[DMSt06] *Vasilios Darlagiannis, Andreas Mauthe, Ralf Steinmetz;* **Sampling Cluster Endurance for Peer-to-Peer based Content Distribution Networks;** Thirteenth Annual Multimedia Computing and Networking (MMCN'06), Januar 2006, S. 60710N-1 - 60710N-14.

Sampling Cluster Endurance for Peer-to-Peer based Content Distribution Networks

Vasilios Darlagiannis[a] and Andreas Mauthe[b] and Ralf Steinmetz[a]

[a]Technische Universität Darmstadt, Multimedia Communications (KOM), Merckstr. 25, D-64283 Darmstadt, Germany;
[b]Lancaster University, Computing Department, Lancaster, LA1 4YR, UK

ABSTRACT

Several types of Content Distribution Networks are being deployed over the Internet today, based on different architectures to meet their requirements (e.g., scalability, efficiency and resiliency). Peer-to-Peer (P2P) based Content Distribution Networks are promising approaches that have several advantages. Structured P2P networks, for instance, take a proactive approach and provide efficient routing mechanisms. Nevertheless, their maintenance can increase considerably in highly dynamic P2P environments. In order to address this issue, a two-tier architecture that combines a structured overlay network with a clustering mechanism is suggested in a hybrid scheme.

In this paper, we examine several sampling algorithms utilized in the aforementioned hybrid network that collect local information in order to apply a selective join procedure. The algorithms are based mostly on random walks inside the overlay network. The aim of the selective join procedure is to provide a well balanced and stable overlay infrastructure that can easily overcome the unreliable behavior of the autonomous peers that constitute the network. The sampling algorithms are evaluated using simulation experiments where several properties related to the graph structure are revealed.

Keywords: Sampling mechanisms, Overlay networks, Peer-to-Peer systems, Stability

1. INTRODUCTION

Content Distribution Networks (CDN) are used to deliver content to potentially very large user **populations.**[1] Within the Internet, different types of CDN are being envisaged ranging from simple Web based applications to sophisticated multimedia entertainment systems, including interactive systems such as multiplayer games and virtual environments. Whereas first generation CDN have mostly focused on Web content, the current, second generation systems also deal with Video-on-Demand (VoD) and audio and video streaming. A number of standardization attempts are being made to address different aspects of content delivery. MPEG-21[2] for example is concerned with the entire workflow of digital content creation, delivery and trading. It covers all interaction with multimedia content and provides a framework for all content related tasks (of which delivery is one). The IETF also addressed content distribution within the CDI (Content Distribution Interworking) working group.[3] Its focus was on large scale content distribution in a Web context including provider interaction.

A special, very successful type of CDN are Peer-to-Peer (P2P) file sharing systems. In 2001 for instance, Napster was the fastest growing application in the Internet's history.[4] Since Napster, a number of unstructured P2P systems have been developed such as Gnutella,[5] eDonkey,[6] but also structured approaches such as Chord,[7] CAN,[8] etc. What these systems share is the idea to have independent, collaborating nodes that organize and share information in a peer-to-peer fashion. Ideally there is no central instance that polices or governs the interaction between the peers as it is the case in client-server interaction. The P2P paradigm basically states that P2P systems are self-organizing systems consisting of equal, autonomous entities where the interaction is governed by rules.

Further author information: (Send correspondence to Vasilios Darlagiannis)
Vasilios Darlagiannis: E-mail: bdarla@KOM.tu-darmstadt.de, Telephone: +49-6151-166155

This paradigm can be applied to a multitude of structures and systems. Apart from file sharing, P2P mechanisms are also proposed for media (mainly video) streaming.[9–12] Here, the focus is on the streaming of media from multiple senders to one receiver exploiting certain media properties (e.g., layered video coding). P2P structures are also being used for the transmission of media to multiple receivers, as in the case of application level multicast. A single tree approach is for instance taken in PeerCast[13] and SpreadIT.[14] In order to achieve better load balancing and improve resilience to node failures, multiple multicast trees are employed in the case of P2PCast[15] and SplitStream.[16] Other questions that are being addressed in the context of P2P based content distribution networks is the replication of files on a large set of peers. This has been labeled Quality of Availability (QoA) and defines a metric for the availability of certain content items within the system.[17] BitTorrent[18] is one of the most popular systems using replication on a wider scale. Though, it uses a central instance (i.e., a web-service to redirect the client to the tracker) the exchange and organization of content is essentially P2P. FastReplica[19] is another replication system for large scale replication that uses a central instance (comparable to the control of surrogate servers in the case of CDI).

The problem most P2P based CDN encounter is the multitude of requirements placed on them. Hence, a number of solutions are very restricted in their approach concentrating on a (sub-)set of the critical requirements. However, this only provides a solution for specific cases and thus, they cannot be applied more widely. In order to build more generic CDN infrastructures based on P2P principles that maintain the advantages of P2P (such as flexibility and dynamicity), it is necessary to take certain requirements into account. Such a system has to be, for instance, able to cope with the inherent heterogeneity of peers since a common denominator approach would make it very inefficient. Further, it has to provide scalability, be incrementally expandable and dependable in its service. It should also balance the load of requests in a way that no hot-spots occur. The aim is to develop principles and methods that can be used to build P2P based content infrastructures that can be used in a multitude of environments and cases. Eventually, the goal is to create a generic infrastructure that can support all kinds of multimedia applications, including content production and delivery networks, interactive multimedia applications, multiplayer games, etc.

