Computer Networks 66 (2014) 112–120

Contents lists available at ScienceDirect

Computer Networks

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

Information-centric networking: The evolution from circuits to packets to content

Jim Kurose

School of Computer Science, University of Massachusetts, Amherst, MA 01003, United States

ARTICLE INFO

Keywords:
Computer networks
Network architecture
Network protocols
Cache networks
Information-centric networks
Performance analysis

ABSTRACT

Today's information-centric networks (ICNs) represent a 100-year evolution of communication networks from circuit-switched networks to packet-switched networks to ICNs, sharing common features with both of these earlier network architectures, but having many unique characteristics of its own. We describe and survey ongoing research and identify challenges in the modeling, design and analysis of information-centric networks and protocols. We discuss performance modeling frameworks and challenges for ICNs, with a particular focus on content flowing through a network of caches, drawing analogies and distinctions from past research in both circuit-switched and packet-switched networks. We also survey the challenges and recent research results associated with finding content in a network of caches and managing the content in those caches. The challenges posed by mobility (of both the end users accessing content as well as content itself) are also discussed.

1. Introduction

Roughly one hundred years ago, Erlang created the mathematical foundations for analyzing the new telecommunication network of his day – the circuit-switched telephone network, providing the Poisson model for telephone call arrivals [22] and the celebrated Erlang formula for estimating the call blocking probability at a telephone network switch [21]. Fifty years later, Len Kleinrock's seminal Ph.D. thesis [48] used queueing theory to analyze and help make the case for a radically new type of telecommunication network – the packet-switched network, introducing the notions of demand-access resource sharing, statistical multiplexing gains, and underlying modeling assumptions (e.g., the Independence Assumption) [49,50] that would provide the technical foundation for performance models of these networks and their protocols for decades to come.

Fifty years later again, the networking field again finds itself at the doorstep of another potentially profound change – the rise of *content-centric or information-centric networks (ICNs)*. While packet-switching network architectures have focused on host-to-host communication – "delivering data between computers or between computers and terminals" [9] – today's Internet is arguably more concerned with connecting people with content and information. Such content includes stored video and audio, web content, software, and more. For example, Cisco's Visual Networking Index noted that Internet video traffic was 64% of all global consumer Internet traffic in 2012 and predicts that by 2017, 51% of all Internet traffic will cross content delivery networks [13]. In such a content-centric worldview, *what* a person wants, rather than *where* it is located, is what matters most; content, rather than the server on which content resides, becomes the starting point.

With an increased focus on content and with nearly 50 years of networking research to draw on, engineers and researchers have begun to design, and in some cases build and deploy, a wide range of content-oriented

E-mail address: kurose@cs.umass.edu

http://dx.doi.org/10.1016/j.comnet.2014.04.002
1389-1286/

networks. At one end of this range are commercially-deployed server-based content distribution networks (CDNs) such as Akamai [63] and proposed federations of CDNs [4] that re-direct a client to a "nearby" copy of content; once the address of a nearby host containing a copy has been determined, however, that content is then retrieved by the client in a traditional host–host manner. At the other end of this range are more radical designs in which network elements route (and typically cache) content objects amongst themselves, as content is forwarded from content publishers to content subscribers [7,8,18,19,25,28,33,38]. Here, content is requested and forwarded on the basis of a name or other content attribute, rather than on the basis of an address, as in traditional host-to-host communication. In these latter approaches, which we will refer to as *information-centric networks (ICNs),* content objects are the units of information processed, forwarded and stored among network elements. Just as calls were central to circuit-switched networks and packets were central to packet-switched networks, content objects are the central units of information in information-centric networks.

In this paper, we describe ongoing research and identify research challenges in the design and analysis of information-centric networks. In contrast to several recent excellent surveys of ICN architecture and protocols [1,78], our focus here will be on challenges related to the operation and performance of ICNs, and the inter-connected network of content caches within an ICN. Focussing on the modeling and performance analysis is perhaps particularly appropriate, given Len Kleinrock's foundational contributions to our understanding of the modeling and performance analysis of packet-switched networks and their protocols. But a focus on ICN cache networks and their protocols and operation is also appropriate. Like Erlang before him (who was an applied mathematician as well as an engineer, known to crawl into manholes in the streets of Copenhagen to make measurements in the local telephone network), Kleinrock is also known as a modeler, protocol designer (of routing, flow and congestion control, packet voice, and numerous wireless and mobile network protocols) and an experimenter (see, e.g., his account of the first ARPAnet remote login from UCLA to SRI [50]). In taking the long view of the history of circuit-switched, packet-switched, and now information-centric networks, it is fitting to acknowledge an individual whose contributions to packet-switched networks have been foundational, but who has also made important contributions to circuit-switched networks (e.g., research on optically switched LANs [51,58]) as well as information-centric networks (e.g., research on caching and prefetching [41,42]). Kleinrock's pioneering work on packet-switched networks bridges 100 years of networking research – a radical transformation of the circuit-switched networks before, and setting the stage for the information-centric networks of tomorrow.

The remainder of this paper is structured as follows. In Section 2, we discuss the performance modeling framework and challenges in ICNs, with a particular focus on content flowing through a network of caches, drawing analogies and distinctions from past research in both circuit-switched and packet switched networks. In Section 3 we discuss challenges in locating content in a network of caches, and managing the content in those caches. Section 4 concludes this paper.

2. Modeling information-centric networks

2.1. Perspective: circuits, packets, and content

Before delving into the challenges of modeling information-centric networks, it will be instructive to retrospectively consider ICN in the context of earlier circuit- and packet-switched networks. Fig. 1 shows a simple 7-node network with a similar topology in a circuit, packet and ICN setting.

