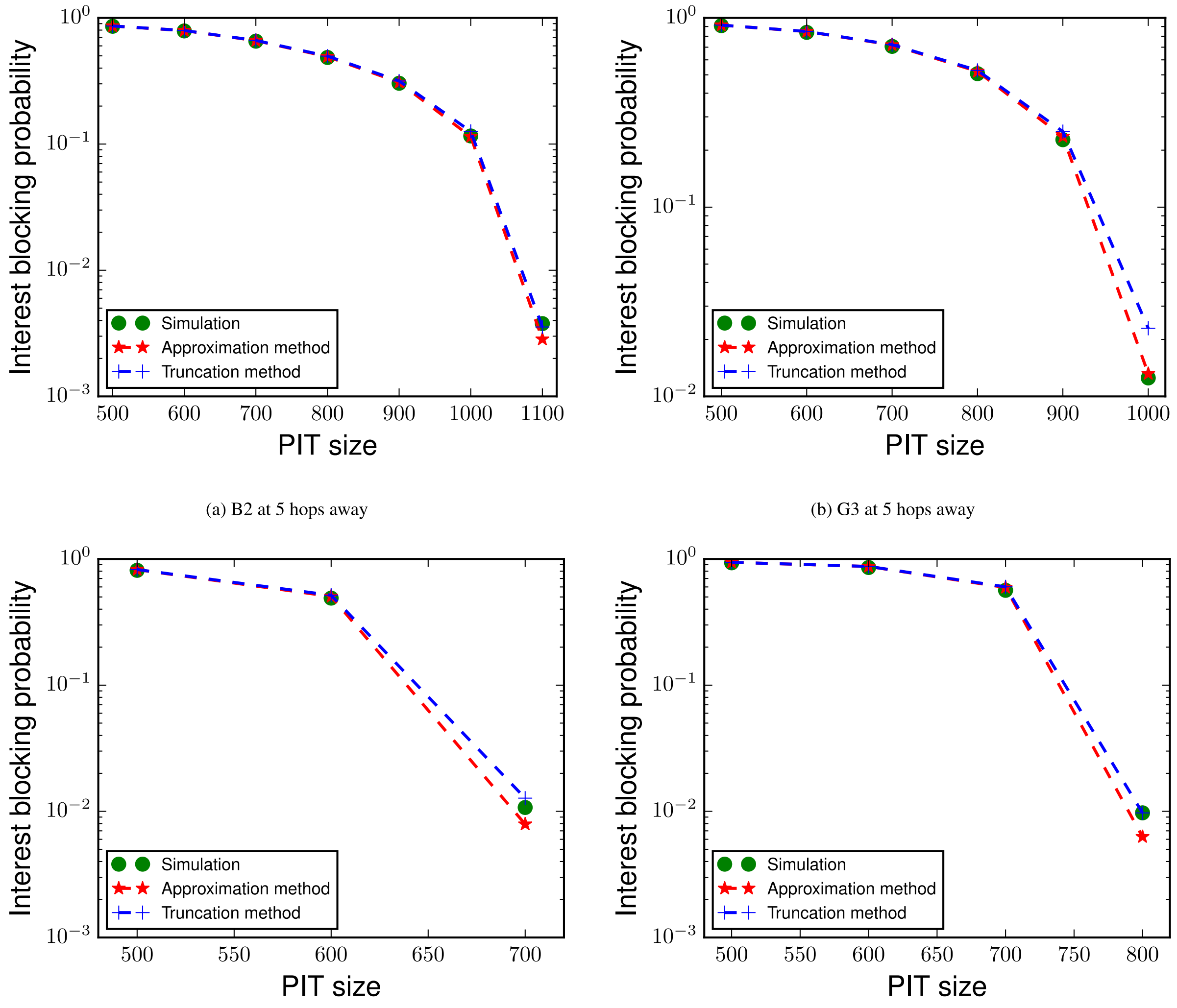
conventional CCN processing [34]. Yalei et al. advance the management of the PIT entry lifetime proposed in [25] to include strategies for storing RTT measurements and PIT entry replacement in the event that the PIT is saturated. In particular, the authors propose a new PIT entry replacement strategy that prioritizes the entries based on how long they have been pending in the PIT. Simulation results show performance improvement over default strategies used in CCN and NDN for PIT entry management. Recent work et al. propose different Bloom Filter-based PIT architectures for forwarding and dynamic allocation of the PIT [35,36]. Another work proposes reinforcement learning-based forwarding in NDN by leveraging online learning to adapt its forwarding decision according to the state of the Pending Information Table (PIT) [37]. Another work proposes, iCAFE, an efficient traffic control algorithm for VANET-based CCN that actively updates the FIB and PIT for effective communications [38].