This paper elaborates on a particular issue in designing overlay networks, the network stability issue. Since peers are autonomous entities, they dynamically participate in the constructed network by joining and leaving. In fact, several empirical observations of the uptime distribution (c.f.[20,21]) indicate that the majority of peers do not stay connected for long time periods. Therefore, structured approaches utilizing proactive mechanisms in order to provide efficient routing mechanisms, require significant signaling to maintain the targeted topology and update the indexing data on the advertised content. The required information exchanged in this process can be further increased if the proactive design of the system replicates the content, too. Omicron[22] addresses this issue with a two-tier architecture combining a structured approach with a clustering mechanism. Omicron constructs clusters of peers that (as a set) form reliable components to develop a stable structured overlay. In order to do so, new peers perform a selective join mechanism, so that the resulting joint reliability of each cluster is above a minimum threshold. The joint reliability of a cluster is called *endurance* to reflect the differences from single peer reliability and network stability. The selective join mechanism is based on random sampling of a subset of clusters in order to decide which one is the weakest. Several algorithms have been investigated in order to evaluate their performance on cluster coverage and the related properties.

Stable P2P networks provide the required infrastructure for different kinds of CDN to operate efficiently. Omicron's design provides the additional aforementioned requirements, such as being scalable, incrementally expandable and dependable, providing evenly distributed workload properties, dealing adaptively with potential hot-spots, supporting heterogeneous populations, etc.

This paper is organized as follows: In Section 2 an overview of the Omicron network is provided focusing on the graph structure, the clustering mechanism and the resulting architecture. The network management mechanism dealing with peers joining the network is discussed in Section 3. The investigated sampling algorithms are provided in Section 4 and the related simulation results are given in Section 5. Finally, Section 6 summarizes the paper and gives an outlook for the future.

2. OMICRON

Omicron *(Organized Maintenance, Indexing, Caching and Routing for Overlay Networks)* is a P2P overlay network aiming to address issues of heterogeneous, large-scale and dynamic P2P environments. Its hybrid, two-tier, DHT-based approach makes it highly adaptable to a large range of applications. Omicron deals with a number of conflicting requirements, such as scalability, efficiency, robustness, heterogeneity and load balance. Issues to consider in this context are:

Topology. The rational in Omicron's approach is to reduce the high maintenance cost by having a small and fixed node degree, thus, requiring small and fixed size routing tables (at least for the majority of peers), while still performing lookup operations at low costs. For this reason the usage of appropriate graph structures (such as de Bruijn graphs,[23] which are further discussed in Section 2.1) is suggested. However, while the small fixed node degree reduces the operational cost, it causes robustness problems.

Clustering mechanism. To address the robustness issue, clusters of peers are formed with certain requirements on their endurance. The clustering mechanism is described in deeper detail in Section 2.2.

Roles. A unique feature of Omicron is the integrated specialization mechanism that assigns particular roles to peers based on their physical capabilities and user behavior. The specialization mechanism provides the means to deal with peer heterogeneity. This scheme fits the contribution of each node to its resource capabilities and aims at the maximization of the cluster efficiency by providing appropriate incentives to peers to take a certain role. As it can be observed from Omicron's name, four different core roles have been identified: *Maintainers (M), Indexers (I), Cachers (C) and Routers (R).* Maintainers are responsible to maintain the overlay network topology, while Indexers handle the relevant indexing structures. Routers forward the queries towards their logical destination and Cachers reduce the overall routing workload by providing replies to popular queries. Roles are additively assigned, meaning that peers do not remove their older roles as they get new ones.

Identification scheme. A dual identification scheme has been introduced for Omicron with a number of advantages. Clusters are assigned a *Globally Unique IDentifier* (GUID) that is used to route requests over the network. Advertised items are assigned a GUID and are located at the clusters whose GUID matches best. Moreover, peers are assigned their own GUID to trace their actions in the system.

2.1. de Bruijn Digraphs

Directed graphs (digraphs) have been extensively used in interconnection networks for parallel and distributed systems design (cf.[24,25]). Digraphs received special attention from the research community aiming to solve the problem of the so-called (k, D) digraph **problem,**[26] where the goal is to maximize the number of vertices (order) N in a digraph of maximum out-degree k and diameter D. Some general bounds relating the order, the degree and the diameter of a graph are provided by the well-known Moore bound.[27] Assume a graph with node degree k and diameter D; then the maximum number of nodes (*graph order*) that may populate this graph is given by Equation 1:

$$N \leq 1 + k + k^2 + \ldots + k^D = \frac{k^{D+1} - 1}{k - 1} \tag{1}$$

Interestingly, the Moore bound is not achievable for any non-trivial graph.[27] Nevertheless, in the context of P2P networks, it is more useful to reformulate Equation 1 in a way that provides a lower bound for the graph *diameter* (D_M), given the node degree and the graph order[28]:

$$D_M = \lceil log_k(N(k-1)+1) \rceil - 1 \leq D. \tag{2}$$

The *average distance* (μ_D) among the nodes of a graph may also be bounded by the following inequality[29] (which is approximated by Loguinov et el.[30]):

$$D_M - \frac{k(k^{D_M} - 1)}{N(k-1)^2} + \frac{D_M}{N(k-1)} \approx D_M - \frac{1}{k-1} \leq \mu_D. \tag{3}$$

An interesting class of digraphs is the so-called lexicographic digraph class,[31] which includes the de Bruijn and Kautz digraphs*. de Bruijn digraphs have asymptotically optimal graph diameter and average node distance.[30] Thereby, they are employed in the design of our work. Figure 1 shows a directed de Bruijn$(2,3)$ graph denoting a graph with a maximum out-degree of 2, where the diameter length is 3 and the graph order is 8. For graphs with fixed out-degree of 2 the maximum number of nodes† is always limited by 2^D. The graph contains 2^{D+1} directed edges in this case. Each node is represented by string of length D ($D = 3$ in this example). Every character of the string can take k different values (2 in this example). In the general case each node is represented by a string such as $u_1u_2...u_D$. The connections between the nodes follow a simple left shift operation from node $u_1(u_2...u_D)$ to node $(u_2...u_D)u_x$, where u_x can take one of the possible values of the characters $(0, k-1)$. The solid lines in the figure denote links where the '0' character is shifted in, while the dotted lines denote links where the '1' character is shifted in.