2.1.1. Circuit switching

In the circuit-switched network shown in Fig. 1(a), calls are the basic unit of work. In the simplest scenario, an arriving call must be allocated a free circuit on each link from source to destination in order for the call to be connected. In the example, three calls are connected (two from A to G, and one from B to F), with each call holding one circuit on each link along the path from source to destination. A call that is unable to receive the full set of source-to-destination circuits is *blocked* and receives no resources. In Fig. 1(a), since all three circuits are occupied on the DE link, a new incoming call routed over the DE link (say from A to F) would be blocked. In this case, dynamic alternate routing [29,31] might be used to route this otherwise-blocked call along the path ADCEF. A successfully-connected call holds its circuits for the call's duration and then simultaneously releases these resources when the call terminates.

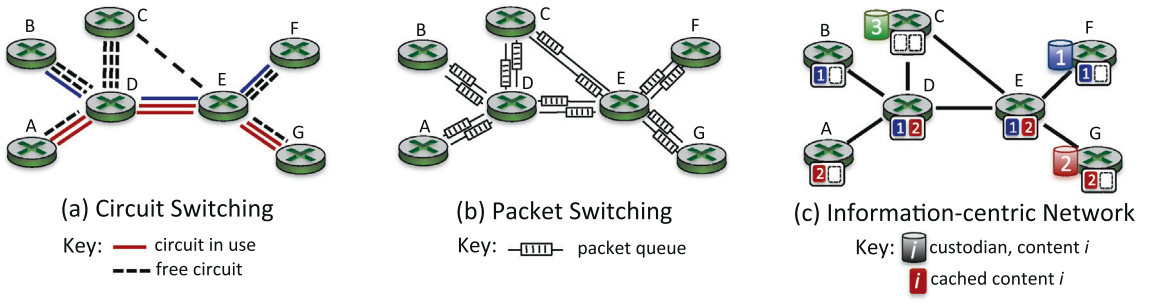

Fig. 1. Circuits, packets and content.

From a modeling standpoint, calls are allocated a *set* of resources (circuits along an end-end-path), which are then freed when the call terminates. The key performance metric is the call blocking rate, with network congestion being manifested by an increased call blocking rate. An exact analysis of blocking in a circuit-switched network is made difficult by the fact that the call arrival process at internal network switches (i.e., for calls that have successfully reserved upstream circuits) is extremely complex, even if the exogenous call arrival process is a simple (e.g., Poisson) process. In particular, a link's call blocking probability (i.e., the probability that a needed circuit is not available to an incoming call) is not independent of the blocking probability at other (e.g., nearby) links. Given these complex couplings among network elements, the call blocking probability along an end-to-end path cannot be obtained by simply considering call blocking at individual links in isolation, and then composing these link-level results. The analysis of blocking networks has a long history [46], including fixed point approximate methods [45,72] (which we will consider again shortly in the ICN context) and asymptotic analyses, with the seminal work of Kelly demonstrating that in a large-scale regime, loss probabilities can be computed *as if* links block independently [44].

2.1.2. Packet switching

In the packet-switched network shown in Fig. 1(b), packets are the basic unit of work. An arriving packet is stored in a router's buffer until it is transmitted over an outgoing link; a packet arriving at a full buffer is typically dropped.

From a modeling standpoint, a packet is allocated one resource (a buffer) while it queues for access to another resource (the link), and only holds these resources until it is forwarded to the next router. Two key performance metrics are packet delay and the throughput of packets across a link, with network congestion being manifested in increased delay and an increased number of dropped packets due to buffer overflow. As noted earlier, Kleinrock's Ph.D. thesis pioneered the use of queueing theory to model the flow of packets among routers. As in the case of circuit-switched networks, the packet departure process from a router is complex, even if the exogenous packet arrival process is simple. For this and other reasons, an exact queueing network analysis is intractable without simplifying assumptions, such as Kleinrock's independence assumption.

In cases where a single congested link dominates performance, one can focus on modeling just that single bottleneck queue in isolation and still obtain valuable performance insights (e.g., [37]). But in cases where a network-level analysis is needed, approximate (e.g., [30]) or bounding [15,16,40,54] performance models are needed.

A more recent trend in packet-switched network analysis is to raise the level of abstraction from the individual packet moving discretely from one queue to another to a *flow* of packets, with a flow being modeled as if it were a fluid flowing from source to destination. This approximation generally becomes increasingly accurate as the ratio between the link transmission rate and packet size increases; since today's common Ethernet transmission rates of 10 Gbps are six orders of magnitude faster than those of the earliest packet-switched networks and since packet sizes have remained constant, the fluid approximation has seen increasing use. A fluid approach also becomes increasingly valuable in simulation as well, as the number of packets-per second that must be simulated in a packet-level simulation (and hence the simulation run time) increases with increasing transmission rates [26]. In fluid models, source-to-destination packet flows, rather than the individual packets within a flow, become the basic unit of work. Fluid models have been used to successfully model the performance of individual packet switching elements in isolation (e.g., [20]), as well as the performance of specific packet-based network mechanisms such as AQM (e.g., [56,61]).

2.1.3. Information-centric networks

Fig. 1(c) illustrates the information-centric network setting. Here, requests for content issued by end users are the basic unit of work. In a simple ICN scenario, each piece of content has a name and a custodian, a known location where that content is permanently stored and from which that piece of content can be requested. End users generate content requests, which are forwarded among content routers towards content custodians. Each content router has a co-located *cache* that stores pieces of content passing through that router on their way to content requestor(s). Before a content router forwards a request for named content towards the content's custodian, it first checks whether it has the requested named content in its cache. If so, that router itself satisfies the request, sending the content to the requestor. Such caches are commonly referred to as Transparent En-Route Caches (TERC) [43] – "transparent" in that neither the requestor nor the custodian are aware of the cache, and "en-route" since cached content is accessed via a request being forwarded on the path to the custodian server.