6.2. PIT analytical modelling

Wang et al. proposes an analytical model of the denial-of-service attack against the PIT to derive the probability that the PIT is full [39]. The authors do not consider a scenario where the PIT becomes full and interests are dropped in the absence of a DDoS attack. This scenario does exist in practice when the rate at which distinct (non-aggregated) interest packets arrive at the PIT is greater than the rate at which the PIT entries are deleted. Another scenario that has not been addressed by Wang et al. is the case when the dropped or timed-out interests may be retransmitted either by the router or the requester [25,30]. Finally, the impact of interest aggregation has not been considered in [39]. More recently, queueing theory is used to propose a new model for transferring content in ICN under the assumption of bursty requests [40]. In this work, the mathematical model is driven for calculating cache and PIT miss rate and is verified via a comparison of the analytical model and the simulation experiments.

To estimate the average and maximum PIT size at a steady state, Carofiglio et al. propose an analytical model of the PIT dynamics [11]. The authors employ a fluid model to characterize the variations in the queue and PIT from which the average and maximum PIT size values are derived. Experimental results show that the proposed model is accurate. However, the authors do not consider heterogeneous flows using different transport mechanisms like a typical inter-networking scenario such as the Internet. This will further help in strengthening the authors' claim of the size of the PIT is small under typical network settings. In addition, the authors consider the worst-case scenario (no caching, no aggregation, no timeout, etc.). Performance analysis of the PIT under other scenarios will provide some useful insights.

To the best of our knowledge, no work has analytically modelled the PIT to estimate the interest packet blocking probability in the absence of a denial-of-service attack. To this end, we present in this paper an



(a) B2 at 5 hops away

(b) G3 at 5 hops away

(c) G6 at 2 hops away

(d) G5 at 1 hop away

Fig. 15. Interest blocking probability with 5% cache size of the universe content at routers: B2, G3, G6, G5, in Fig. 3.

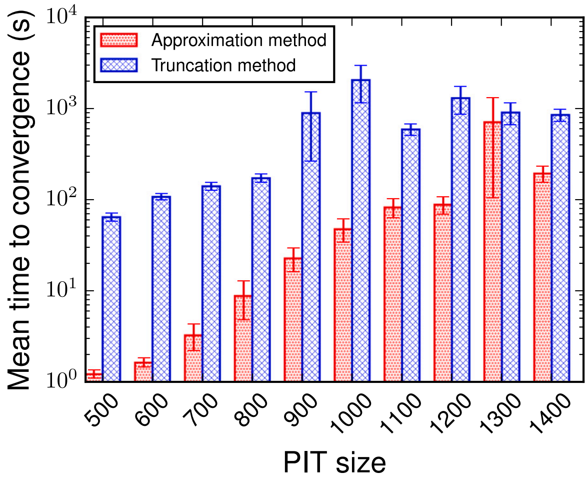


Fig. 16. Mean time taken to reach convergence for different sizes of the PIT including the standard error of the mean.

analytical model of the PIT for estimating the interest packet blocking probability.

7. Discussion and conclusion

This paper presented two analytical models of the Pending Interest Table to estimate the interest-blocking probability at a given node in a content-centric network using a 2-dimensional continuous-time Markov chain. We employ two approaches to obtain the model solutions: truncation and approximation methods. We use results from ns-3 simulation (ndnSIM module) to validate the models. In addition, we have shown that the PIT can represent a bottleneck in the presence of a high traffic load in a content-centric network even in the absence of bottlenecked links. As a result, the interest packet blocking probability estimated by the model can be used in managing the PIT occupancy, which can be used in turn to design effective and efficient congestion control mechanisms for CCN and other PIT-based ICN proposals.