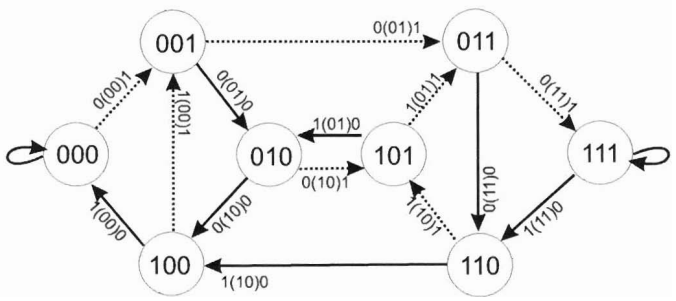


Figure 1. Directed de Bruijn$(2,3)$ graph.

de Bruijn graphs have been suggested to model the topology of several P2P systems, however, we exploited them in an innovative way. Considerable examples of P2P systems that use de Bruijn graphs are Koorde,[32] D2B[33] and Optimal Diameter Routing Infrastructure (ODRI).[30]

2.2. Clustering

Clusters have been introduced into the design of P2P systems in a variety of approaches. JXTA defines the concept of PeerGroups[34] to provide service compatibility and to decompose the large number of peers into more manageable groups. Further, SHARK[35] cluster peers based on the common interests of users. Also, Considine[36] proposes multiple cluster-based overlays for Chord. The cluster construction in the latter proposal is based on network proximity metrics aiming to reduce the end-to-end latency. Furthermore, even hierarchical approaches like eDonkey and KaZaA might be considered as clustering approaches‡ to a certain extent, where normal peers are clustered around the super-peers. The purpose of this "clustering" is to transform the costly all-to-all communication pattern into a more efficient scheme. However, by doing so it introduces additional load-balancing concerns. In fact, this is a more general issue that appears in every acyclic hierarchical organization (i.e., tree-like organization). Thus, the cluster organization must be restricted to non acyclic structures in order to provide even distribution of responsibilities.

A desirable property of each overlay network is acquiring a topology that remains as *stable* as possible over the time and minimize the related required communication cost to maintain the targeted structure. However, in highly dynamic P2P systems that consist of *unreliable* peers, the desirable stability cannot be attained. This is the most crucial motivating factor for introducing the concept of clusters in the architectural design of the Omicron overlay network. Clusters can be considered as an essential abstraction, which can be used to absorb

*de Bruijn graphs are less dense than Kautz graphs but they are more flexible since they do not have any limitations on the sequence of the represented symbols in every node.

†The Moore bound determines always maximum upper bounds on the size of the graphs that are not reachable for non-trivial cases.

‡The calculated clustering coefficient for these networks is relatively high.

the high peer attrition rate and accomplish high network stability. They can be considered as an equivalent mechanism to the suspensions used in vehicles to absorb shocks from the terrain. In order to make more clear the involved concepts, it is required to define *peer reliability*, *network stability* and *cluster endurance*. We define network stability as follows:

DEFINITION 2.1. *Network stability $S_N(t)$ is the probability that the topology of the network remains unmodified until some time t.*

A definition for peer reliability is given below.

DEFINITION 2.2. *Peer reliability $R_P(t)$ is the probability that the peer remains connected until some time t.*

Assuming that the lifespan of a peer is modeled with the random variable X, then the reliability of the peer is given by:

$$R_P(t) = Pr\{X > t\} = 1 - F(t). \tag{4}$$

where F(t) is the CDF (cumulative distribution function) of peer's lifetime.

On the other hand, a cluster is a virtual entity composed by several peers. We define the endurance of a cluster as follows:

DEFINITION 2.3. *Cluster endurance $E_C(t)$ is the probability that at least one peer of the cluster will remain connected until some time t.*

The endurance of clusters is calculated by the following equation.

$$E_C(t) = 1 - \prod_{i=1}^{K}(1 - F_i(t)), \tag{5}$$

where K is the size of the cluster and $F_i(t)$ is the CDF of the ith peer in the cluster.

Clustering algorithms aim mainly at *"partitioning items into dissimilar groups of similar items"*. They require the definition of a *metric* to estimate the similarity of the items in order to perform the partitioning procedure. Since an overlay network is a virtual network, there is a lot of freedom in defining the optimal partitioning metric. In the context of P2P overlay networks, the similarity of the peers forming an individual cluster is that all the members are responsible for the same part of the *address space*, which does not necessarily define a meaningful metric for assigning peers to clusters. Therefore, the proposed clustering algorithm requires criteria other than the usual similarity metrics. The key factors that motivate the construction/deconstruction of clusters are the following:

- Clusters should fulfill the endurance requirements.
- Clusters should have the smallest possible size in order to reduce the intra-cluster communication complexity.
- Clusters should be divided when it is possible to create d other endurable clusters, where d is the degree of the employed de Bruijn graph. Selective division should be applied to maximize the endurance of each new cluster.
- Clusters should be merged when their estimated endurance is lower than a predefined endurance threshold.

Moreover, a hysteresis-based mechanism is required to avoid oscillations in splitting and merging clusters. In order to describe the membership of peers in the clusters the ClusterMap concept has been used. A *ClusterMap* includes the peers participating in a cluster. ClusterMaps may be realized as tables collecting entries for each member peer. Every entry may hold information about peers' GUID, their role in the system, the observed reliability and other useful information that could be used by every peer of a cluster to effectively construct its local routing table. In fact, ClusterMaps are supersets of Routing Tables, including the potential peers of clusters that may become neighbors of a particular peer. ClusterMaps are periodically disseminated to neighbor clusters as well as the cluster itself.