As content is transmitted along a path from content sender to content requestor, it is cached at content routers along the return path. For example, the scenario shown in Fig. 1(c) could have resulted from node A requesting content item 2 from custodian node G (causing content item 2 to be cached at nodes G, E, D, and A) and node B requesting content item 1 from custodian node F (causing content item 1 to be cached at nodes F, E, D, and B). If a request for content item 1 were to be made at node A, that request would be satisfied by node D, and content item 1 would be added to the cache at node A.

From a modeling standpoint, an ICN can be modeled as a *network of caches,* with requests being forwarded by a content router towards a custodian only when a request "misses" in that router's cache. When a content request "hits" (i.e., finds the requested content resident in the router's local cache) – either at an intermediate cache or at the custodian – that content is then cached at all routers on the return path to the requestor. Storage space in the caching content routers is thus a key ICN resource. Since requested content is cached at all routers between the requestor and the content router that satisfies the request, a single content request results in multiple resources (storage space in multiple caches) being allocated, in much the

same way that a single call in a circuit-switched network results in multiple circuits being allocated between switches. This fundamental similarity suggests that cache network analysis may share more in common with the analysis of its oldest ancestor (circuit-switched networks) than with its most recent ancestor (packet-switched networks).

Resource allocation in an ICN does, however, have one significant difference from resource allocation in a circuit-switched network. When a piece of content is to be cached at an already full cache, a previously-cached piece of content must first be evicted. For example, if Node G were to request content item 3 from custodian node C in Fig. 1(c), then when content item 3 is returned to node G via node E, node E (with a cache that can hold two cached items, and is full) must decide which of its two currently cached content items (1,2) to evict. Fig. 2 shows the case that content router E has replaced item 1 in its cache with item 3; nodes A and G have also cached item 3 but did not need to evict existing content to do so.

2.2. Modeling networks of caches

The workload model for a network of caches is substantially richer than that of either circuit-switched or packet-switched networks. This richness and complexity results principally from the fact that each piece of content has a name, i.e., a distinct identity, with content being requested by name. Thus, while buffered packets are indistinguishable from each other in a packet-switched network model (and similarly circuits in use are indistinguishable in a circuit-switched model), each piece of named content is distinct. In ICN models, each piece of content, i, is thus typically characterized by a "popularity," p_i, which we can think of informally as the probability that a randomly selected content request is requesting content item i. Accesses to web content circa 2000 were empirically found to follow a Zipf distribution, where p_i is proportional to $1/i^{\alpha}$ and i is typically less than 1 [5]. The Zipf parameter α allows for a family of distributions to be modeled, with an α value approaching zero resulting in a uniform popularity distribution, and larger values of α producing distributions with increasingly lighter tails. Perhaps as a result of similarities between web and ICN content access, ICN models have typically adopted Zipf-like popularity distributions.

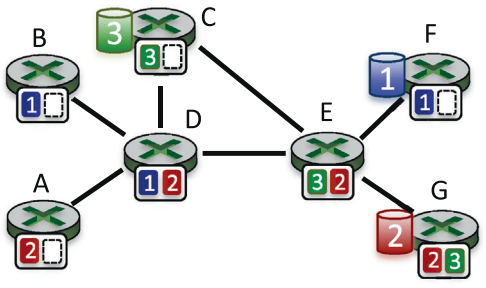

Fig. 2. The ICN scenario from Fig. 1, following the request (by node G) and delivery of content item 3 from node C.

Diving deeper into the ICN workload model, let us focus next on the stream of arriving exogenous requests for content item i, and the probability that an incoming reference at time t is for content item i. For more than 40 years [2], a common model adopted by researchers studying caches – whether memory hardware caches, disk caches, web caches or ICN caches – is the *independent reference model (IRM)*. IRM states that the probability that a request made at time t is for item i is simply p_i, independent of t, independent of the history of past requests for item i and independent of requests for other content items, j. This independence assumption is as fundamental for cache modeling as the memoryless assumption of exponential packet/circuit inter-arrival times (Erlang referred to this memoryless assumption as calls being "distributed quite accidentally during the day") are for modeling packet- and circuit-switched networks, and Kleinrock's independence assumption is for modeling packet service/link-transmission time in packet-switched networks.

A final distinguishing characteristic of cache networks is the cache replacement policy employed when a content item currently in the cache must be removed to make space for a newly arrived content item. Least recently used (LRU) cache replacement is commonly assumed in ICN models (e.g., [27,52,71]) and has been advocated for use in specific ICN architectures, e.g., [38]. An exact analysis of LRU caching, even in a standalone cache and under IRM assumptions is difficult, and so approximation techniques have been developed [11,17,27]. In more static scenarios, cache content replacement is performed on the basis of measured or anticipated content popularity over a longer time scale (e.g., on a daily basis, as is done in some video CDNs [62]), rather than on a fine time-scale, per-request basis. In this case, static caching of the most frequently requested content has been proposed [74] and shown to be optimal under mild assumptions [57]. Ref. [64] surveys web cache replacement policies.

2.3. Analyzing networks of caches: a fixed-point approximation

Armed with an understanding of content popularity, IRM, and cache replacements policies, let us next consider the modeling and analysis of *networks* of caches. As in the case of packet- and circuit-switched networks, the primary complicating factor here is that the output process from a network element (content requests that have "missed" in the cache) is complex. In particular, even if the input request process to a cache is IRM, the output process is not. This can been seen informally by noting that if an LRU cache experiences a miss for content item i and forwards a request for i at time t (and then that item is subsequently located, downloaded and cached at this cache), then the next request forwarded by the cache as the result of the very next miss is very unlikely to be for content item i, since i was just added to the cache as a result of the request miss at time t. The output miss process from a cache for a given piece of content is a component of the arrival stream at the upstream cache on the path to the content custodian for that piece of content, and thus a complex

output process at one cache results in a complex arrival processes at other caches. For example, requests for content items 1 and 2 arriving at cache E in Fig. 2 from cache D and destined to content custodians F and G will only occur following a cache miss at downstream cache D.