In addition, we presented a comparative performance analysis of the two approaches for obtaining the model solutions. In most of the results, the approximate model achieves a better convergence time. It is also more accurate than the truncation model approach. We also showed the impact of interest retransmission ceases to be significant at a relatively large size of the PIT. Our models are still valid for deterministic, uniform and exponential per-user interest inter-arrival times, especially for nodes that are several hops away from the access routers. Thanks to the in-network content filtering effect caused by caching and interest aggregation. Results from our proposed analytical models and simulations show that PIT congestion can be alleviated by increasing the size of the cache, especially at the access routers.

Given the values of the following parameters: traffic load, PIT hit rate, cache hit rate and PIT size, we can estimate the probability that an arriving interest finds the PIT saturated and is eventually dropped. A CCN network designer can use our models to dimension the PIT for given loss rates and traffic load. In addition, as argued earlier, since there is a direct relationship between the number of interests forwarded upstream and the number of packets received by the CCN router, controlling the interest rate would have a direct impact on the data rate, therefore our model could be invoked to design an effective and efficient traffic control mechanism for CCN. We shall explore this and other approaches to using our model in the real network in the future.

Additional work is needed in modelling the content popularity and in addressing the interaction between the cache dynamics and the PIT occupancy as the content store and PIT have a joint impact on the performance of content-centric networking. We also leave these for future work.

CRediT authorship contribution statement

Amuda James Abu: Conceptualization, Methodology, Writing – original draft, Experiments, Visualization, Validation. **Brahim Bensaou:** Supervision, Conceptualization, Investigation, Writing – review & editing, Validation. **Ahmed M. Abdelmoniem:** Conceptualization, Writing – review & editing, Validation, Correspondence.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, B. Zhang, Named Data Networking, Tech. Rep. NDN-0019, 2014, Revision 1, NDN, http://named-data.net/wp-content/uploads/2014/04/tr-ndn-0019-ndn.pdf.

[2] A. Detti, N. Blefari Melazzi, S. Salsano, M. Pomposini, Conet: A content centric inter-networking architecture, in: Proceedings of the ACM SIGCOMM Workshop on Information-centric Networking, 2011, pp. 50–55.

[3] T. Sasu, A. Mark, V. kari, Towards the Future Internet: A European Research Perspective, IOS Press, Amsterdam, 2009, Ch. The Publish/Subscribe Internet Routing Paradigm (PSIRP):Designing the Future Internet Architecture.

[4] C. Dannewitz, D. Kutscher, B. Ohlman, S. Farrell, B. Ahlgren, H. Karl, Network of information (netinf) – An information-centric networking architecture, Comput. Commun. 36 (7) (2013) 721–735.

[5] T. Koponen, M. Chawla, B.-G. Chun, A. Ermolinskiy, K.H. Kim, S. Shenker, I. Stoica, A data-oriented (and beyond) network architecture, SIGCOMM Comput. Commun. Rev. 37 (4) (2007) 181–192.

[6] V. Jacobson, D.K. Smetters, J.D. Thornton, M. Plass, N. Briggs, R. Braynard, Networking named content, Commun. ACM 55 (1) (2012) 117–124.

[7] PARC, The ccnx software source v1.0, 2017, http://http://blogs.parc.com/ccnx/ccnx-downloads//.

[8] A.J. Abu, B. Bensaou, Modelling the pending interest table occupancy in ccn with interest timeout and retransmission, in: Proceedings of the 40th IEEE International Conference on Local Computer Networks, 2015, pp. 246–254.

[9] A.J. Abu, B. Bensaou, A.M. Abdelmoniem, A markov model of ccn pending interest table occupancy with interest timeout and retries, in: Proceedings of the IEEE International Conference on Communications, 2016, pp. 1–6.

[10] G. Carofiglio, M. Gallo, L. Muscariello, M. Papalini, S. Wang, Optimal multipath congestion control and request forwarding in information-centric networks, in: Proceedings of the 21st IEEE International Conference on Network Protocols, Gottingen, Germany, 2013.

[11] G. Carofiglio, M. Gallo, L. Muscariello, D. Perino, Pending interest table sizing in named data networking, in: Proceedings of the 2nd International Conference on Information-Centric Networking, 2015, pp. 49–58.

[12] L. Saino, C. Cocora, G. Pavlou, Cctcp: A scalable receiver-driven congestion control protocol for content centric networking, in: Proceedings of the IEEE ICC, Budapest, Hungary, 2013.