2.3. Two-tier Network Architecture

The suggested *two-tier* network architecture is a major step towards accomplishing the fulfillment of the targeted requirements. It enables the effective usage of de Bruijn graphs by successfully addressing their shortcomings. In fact, the successful "marriage" of two different topology design techniques (in a combination of a *tightly structured macro level* and a *loosely structured micro level*) provides a hybrid architecture with several advantages.

1. **Tightly structured macro level.** Adopting the topological characteristics of de Bruijn graphs, the macro level is *highly symmetrical* enabling *simple routing* mechanisms. Composed of endurable components, it results in a relatively *stable* topology with *small diameter* and *fixed node degree*.
2. **Loosely structured micro level.** On the other hand, the micro level provides the desirable characteristics to the macro level by following a more loosely structured topology with a great degree of *freedom* in the neighbor selection. This freedom may be invested on regulating and achieving a *finer load balance*, offering an effective mechanism to handle potential *hot spots* in the network traffic. Moreover, *locality-aware* neighbor selection may be used to maximize the matching of the virtual overlay network to the underlying physical network. Finally, *redundancy* may be developed in this micro level supplying seamlessly *fault-tolerance* to the macro level.

An example of the hybrid topology is illustrated in Figure 2. The structured macro level is a de Bruijn$(2,3)$ digraph. Two nodes (representing peer clusters) are "magnified" to expose the micro level connectivity pattern between them. Two different connection types are shown: inter-cluster connections and intra-cluster connections.

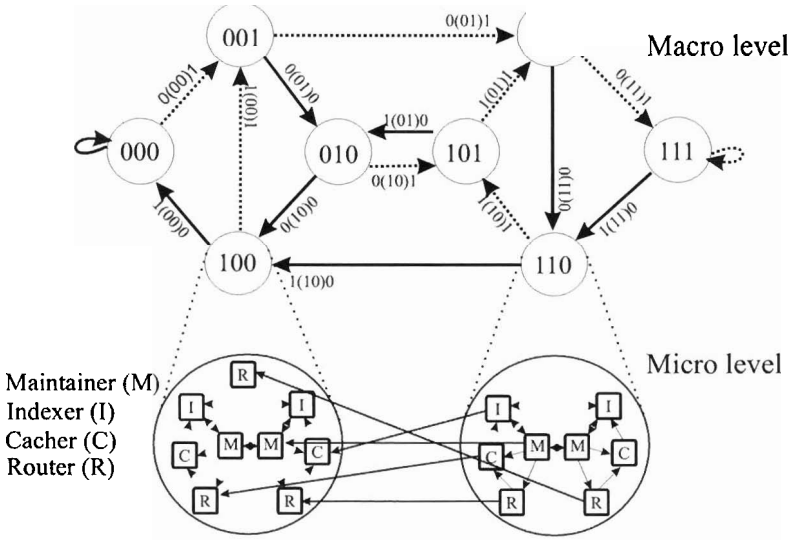


Figure 2. Omicron Overlay Network

3. OVERLAY NETWORK MANAGEMENT

In the resulting two-tier architecture, two entities are used to construct the network topology: individual peers and clusters of peers. Thereby, a set of efficient procedures need to be defined to handle the dynamic participation of the peers in the system and the resulting consequences in both the endurance and the maintenance cost of clusters. Three crucial requirements are driving the developed solutions: *dependability*, *load-balance* and *efficiency*.

When new peers request to join an Omicron-based P2P system, Maintainers perform a number of operations in order to place the new peers in the network. The purpose of their actions is to achieve a well balanced topology where clusters have sufficient endurance and the total workload is minimized and well distributed.

Obviously, the optimal selection can be made when the endurance of every cluster of the network is globally known. However, such a solution raises scalability issues as the size of the network increases considerably. Therefore, an alternative approach has been investigated where Maintainers perform a random walk collecting the endurance of each cluster in the path in order to decide which one is the best selection to direct the new peer to.

More particularly, a variety of bootstrap phases may be assumed, providing an initial online peer P_Y that triggers the mechanism to accept the newly joining peer P_X. Without loss of generality it can be assumed that P_Y has been assigned the Maintainer role (otherwise the request has to be simply redirected to another peer $\acute{P}_Y$ of the same cluster that has been assigned the Maintainer role).

Upon the reception of the joining request P_Y triggers a *sampling* procedure using an inter-cluster *random walk*. It contacts a Maintainer P_Z of a randomly selected neighbor cluster, which is recursively repeating this step making a random walk of length $w = \alpha \cdot log(C_S)$, where C_S is the number of clusters in the system and α is a weight. It should be noted that w is asymptotically equal to the diameter of the inter-cluster overlay network. The goal of the procedure is to equally distribute the new peers in the deployed clusters considering the internal state of each cluster, i.e., its endurance and its size. By performing a random walk of length at least equal to the diameter of the network, every cluster has a probability of being included in the sampling procedure of each join request. Moreover, having a logarithmic number of samples provides a fairly good approximation of collecting the state of all clusters, which would have been very costly in terms of communication traffic. Thus, the selected approach can provide a well-balanced outcome with a low cost.