In order to analyze the hits rates at various network caches, the following fixed-point approximation approach similar to that used in circuit- and packet-switched networks was developed in [70]:

- Let $r_{i,v}$ be the combined incoming rate of requests for content item i at node v and let $m_{i,v}$ be the miss rate for i at node v. Then the rate of content requests at each node can be expressed as

$$r_{i,v} = \lambda_{i,v} + \sum_{v':i\in R(v';v)} m_{i,v'} \quad (1)$$

  where $\lambda_{i,v}$ is the exogenous arrival rate of requests for content item i at content router v, and v' is the set of content routers that forward a miss for i to content router v.
- Under the IRM model, the miss rate for item i at content router v can be expressed:

$$m_{i,v} = r_{i,v}(1 - q_{i,v}) \quad (2)$$

  where $q_{i,v}$ is the hit probability for item i at content router v.
- Assuming IRM arrivals at v, approximate techniques such as [17,11], developed for computing the hit probabilities for a single standalone cache, can be used to compute the hit probability for item i at content router v:

$$q_{i,v} = hit\left(\left\{\frac{r_{i,v}}{\sum_j r_{j,v}}\right\}, |v|\right) \quad (3)$$

  where $\frac{r_{i,v}}{\sum_j r_{j,v}}$ is the probability that a request arrival at v is for content item i, and $hit()$ is the approximation technique (e.g., [11,17]) used to compute the hit rate for i at v for a given cache replacement policy, a cache of size $|v|$ and the given request arrival rates.

Eqs. (1)–(3) provide a set of fixed-point equations that can be solved iteratively to compute hit and miss probabilities for all content items at all caches. It is, however, an approximate model. There are multiple possible sources of inaccuracy, including those introduced by the underlying approximate model used to calculate the hit/miss probabilities at each individual cache in isolation and the IRM assumption. Ref. [70] identifies IRM as the primary cause for inaccuracies, but that in the numerical cases studied, the relative error is less than 15%, with accuracy improving as the number of caches and connectivity among caches scales. We conjecture that in asymptotic large scale regimes, hit probabilities can be computed exactly, as if hits/misses at all caches were independent and the output processes were IRM, similar in spirit the earlier results for computing blocking probabilities in circuit-switched networks [44].

2.4. Other approaches for analyzing cache networks

While an approximate fixed-point method can be used to analyze cache networks with arbitrary topologies, a number of researchers have also specifically investigated caching in tree topologies [6,11,67], a scenario particularly appropriate in hierarchical (e.g., web) caching networks where content custodians are located at the root of a tree of caches.

Additionally, researchers have recently begun investigating the use of fluid models for ICN network analysis [6]; [14] presents techniques for fluid analyses of related peer-to-peer caching networks. Models that compute provable bounds on the performance of cache networks, using a bounding characterization for content request traffic flows can be found in [71]. This work was inspired by Cruz's (σ, ρ) calculus [15,16] for computing delay bounds in packet-switched networks.

3. Locating content and caching content

In this section, we identify additional challenges in the design and operation of cache networks, with a focus on issues affecting ICN performance. For a discussion of security and privacy challenges, see [60,73,77] and [3,53], respectively; see [1,78] for broader surveys of ICNs and their challenges.

3.1. Where to cache content?

In our discussion above and in initial ICN proposals (e.g., [39]), it was assumed that caching was universal – that all content routers performed caching and that all content was cached along download paths.

Even if all content routers can perform caching, it can be advantageous to *selectively* cache content along the download path, rather than at each and every content router. In Fig. 1(c), for example, rather than caching content item 1 at both D and E (both of which are on the paths from A and B to F, the custodian for content item 1), it might be preferable to cache content item 1 at just one of these two routers, allowing a different piece of content, that would otherwise not be cached, to be cached at the other of these two routers. Che exploited such observations in proposing selective caching in tree topologies [11], showing that cooperative caching decreased the storage space needed to achieve a given hit rate by a factor of two over uncooperative hierarchical caching in a two-level tree. Similarly, probabilistic content caching along download paths was investigated in [10,66] and shown to improve performance. Explicit and implicit coordination among caches (e.g., enroute caching of downloaded content only if that content is not stored in a nearby cache, or based on the content router's distance from the content custodian) was investigated in [12,59].

An even more fundamental question than whether to cache a particular piece of content at a particular cache is where to perform caching in the first place. Does it make sense to cache ubiquitously, or might it be possible to gain most of the advantages of pervasive caching by strategically locating caches within the network? Ref.

[10] observes that if a content router is on a large number of paths from requesting nodes to content custodians, then it is likely to receive more request traffic, and hence by allocating storage and caching at these more "central" nodes, cache hit rates can be improved over the case of uniform, heterogeneous caching. Refs. [25,68] argue for caching at the network edge, close to the end users, rather than ubiquitously at all content routers. Building on this observation, Ref. [25], presents a proof-of-concept design for an incrementally deployable application-level ICN architecture.

3.2. How to find cached content?

A complementary challenge to the question of where/when to cache content is the question of how to find content in the cache network. In the simple ICN scenarios discussed in Section 2, for example, cached content is found serendipitously while routing a content request along the shortest path from content requestor to content custodian; content cached near to, but not directly on, this shortest path would not be found. In Gnutella's unstructured peer-to-peer network [34], content is located via a flooding ring search to locate the closest peer with a copy of the requested content.