[13] X. Qiao, H. Wang, P. Ren, Y. Tu, G. Nan, J. Chen, M.B. Blake, Interest packets scheduling and size-based flow control mechanism for content-centric networking web servers, Future Gener. Comput. Syst. 107 (2020) 564–577.

[14] Ahmed M. Abdelmoniem, Brahim Bensaou, Enforcing transport-agnostic congestion control in sdn-based data centers, in: 2017 IEEE 42nd Conference on Local Computer Networks (LCN), 2017.

[15] D.R. Cox, P.A.W. Lewis, The Statistical Analysis of Series of Events, Methuen, London, 1966.

[16] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, Web caching and zipf-like distributions: Evidence and implications, in: IEEE INFOCOM, 1999, pp. 126–134.

[17] N. Spring, R. Mahajan, D. Wetherall, T. Anderson, Measuring isp topologies with rocketfuel, IEEE/ACM Trans. Netw. 12 (1) (2004) 2–16.

[18] M. Busari, C. Williamson, Prowgen: A synthetic workload generation tool for simulation evaluation of web proxy caches, Comput. Netw. 38 (6) (2002) 779–794.

[19] A. Afanasyev, I. Moiseenko, L. Zhang, ndnSIM: NDN simulator for NS-3, Technical Report NDN-0005, NDN, 2012, URL http://named-data.net/techreports.html.

[20] K. Katsaros, G. Xylomenos, G.C. Polyzos, Multicache: An overlay architecture for information-centric networking, Comput. Netw. 55 (4) (2011) 936–947.

[21] I. Psaras, R.G. Clegg, R. Landa, W.K. Chai, G. Pavlou, Modelling and evaluation of ccn-caching trees, in: NETWORKING 2011: 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011, Proceedings, Part I 10, Springer, 2011, pp. 78–91.

[22] C. Giovanna, G. Massimo, M. Luca, D. Perino, Modeling data transfer in content-centric networking, in: Proceedings of the 23rd International Teletraffic Congress, 2011, pp. 111–118.

[23] M. Marsan, G. de Carolis, E. Leonardi, R. Lo Cigno, M. Meo, Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials, IEEE J. Sel. Areas Commun. 19 (2) (2001) 332–346.

[24] W.J. Stewart, Probability, Markov Chains, Queues, and Simulation, Princeton University Press, New Jersey, USA, 2009.

[25] A.J. Abu, B. Bensaou, J.M. Wang, Interest packets retransmission in lossy ccn networks and its impact on network performance, in: Proceedings of the 1st International Conference on Information-centric Networking, 2014, pp. 167–176.

[26] M. Virgilio, G. Marchetto, R. Sisto, Pit overload analysis in content centric networks, in: Proceedings of the 3rd ACM SIGCOMM Workshop on Information-centric Networking, ICN '13, New York, NY, USA, 2013, pp. 67–72, http://dx.doi.org/10.1145/2491224.2491225.

[27] G. Carofiglio, M. Gallo, L. Muscariello, Icp: Design and evaluation of an interest control protocol for content-centric networking, in: Proceedings of IEEE INFOCOM Workshop on Emerging Design Choices in Name-Oriented Networking, Orlando, FL, 2012, pp. 304–309.

[28] A. Kazi, H. Badr, Some observations on the performance of ccn-flooding, in: Computing, Networking and Communications (ICNC), 2014 International Conference on, 2014, pp. 334–340.

[29] S.H. Bouk, S.H. Ahmed, M.A. Yaqub, D. Kim, M. Gerla, Dpel: Dynamic pit entry lifetime in vehicular named data networks, IEEE Commun. Lett. 20 (2) (2016) 336–339.

[30] A. J. Abu, B. Bensaou, A. M. Abdelmoniem, Inferring and Controlling Congestion in CCN via the Pending Interest Table Occupancy, in: Proceedings of IEEE Local Computer Networks, LCN, 2016, pp. 433–441.

[31] Y. Tan, Q. Li, Y. Jiang, S. Xia, Rapit: Rtt-aware pending interest table for content centric networking, in: 2015 IEEE 34th International Performance Computing and Communications Conference, IPCCC, 2015, pp. 1–8.

[32] W.U.I. Zafar, M.A.U. Rehman, F. Jabeen, S. Ghouzali, Z. Rehman, W. Abdul, Context-aware pending interest table management scheme for ndn-based vanets, Sensors 22 (11) (2022).