After performing the sampling procedure with the random walk, the appropriate cluster is selected. In this phase, a newly joined peer is considered unreliable and it is merely assigned the Router (and optionally the Cacher) role. Thus, the selected cluster is the one with the least number of unreliable peers so that the load for the Maintainers of the clusters is fairly equal. The Maintainer of the lastly visited cluster included in the random walk indicates to the newly joined peer the selected cluster of which it should become member of (via an `Accept` message). Afterwards, P_X is asking the provided maintainer of the target cluster to connect and become a cluster member. As a reply, the Maintainer provides updated ClusterMap structures of the cluster itself and the neighbor clusters so that the P_X can correctly build its routing table.

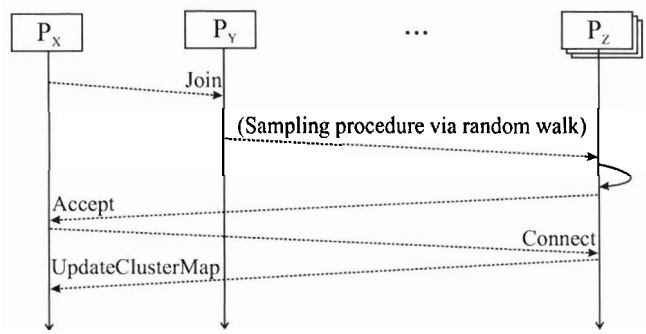


Figure 3. Join sequence diagram.

The whole process is illustrated in Figure 3 by a sequence diagram. Peer P_Z represents the Maintainers that participate in the sampling random walk. The selected Maintainer receives the *Connect* message and replies by providing the necessary ClusterMap structures. A further issue related to the network management is the way the structured macro level expands or shrinks in order to fit to the network size. For this purpose, a decentralized algorithm to split and merge the clusters is described in.[22] Moreover, a similar random walk mechanism may be applied when peers are becoming reliable enough to be assigned more critical roles (i.e., Indexers and Maintainers).

4. INVESTIGATED SAMPLING ALGORITHMS

In this section, we describe four different algorithms to accomplish effective cluster coverage at low cost. Three of them are probabilistic and one is deterministic. All algorithms start randomly at any peer, assuming that new peers randomly select their first contact to send the requests to. Even if the employed bootstrap phase does not comply with this assumption, it can be easily achieved by performing an additional random walk before the sampling phase begins.

4.1. Probabilistic Algorithms

The probabilistic algorithms differ both in length and neighbor selection policy. Their description is provided in the following list.

1. **Random destination**. This algorithm starts from any random peer, which randomly selects the final destination. This has the advantage of simplicity since it does not differ from a typical query routing procedure. However, the number of covered clusters is equal to the average query length. Therefore, the average random walk length is given by Equation 3, which is shorter than the network diameter. The achieved cluster coverage is very similar to the assigned routing workload.[37]
2. **Short random walk**. This algorithm starts from any random peer by randomly selecting only the next peer to follow among the neighbors found in the routing table. The procedure is recursively applied until a random walk of length equal to the diameter D of the network is reached. This algorithm has the advantages of (i) equal length random walks and (ii) better coverage distribution than the previous algorithm since seldom reached clusters are more likely to be visited. However, its implementation is more complex. It requires a non-oblivious routing mechanism to avoid cycles in the random walk. Clusters should be visited only once.
3. **Long random walk**. This algorithm is very similar to the previous one. The only difference is that the required length of the random walk must be twice the length of the diameter ($2 \cdot D$). It is expected that the longer random walk combined with the cycle avoidance restriction will provide a much better cluster coverage. The disadvantage of this algorithm is that it costs twice as much as the short random walk.

4.2. Deterministic R-Shift Algorithm

The aforementioned algorithms perform random walks in order to sample the endurance of the clusters. In this section, a deterministic algorithm is investigated in order to evaluate such an alternative. It is assumed that the deterministic walk begins randomly at any peer (similarly to the random alternatives).

There are certain restrictions and guidelines in designing an effective deterministic walk appropriate to effectively sample the endurance of clusters.

1. The length of each walk must be as close as possible to its maximum value (i.e., the diameter of the network).
2. The length of each walk should not differ considerably (independently of the position of the initial cluster).
3. The cluster coverage should be as wide and as evenly distributed as possible.

There are several algorithms that can fulfill the aforementioned requirements. We have designed one that is as simple as possible. It is called *"R-Shift"* algorithm. Basically, each peer deterministically select the final destination by applying a *right-shift* operation at the GUID of its cluster (note that the conventional routing in Omicron utilizes *left-shift* operations). The new symbol at the right end of the GUID must be different than the symbol at the left end before the right-shift operation. Formally, this operation can be expressed as follows.

$$rshift(u_1 u_2 ... u_D) = u_2 ... u_D \overline{u_1}, \tag{6}$$

where the $\overline{u_1}$ is an operation that provides a different symbol from the available alphabet (deterministically). For example, for binary de Bruijn graphs, it holds that $\overline{0} = 1$ and $\overline{1} = 0$. It is guaranteed that using the R-Shift

algorithm all the clusters will be included in the sampling since Equation 6 provides a direct and unique mapping of the input cluster to the output cluster.

This algorithm is graphically illustrated in Figure 4, which displays a de Bruijn(2, 4) digraph. Two different deterministic walks are shown. Assume that the two walks start at nodes (0011) and (0101), respectively. Thereby, the R-Shift algorithm produces the sequence (0011) $\rightsquigarrow$ (0110) $\rightsquigarrow$ (1100) $\rightsquigarrow$ (1000) $\rightsquigarrow$ (0001) for the first case, which is traversed by `Msg1`. Similarly, the sequence (0101) $\rightsquigarrow$ (1010) $\rightsquigarrow$ (0100) $\rightsquigarrow$ (1001) $\rightsquigarrow$ (0010) is traversed by **`Msg2`**. However, in certain cases the length of the path is smaller than the diameter, e.g., (1101) $\rightsquigarrow$ (1011) $\rightsquigarrow$ (0110).