Between these alternatives of serendipitous on-shortest-path search and flooding search, many mechanisms have been proposed for more intentionally searching for cached content in the interconnected network of caches. In the Breadcrumbs approach [69], when a piece of content is removed from a content router's cache, the router still maintains a pointer to the upstream/downstream content routers from/to which that content was forwarded (and cached); note that this most recent content-forwarding path may well differ from the content-request forwarding path. An arriving content request that misses in the local cache may then be detoured to follow this path of "breadcrumbs" to other content routers that have recently stored and forwarded the requested content. Summary Cache [24] presents a web-caching technique for maintaining an efficient, compact summary of the identity of items stored in nearby caches. Using Summary Cache in an ICN, an arriving content request that misses in the local cache could then be detoured to a neighboring cache, as in Breadcrumbs. In both cases, detouring is "best effort" – a detoured request is not guaranteed to find requested content in a neighbor cache. When a content router may be aware of multiple copies of nearby cached content, requests can be forwarded to a particular copy based on numerous different criteria [23].

While many mechanisms have been proposed for a more intentional search for cached content, many fundamental challenges still remain. In most past research, the hop-count between the requestor and the content router that satisfies the request has been the primary performance metric of interest, with closer cache hits being preferred. But ultimately, the download times of the requested content – and the download times of other content requests that share bottleneck network bandwidth with that requested content – are the primary application-level performance metrics of interest. This suggests that content search should be based on the amount of available bandwidth on the shared bottleneck link on the download path between the content requestor and the content router that satisfies that request (which then determines content download time). There will also be "knock-on" effects on the download times of ongoing and future content downloads on other download paths that intersect this download path. The effect of caching on the throughput capacity of wireless networks has recently been investigated in [32], but per-request cache network search to minimize download times, remains largely unaddressed.

More fundamentally, the question of how much effort and overhead the network should put into content search – both for routing content search as well as distributing state that is used by content search – remains unanswered. Clearly, as the amount of network resources used by content download increases, the more advantageous it becomes to use additional resources to locate the "best" cached copy of the requested content.

3.3. Mobility

With the global number of active mobile broadband subscriptions now exceeding the number of wired broadband subscriptions by a factor of three [75] and reports that "by 2016, wired devices will account for only 39% of IP traffic, while WiFi and mobile devices will account for 61%" [13], support for device mobility is a critical component of any future Internet architecture [65]. Here "mobility" must be carefully, but broadly, construed – a user physically moving among access points or base stations within the same subnet retains its IP address; this user is stationary from an addressing and inter-network routing point of view. On the other hand, a physically stationary user shifting among multiple devices attached via contemporaneous connections to different networks will change access networks and the IP address to which his/her identity is associated; this physically stationary use is "mobile" from a network attachment point of view [79]. Similarly, if content changes name, that content is functionally "mobile," since a new content name will be used for name-based content routing [35].

Research challenges at the intersection between ICN and mobility are still relatively unexplored. Ref. [76] considers information dissemination in a linear V2V network using the NDN ICN architecture [38], focusing primarily on the impact of protocol timer values on performance. Proxies and/or indirection points (such as the HLR in cellular networks and home agent in Mobile IP) have been a common feature of many architectures supporting mobility, including recent proposals for NDN-like architectures. Refs. [36,55] both adopt a proxy-based approach and rely on underlying tunneling or the existence of IP addresses to deliver content. Kim [47] proposes the use of an indirection point where mobile content publishers and subscribers can register mobility-related name changes or query for new names associated with a mobile publisher. Ref. [35] presents a quantitative methodology and results comparing different location-independent architectures (including named-based ICN) using a common set of metrics that include update cost at routers, path stretch, and forwarding

table size in scenarios where either devices or content may be mobile.

4. Summary and conclusion

In this paper, we have described and surveyed ongoing research and identified challenges in the modeling, design and analysis of information-centric networks and protocols. Drawing analogies, as well as distinctions, from past research in both circuit-switched and packet-switched networks, our particular focus was networks of caches, and the content flowing through, and stored in, these caches. The additional research challenges we discussed included finding content in a network of caches, managing the content in those caches, and handling the mobility of both the end users accessing content as well as content itself. We noted that ICN design space is large, ranging from commercially-deployed content distribution networks to more radical designs in which network elements route and content objects amongst themselves.

The evolution of large scale networks (whether for communication, electrical or transportation) often occur in a style of punctuated equilibrium slow evolutionary changes between short periods of profound change. The telephone network evolved from human control to wired-logic control to stored-program control of electronic switches to today's IP-based control networks. Whether tomorrow's ICN results from a smooth evolution of today's CDNs or from more profound architectural change, content storage, replication and access will play a central in the communication networks of the future.

Acknowledgements

This material is based upon work supported by the US National Science Foundation under grants CNS-1117764 and CNS-1040735. The author, an academic half-grandson of Len Kleinrock, appreciates the many insightful and enjoyable discussions over the years about cache networks with his former Ph.D. student, Elisha Rosensweig (now an academic great grandson of Len Kleinrock).