[33] G. Manap, S. Bilgili, A.K. Demir, On the forwarding information base sizing in named data networking of things, in: 2022 2nd International Conference on Computing and Machine Intelligence, ICMI, 2022.

[34] R. Ravindran, G. Wang, X. Zhang, A. Chakraborti, Supporting dual-mode forwarding in content-centric network, in: Proceedings of the 2012 IEEE International Conference onAdvanced Networks and Telecommuncations Systems, ANTS, 2012, pp. 55–60.

[35] S. Jang, H. Byun, H. Lim, Dynamically allocated bloom filter-based pit architectures, IEEE Access 10 (2022) 28165–28179.

[36] N. Dutta, An approach for fib construction and interest packet forwarding in information centric network, Future Gener. Comput. Syst. 130 (2022) 269–278.

[37] Y. Mordjana, B. Djamaa, M.R. Senouci, A q-learning based forwarding strategy for named data networking, in: 2021 International Conference on Networking and Advanced Systems, ICNAS, 2021.

[38] A. Siddiqua, M.A. Shah, H.A. Khattak, I. Ud Din, M. Guizani, Icafe: Intelligent congestion avoidance and fast emergency services, Future Gener. Comput. Syst. 99 (2019) 365–375.

[39] K. Wang, J. Chen, H. Zhou, Y. Qin, H. Zhang, Modeling denial-of-service against pending interest table in named data networking, Int. J. Commun. Syst. 27 (12) (2014) 4355–4368.

[40] H. Xu, H. Wang, J. Hu, Z. Yu, Modeling short-form video transfer in information centric network, in: 2021 20th International Conference on Ubiquitous Computing and Communications, 2021.

Amuda James Abu: obtained his Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology in 2017. He is currently a researcher with Environment and Climate Change Canada, Canada. Previously, he worked as a researcher in the industry for several years such as the Applied Science and Technology Research Institute, (ASTRI), HK, Huawei Research, HK and Huawei Research, Canada. His research interests are internet traffic engineering, congestion control in data networks, QoS in wired and wireless networks, future internet architectures, and network resource management.

Brahim Bensaou: (Senior Member, IEEE, Member, ACM) received his Bachelor of Engineering in Computer Science from the University of Science and Technology Houari Boumediene of Algiers, Algeria in 1982, and a DEA degree from University Paris XI in Computer Science in 1988. He earned his Doctorate in Computer Science from the University Paris VI in 1993. He is a tenured faculty member at the Computer science and Engineering department of Hong Kong University of Science and Technology. Formerly, he held positions of research assistant at France Telecom Research labs, Research Associate at HKUST, and Senior Member of Technical Staff at the National R&D Centre for Wireless Communications in Singapore where he was the leader of the strategic research group on wireless networking. His research is in general centred around computer networking, the Internet and wireless networks, in particular, Congestion Control and resource allocation, Information-centric Networking, Energy efficiency, and Performance evaluation. He published on these areas extensively in prominent conferences and journals, received numerous research grants, supervised many postgraduate research theses including both Ph.D. and Masters and holds 3 granted US patents, of which one is licensed and used in a standard protocol.

Ahmed M. Abdelmoniem: (Member ACM, IEEE, USENIX) received his Ph.D. in Computer Science and Engineering from Hong Kong University of Science and Technology, Hong Kong in 2017 and his B.Sc. and M.Sc. degree in Computer Science from Assiut University, Assiut, Egypt, in 2007 and 2012 respectively. He is an Assistant Professor at the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK and the Faculty of Computers and Information, Assuit University, Egypt. Formally, he held the position of a Research Scientist at KAUST, Saudi Arabia and a Senior Researcher with Huawei's Future Networks Lab, Hong Kong. He is an investigator on projects totalling USD 1.5Mil in funding including being the PI of an UKRI-EPSRC New Investigator Award. was awarded the prestigious Hong Kong Ph.D. Fellowship from the Research Grant Council (RGC) of Hong Kong in 2013 to pursue his PhD. He has published numerous papers in top venues and journals in the areas of distributed machine learning, computer and wireless networks, traffic engineering and congestion control. His current research interests are in the areas of optimizing systems supporting distributed machine learning and cloud/data-centre networking with an emphasis on performance, practicality, and scalability.