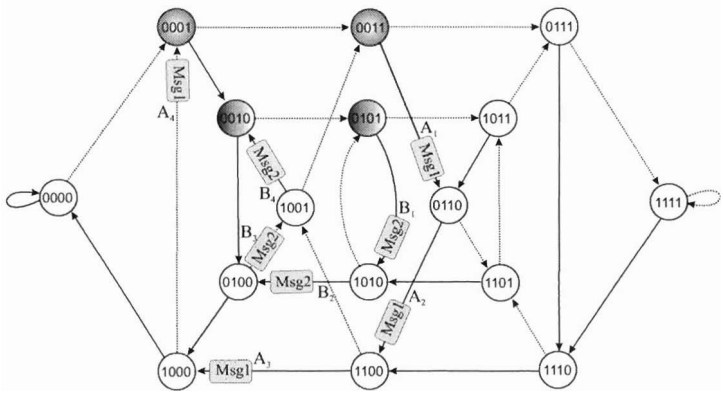


Figure 4. Deterministic R-Shift algorithm.

The described algorithms have different inter-cluster communication cost that determines the sampling walk length. Table 1 summarizes this cost.

Table 1. Sampling algorithms routing cost.

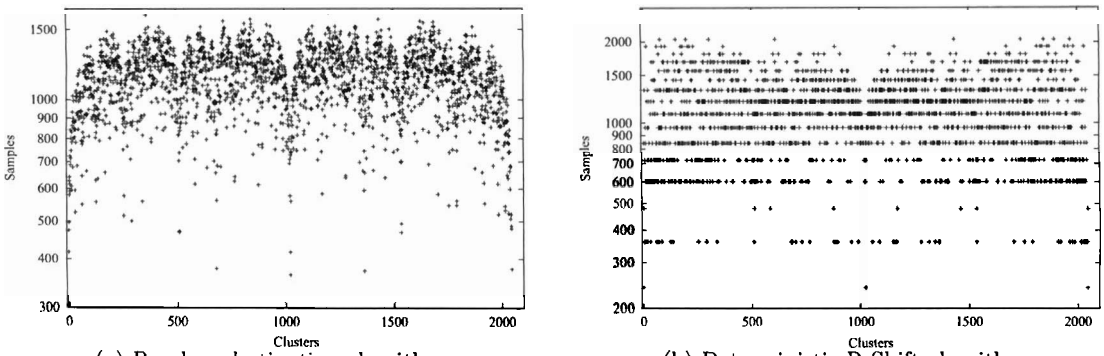
Sampling algorithm	Expected walk length
Random destination	$D_{DB} - \frac{1}{k-1}$
Short random walk	D_{DB}
Long random walk	$2 \cdot D_{DB}$
R-Shift	$\approx D_{DB}$

5. EVALUATION

The evaluation has been performed by simulation experiments using the general purpose discrete event simulator for P2P overlay networks described in.[38] The constructed overlay network forms an Omicron network. In the involved operations participate mostly the Maintainers.

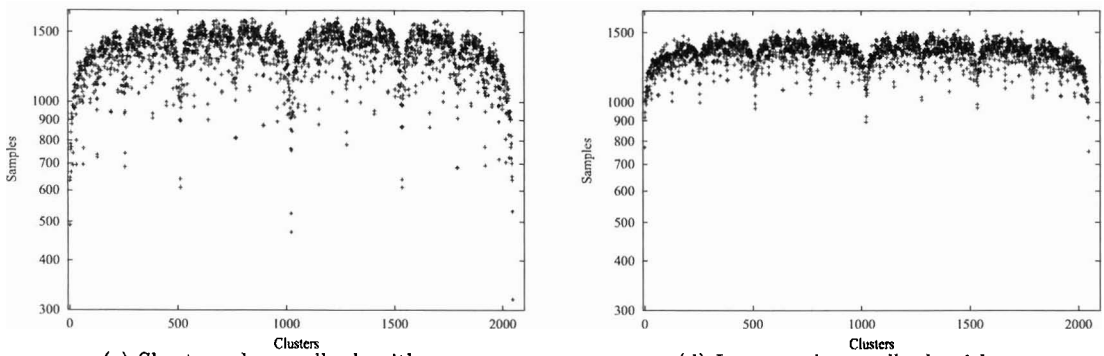
5.1. Cluster Coverage

The most critical aspect we need to evaluate is the ability of the four sampling techniques to visit evenly each cluster of the network. Figure 5 provides the results for a specific Omicron configuration where the structured de Bruijn network is composed of 2048 clusters and the inter-cluster degree is $k = 2$. Figure 5(a) describes the coverage distribution (the number of times each cluster is sampled) for the random destination algorithm. Similar results are provided in Figure 5(b), Figure 5(c) and Figure 5(d) for the R-Shift algorithm, the short random walk algorithm and the long random walk algorithm, respectively.



(a) Random destination algorithm.

(b) Deterministic R-Shift algorithm.



(c) Short random walk algorithm.

(d) Long random walk algorithm.