References

[1] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, B. Ohlman, A survey of information-centric networking, Commun. Magaz., IEEE 50 (7) (2012) 26–36.
[2] A.V. Aho, P.J. Denning, J.D. Ullman, Principles of optimal page replacement, J. ACM 18 (1) (1971) 80–93.
[3] S. Arianfar, T. Koponen, B. Raghavan, S. Shenker, On preserving privacy in content-oriented networks, in: Proceedings of the ACM SIGCOMM Workshop on Information-Centric Networking, ICN'11, ACM, New York, NY, USA, 2011, pp. 19–24.
[4] A. Balachandran, V. Sekar, A. Akella, S. Seshan, Analyzing the potential benefits of cdn augmentation strategies for internet video workloads, in: Proceedings of the 2013 Conference on Internet Measurement Conference, IMC'13, ACM, New York, NY, USA, 2013, pp. 43–56.
[5] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, Web caching and zipf-like distributions: evidence and implications, in: INFOCOM'99, Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, vol. 1, March 1999, pp. 126–134.
[6] G. Carofiglio, M. Gallo, L. Muscariello, D. Perino, Modeling data transfer in content-centric networking, in: 23rd International Teletraffic Congress (ITC), 2011, September 2011, pp. 111–118.
[7] G. Carofiglio, G. Morabito, L. Muscariello, I. Solis, M. Varvello, From content delivery today to information centric networking, Comp. Netw. 57 (16) (2013) 3116–3127.
[8] A. Carzaniga, A.L. Wolf, Forwarding in a content-based network, in: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM'03, ACM, New York, NY, USA, 2003, pp. 163–174.
[9] V. Cerf, R. Kahn, A protocol for packet network intercommunication, IEEE Trans. Commun. 22 (5) (1974) 637–648.
[10] W.K. Chai, D. He, I. Psaras, G. Pavlou, Cache less for more in information-centric networks, in: Proceedings of the 11th International IFIP TC 6 Conference on Networking, IFIP'12, vol. Part I, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 27–40.
[11] H. Che, Y. Tung, Z. Wang, Hierarchical web caching systems: modeling design, and experimental results, IEEE J. Select. Areas Commun. 20 (7) (2002) 1305–1314.
[12] K. Cho, M. Lee, K. Park, T. Kwon, Y. Choi, S. Pack, Wave: popularity-based and collaborative in-network caching for content-oriented networks, in: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2012, March 2012, pp. 316–321.
[13] Cisco, Cisco Visual Networking Index, 2012 <http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights/index.html>.
[14] F. Clevenot, P. Nain, A simple fluid model for the analysis of the squirrel peer-to-peer caching system, in: Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2004, vol. 1, March 2004, p. 95.
[15] R. Cruz, A calculus for network delay. i. Network elements in isolation, IEEE Trans. Inform. Theory 37 (1) (1991) 114–131.
[16] R. Cruz, A calculus for network delay. ii. Network analysis, IEEE Trans. Inform. Theory 37 (1) (1991) 132–141.
[17] A. Dan, D. Towsley, An approximate analysis of the LRU and FIFO buffer replacement schemes, SIGMETRICS Perform. Eval. Rev. 18 (1) (1990) 143–152.
[18] 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.
[19] M. Diallo, V. Sourlas, P. Flegkas, S. Fdida, L. Tassiulas, A content-based publish/subscribe framework for large-scale content delivery, Comput. Netw. 57 (4) (2013) 924–943.
[20] A.I. Elwalid, D. Mitra, Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic, in: Proceedings of the Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies on One World Through Communications, IEEE INFOCOM '92, vol. 1, IEEE Computer Society Press, Los Alamitos, CA, USA, 1992, pp. 415–425.
[21] A. Erlang, Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, in: E. Brockmeyer, H. Halstkom, A. Jensen (Eds.), The life and works of A.K. Erlang, Transactions of the Danish Academy of Technical Sciences (ATS), No. 2 1948. Originally published in Danish in Elektrotkeknikeren, vol. 13, 1917, pp. 138–155.
[22] A. Erlang, The theory of probabilities and telephone conversations, in: E. Brockmeyer, H. Halstkom, A. Jensen (Eds.), The Life and Works of A.K. Erlang, Transactions of the Danish Academy of Technical Sciences (ATS), No. 2 1948, Originally published in Danish in Nyt Tidsskrift for Matematik B, vol. 20, 1909, pp. 131–137.
[23] S. Eum, K. Nakauchi, M. Murata, Y. Shoji, N. Nishinaga, Catt: potential based routing with content caching for icn, in: Proceedings of the Second Edition of the ICN Workshop on Information-Centric Networking, ICN '12, ACM, New York, NY, USA, 2012, pp. 49–54.
[24] L. Fan, P. Cao, J. Almeida, A.Z. Broder, Summary cache: a scalable wide-area web cache sharing protocol, IEEE/ACM Trans. Netw. 8 (3) (2000) 281–293.
[25] S.K. Fayazbakhsh, Y. Lin, A. Tootoonchian, A. Ghodsi, T. Koponen, B. Maggs, K. Ng, V. Sekar, S. Shenker, Less pain, most of the gain: Incrementally deployable icn, SIGCOMM Comput. Commun. Rev. 43 (4) (2013) 147–158.
[26] D.R. Figueiredo, B. Liu, Y. Guo, J. Kurose, D. Towsley, On the efficiency of fluid simulation of networks, Comput. Netw. 50 (12) (2006) 1974–1994.
[27] C. Fricker, P. Robert, J. Roberts, A versatile and accurate approximation for LRU cache performance, in: Proceedings of the 24th International Teletraffic Congress, ITC '12, 2012, pp. 8:1–8:8.
[28] A. Ghodsi, S. Shenker, T. Koponen, A. Singla, B. Raghavan, J. Wilcox, Information-centric networking: seeing the forest for the trees, in: Proceedings of the 10th ACM Workshop on Hot Topics in Networks, HotNets-X, ACM, New York, NY, USA, 2011, pp. 1:1–1:6.

[29] R. Gibbens, F. Kelly, P. Key, Dynamic alternative routing – modelling and behaviour, in: M. Bonatti (Ed.), Teletraffic Science, ITC12, Elsevier, Amsterdam, 1989, pp. 1019–1025.
[30] R.J. Gibbens, S.K. Sargood, C.V. Eijl, F.P. Kelly, H. Azmoodeh, R.N. Macfadyen, N.W. Macfadyen, Fixed-point models for the end-to-end performance analysis of IP networks, in: Proceedings of the 13th ITC Specialist Seminar: IP Traffic Measurement, Modeling and Management, September 2000, 2000.
[31] A. Girard, Routing and Dimensioning in Circuit-Switched Networks, 1st ed., Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.
[32] S. Gitzenis, G. Paschos, L. Tassiulas, Asymptotic laws for joint content replication and delivery in wireless networks, IEEE Trans. Inform. Theory 59 (5) (2013) 2760–2776.
[33] M. Gritter, D.R. Cheriton, An architecture for content routing support in the internet, in: Proceedings of the 3rd Conference on USENIX Symposium on Internet Technologies and Systems, USITS'01, vol. 3, USENIX Association, Berkeley, CA, USA, 2001, p. 4.
[34] L. Guo, S. Jiang, L. Xiao, X. Zhang, Exploiting content localities for efficient search in p2p systems, in: Proceedings of the 18th International Symposium on Distributed Computing, 2004, pp. 729–742.
[35] Z. Guo, A. Venkataramani, J. Kurose, S. Heimlicher, Towards a Quantitative Comparison of the Cost-Benefit Trade-Offs of Location-Independent Network Architectures. Technical report, School of Computer Science, University of Masachusetts, Amherst MA 01003, 2014.
[36] F. Hermans, E. Ngai, P. Gunningberg, Global source mobility in the content-centric networking architecture, in: Proceedings of the 1st ACM Workshop on Emerging Name-Oriented Mobile Networking Design – Architecture, Algorithms, and Applications, NoM '12, ACM, New York, NY, USA, 2012, pp. 13–18.
[37] C. Hollot, V. Misra, D. Towsley, W. Gong, Analysis and design of controllers for AQM routers supporting TCP flows, IEEE Trans. Autom. Control 47 (6) (2002) 945–959.
[38] V. Jacobson, D.K. Smetters, J.D. Thornton, M. Plass, N. Briggs, R. Braynard, Networking named content, Commun. ACM 55 (1) (2012) 117–124.
[39] V. Jacobson, D.K. Smetters, J.D. Thornton, M.F. Plass, N.H. Briggs, R.L. Braynard, Networking named content, in: Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies, CoNEXT '09, ACM, New York, NY, USA, 2009, pp. 1–12.
[40] Y. Jiang, A basic stochastic network calculus, in: Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM '06, ACM, New York, NY, USA, 2006, pp. 123–134.
[41] Z. Jiang, L. Kleinrock, An adaptive network prefetch scheme, IEEE J. Select. Areas Commun. 16 (3) (1998) 358–368.
[42] Z. Jiang, L. Kleinrock, Web prefetching in a mobile environment, Pers. Commun., IEEE 5 (5) (1998) 25–34.
[43] Y. Jin, W. Qu, K. Li, A survey of cache/proxy for transparent data replication, in: Second International Conference on Semantics, Knowledge and Grid, 2006, SKG '06, November 2006, pp. 35–35.
[44] F.P. Kelly, Blocking probabilities in large circuit-switched networks, Advan. Appl. Probab. 18 (1986) 473–505.
[45] F.P. Kelly, Fixed point models of loss networks, The ANZIAM J. 31 (1989) 204–218. 10.
[46] F.P. Kelly, Loss networks, The Ann. Appl. Probab. 1 (3) (1991) 319–378. 08.
[47] D.-h. Kim, J.-h. Kim, Y.-s. Kim, H.-s. Yoon, I. Yeom, Mobility support in content centric networks, in: Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking, ICN '12, ACM, New York, NY, USA, 2012, pp. 13–18.
[48] L. Kleinrock, Communication Nets; Stochastic Message Flow and Delay, McGraw-Hill Book Company, New York, 1964 (Out of Print.) Reprinted by Dover Publications, 1972 and in 2007. Published in Russian, 1971, Published in Japanese, 1975.
[49] L. Kleinrock, On the modeling and analysis of computer networks, Proc. IEEE 81 (8) (1993) 1179–1191.
[50] L. Kleinrock, Creating a mathematical theory of computer networks, INFORMS, Operat. Res. 50 (1) (2002) 125–131.
[51] L. Kleinrock, M. Gerla, N. Bambos, J. Cong, E. Gafni, L. Bergman, J. Bannister, S. Monacos, T. Bujewski, P.-C. Hu, B. Kannan, B. Kwan, E. Leonardi, J. Peck, P. Palnati, S. Walton, The supercomputer supernet testbed: a WDM based supercomputer interconnect, JSAC/JLT Spec. Iss. Opt. Technol. Netw. 14 (6) (1996).
[52] N. Laoutaris, H. Che, I. Stavrakakis, The LCD interconnection of LRU caches and its analysis, Perform. Eval. 63 (7) (2006) 609–634.
[53] T. Lauinger, N. Laoutaris, P. Rodriguez, T. Strufe, E. Biersack, E. Kirda, Privacy risks in named data networking: What is the cost of performance?, SIGCOMM Comput Commun. Rev. 42 (5) (2012) 54–57.
[54] J.-Y. Le Boudec, P. Thiran, Network Calculus: A Theory of Deterministic Queuing Systems for the Internet, Springer-Verlag, Berlin, Heidelberg, 2001.
[55] J. Lee, D. Kim, M. Wuk Jang, B.-J. Lee, Proxy-based mobility management scheme in mobile content centric networking (ccn) environments, in: IEEE International Conference on Consumer Electronics (ICCE), 2011, January 2011, pp. 595–596.
[56] Y. Liu, F.L. Presti, V. Misra, D.F. Towsley, Y. Gu, Scalable fluid models and simulations for large-scale ip networks, ACM Trans. Model. Comput. Simul. 14 (3) (2004) 305–324.
[57] Z. Liu, P. Nain, N. Niclausse, D. Towsley, Static caching of web servers, in: Multimedia Computing and Networking 1998, San Jose, California, January 1998.
[58] J.C. Lu, L. Kleinrock, Performance analysis of single-hop wavelength division multiple access networks, J. High-Speed Netw. 1 (1) (1992) 61–77.
[59] Z. Ming, M. Xu, D. Wang, Age-based cooperative caching in information-centric networks, in: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2012, March 2012, pp. 268–273.
[60] S. Misra, R. Tourani, N.E. Majd, Secure content delivery in information-centric networks: design, implementation, and analyses, in: Proceedings of the 3rd ACM SIGCOMM Workshop on Information-Centric Networking, ICN '13, ACM, New York, NY, USA, 2013, pp. 73–78.
[61] V. Misra, W.-B. Gong, D. Towsley, Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to red, SIGCOMM Comput. Commun. Rev. 30 (4) (2000) 151–160.
[62] Netflix, Openconnect Deployment Guide, 2012 <https://netflix.hs.llnwd.net/e1/us/layout/signup/deviceinfo/OpenConnectDeploymentGuide-v2.4a.pdf>.
[63] E. Nygren, R.K. Sitaraman, J. Sun, The akamai network: a platform for high-performance internet applications, SIGOPS Oper. Syst. Rev. 44 (3) (2010) 2–19.
[64] S. Podlipnig, L. Böszörmenyi, A survey of web cache replacement strategies, ACM Comput. Surv. 35 (4) (2003) 374–398.
[65] M.F. project, Homepage <http://mobilityfirst.winlab.rutgers.edu/>.
[66] I. Psaras, W.K. Chai, G. Pavlou, Probabilistic in-network caching for information-centric networks, in: Proceedings of the Second Edition of the ICN Workshop on Information-Centric Networking, ICN'12, ACM, New York, NY, USA, 2012, pp. 55–60.
[67] I. Psaras, R.G. Clegg, R. Landa, W.K. Chai, G. Pavlou, Modelling and evaluation of ccn-caching trees, in: Proceedings of the 10th International IFIP TC 6 Conference on Networking, NETWORKING'11, vol. Part I, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 78–91.
[68] J. Roberts, On the performance of caching in information-centric networks, in: 17th International GI/ITG Conference on Measurement, Modelling and Evaluation of Computing Systems, March 2104.
[69] E. Rosensweig, J. Kurose, Breadcrumbs: efficient, best-effort content location in cache networks, in: INFOCOM 2009, IEEE, April 2009, pp. 2631–2635.
[70] E. Rosensweig, J. Kurose, D. Towsley, Approximate models for general cache networks, in: INFOCOM, 2010 Proceedings IEEE, March 2010, pp. 1–9.
[71] E.J. Rosensweig, J. Kurose, A network calculus for cache networks, in: INFOCOM, IEEE, 2013, pp. 85–89.
[72] K.W. Ross, Multiservice Loss Models for Broadband Telecommunication Networks, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1995.
[73] D. Smetters, V. Jacobson, Securing Network Content. Technical report, Palo Alto Research Center. 3333 Coyote Hill Road. Palo Alto, CA 94304, 2009.
[74] I. Tatarinov, A. Rousskov, V. Soloviev, Static caching in web servers, in: Proceedings of the Sixth International Conference on Computer Communications and Networks, 1997, September 1997, pp. 410–417.
[75] I.T. Union. ICT Statistics Home Page, 2013 <http://www.itu.int/en/ITU-D/Statistics/Pages/stat>.
[76] L. Wang, A. Afanasyev, R. Kuntz, R. Vuyyuru, R. Wakikawa, L. Zhang, Rapid traffic information dissemination using named data, in: Proceedings of the 1st ACM Workshop on Emerging Name-Oriented Mobile Networking Design – Architecture, Algorithms, and Applications, NoM '12, ACM, New York, NY, USA, 2012, pp. 7–12.