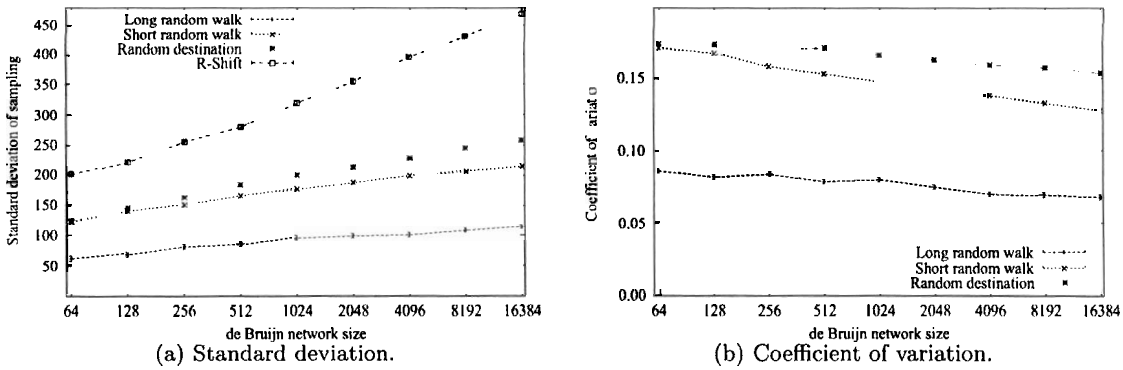
Figure 5. Sampling distribution using random walks.

As expected, the long random walk algorithm provides the most evenly distributed sampling where the majority of samples differ less than 10% from the mean value. The reason for this lies on the fact that more clusters are sampled at every join, which is combined with the cycle-removal mechanism (non-oblivious routing mechanism). Therefore, clusters that are seldom sampled by other algorithms have a higher probability of sampling by this algorithm.

Aiming at providing additional results on the ability of each algorithm to evenly sample the network clusters, further experiments have been performed. In these experiments, the size of the structured macro level (de Bruijn network) is modified between 64 and 16,384 clusters. For each experiment, a number of join requests is generated that is related to the size of the network and the utilized algorithm. The target is to generate approximately equal workload for all of the algorithms.

The quantity of interest in these experiments is the evaluation of the *standard deviation* of the cluster sampling distribution for each algorithm. Figure 6(a) summarizes the results of the experiments. It should be noted that the x-axis scales logarithmically in order to provide a more comprehensive view. As it can be observed, the long random walk algorithm achieves the smallest standard deviation for the complete range of the evaluated network sizes. Moreover, the short walk algorithm achieves better performance compared to the random destination algorithm as the network grows. The standard deviation of the R-Shift algorithm grows considerably more compared to the three probabilistic alternatives.

However, the standard deviation provides an absolute value for the effect. In many cases, it is more important to observe a relative metric that relates the standard deviation with the mean value. Such a metric is the *coefficient of variation*, which is defined as $CV = \sigma/\mu$, where σ is the standard deviation and μ is the mean



(a) Standard deviation. (b) Coefficient of variation.

Figure 6. Cluster sampling in de Bruijn networks.

value. It should be noted that the mean value of the deterministic algorithm differed from the provided mean values set. Therefore, it is not included in the provided results. Figure 6(b) summarizes the results on the coefficient of variation. As it can be observed, the long random walk algorithm accomplishes always a coefficient of variation less than 10%. Also, it is interesting to notice the decreasing rate of CV for the short random walk algorithm. It can be stated that for very large network size, the performance of the short random walk algorithm approaches asymptotically the performance of the long random walk algorithm.

5.2. Cluster Sampling Inter-arrival Distribution

The aggregated behavior of the cluster sampling algorithms can be observed with the experiments performed in the previous sections. However, it is essential to evaluate how often each particular cluster is revisited in subsequent random walks.

Self-connected clusters (i.e., with GUID 111...1 or 000...0) are the least frequently involved clusters in the routing procedure,[37] which can also be observed in Figure 5. Further, by examining the details of the collected results, it has been noticed that clusters with GUID, e.g., (011001001) or (0110010011) or (01100100110) are among the most frequently visited clusters for each sampling algorithm (GUIDs with lack of "patterns" in their digit sequence).

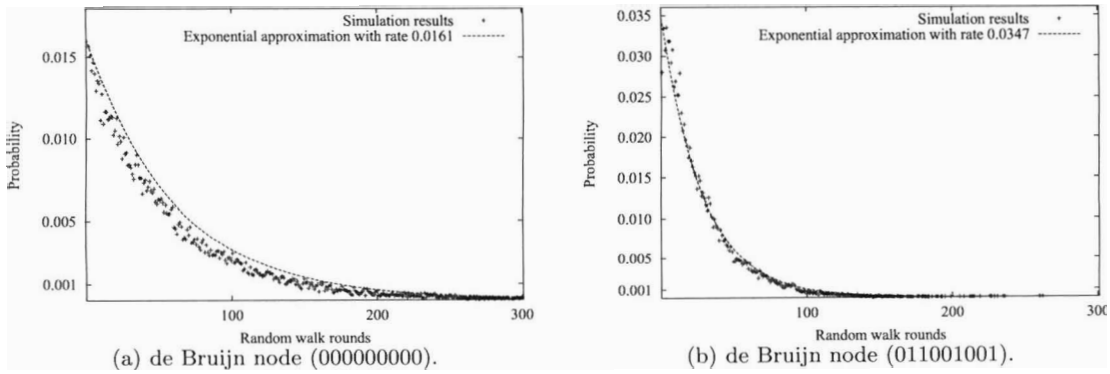
Let us define *each random walk experiment* as a *"round"*. The *event* of interest E_C is *"how many random walks (rounds) are necessary until a particular cluster C is sampled"*. Such a quantity can provide vital information on whether the sampling algorithm is adequate for its need. Therefore, further experiments have been performed aiming to evaluate E_C. Collecting the experiment results for these clusters, the diagrams of Figure 7 have been drawn to show the probability that the particular cluster will be sampled after a certain number of sampling walks. In order to generate these measurements, the long random walk algorithm has been employed.