[77] M. Xie, I. Widjaja, H. Wang, Enhancing cache robustness for content-centric networking, in: INFOCOM, 2012 Proceedings IEEE, March 2012, pp. 2426–2434.
[78] G. Xylomenos, C. Ververidis, V. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. Katsaros, G. Polyzos, A survey of information-centric networking research, Commun. Surv. Tut., IEEE, PP (99) (2013) 1–26.
[79] S. Yang, J. Kurose, A. Venkataramani, S. Heimlicher, User Transitioning Among networks – A Measurement and Modeling Study. Technical report, School of Computer Science, University of Masachusetts, Amherst MA 01003, 2014.

Jim Kurose received a B.A. degree in physics from Wesleyan University and a Ph.D. degree in computer science from Columbia University. He is currently Distinguished University Professor in the School of Computer Science at the University of Massachusetts Amherst. Professor Kurose has been a Visiting Scientist at IBM Research, INRIA, Institut EURECOM , U. Paris, the Laboratory for Information, Network and Communication Sciences, and Technicolor Research. His research interests include network protocols and architecture, network measurement, sensor networks, multimedia communication, and modeling and performance evaluation. Dr. Kurose was the founding Editor-in-Chief of the *IEEE/ACM Transactions on Networking*, and has been Technical Program Co-Chair for *IEEE Infocom, ACM SIGCOMM, ACM SIGMETRICS, ACM Internet Measurement Conference, and ACM e-Energy*. He has won several conference best paper awards and received the IEEE Infocom Achievement Award and the ACM Sigcomm Test of Time Award. He has also received a number of awards for his educational activities, including the IEEE Taylor Booth Education Medal. He is a Fellow of the IEEE and the ACM. With Keith Ross, he is the co-author of the textbook, *Computer Networking, a top down approach (6th edition), published by Pearson.*