It is interesting to observe that the cluster sampling inter-arrival distribution can be closely approximated by an *exponential* distribution of the form $f(x) = \lambda x^{-\lambda x}, x \geq 0$. The reason for such behavior can be explained as follows. Let us call V_C the event of interest, which is visiting a particular cluster C during a random walk. V_C is a *Bernoulli* random variable:

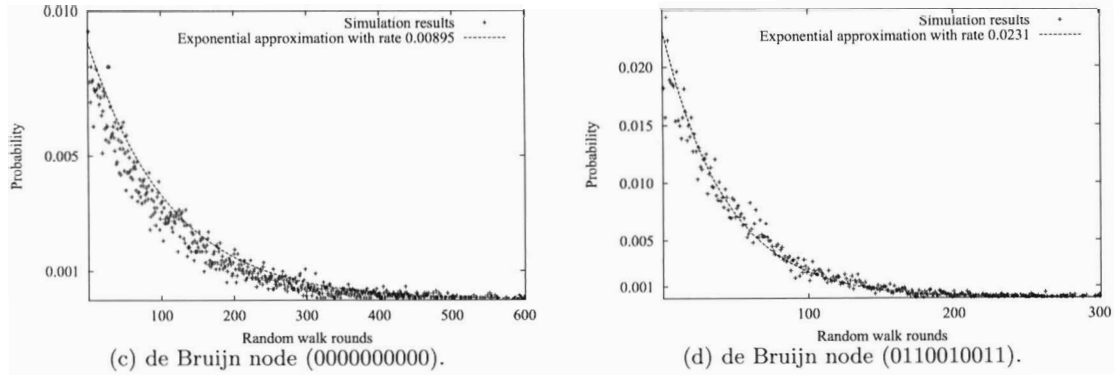
$$g(x) = \begin{cases} g(0) = Pr\{V_C = 0\} = 1 - p, \\ g(1) = Pr\{V_C = 1\} = p, \end{cases} \tag{7}$$

where p is the probability of success that depends on the cluster position in the digraph. Each random walk is an independent event. If we let X be the number of the performed events until a success occurs, then X is said to be a *geometric* random variable with parameter p. Its PMF is given by:

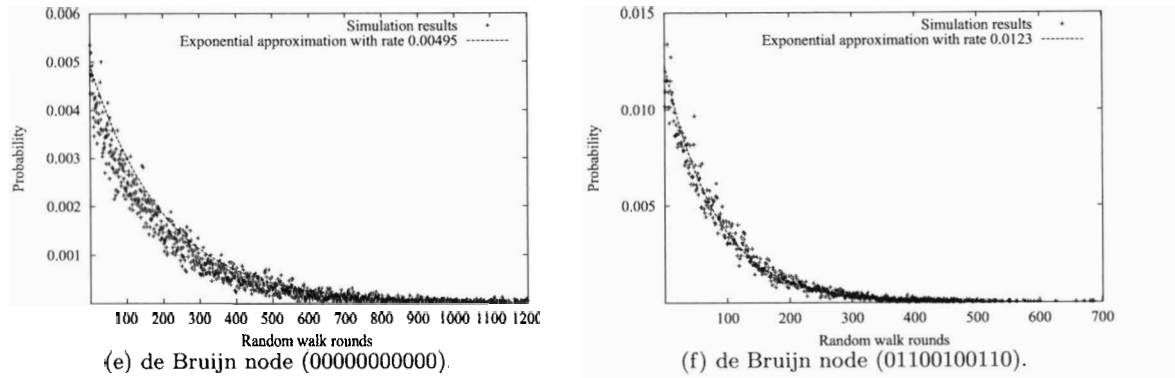
$$h(n) = Pr\{X = n\} = (1-p)^{n-1}p, \qquad n = 1, 2, \ldots \tag{8}$$



(a) de Bruijn node (000000000). (b) de Bruijn node (011001001).



(c) de Bruijn node (0000000000). (d) de Bruijn node (0110010011).



(e) de Bruijn node (00000000000). (f) de Bruijn node (01100100110).

Figure 7. Cluster sampling inter-arrival distribution.

The geometric distribution is the discrete equivalent of the exponential distribution. Therefore, the cluster sampling inter-arrival distribution can be approximated well with the exponential distribution.

As it can be seen from the approximated rates of Figure 7, seldom visited clusters have a lower rate than frequently visited clusters. Also, as the size of the network gets larger, the approximated rates are getting smaller, which is expected since more clusters are available for sampling. However, the peer join rate is getting higher, providing the necessary lower bound for the sampling rate.

6. CONCLUSIONS

Omicron is a hybrid overlay network that has been designed to meet several critical requirements for P2P systems. Omicron fits adequately to the needs of a great multitude of P2P based CDNs. In this paper, we have focused on the sampling mechanism of Omicron that has been employed in order to provide a well balanced and stable network. The maintenance overhead that is required to ensure network stability through cluster endurance has been evaluated with multiple sampling algorithms. Both deterministic and probabilistic algorithms have been employed where the latter showed better cluster coverage capabilities. In addition, the cluster sampling inter-arrival distributions have been estimated revealing an interesting distribution property that can be further exploited by analytical means to obtain a stochastically described P2P network.

In the future, we plan to investigate in detail the properties of the utilized sampling mechanisms as they are applied to different overlay networks. Sampling techniques are of a more general interest and can be applied to other quantities that cluster endurance. They fit well in the distributed nature of P2P systems where no central component exists. Their exploration is vital to develop efficient systems and frameworks that can be deployed in the context of decentralized CDN infrastructures.

ACKNOWLEDGMENTS

This work has been partly supported by the European Union under the E-Next Project FP6-506869.

REFERENCES

1. T. Plagemann, V. Goebel, A. Mauthe, L. Mathy, T. Turletti, and G. Urvoy-Keller, "From Content Distribution networks to content networks - Issues and Challenges," *Computer Communications, ENEXT Special Issue* , to appear in September 2005.
2. I. JTC1/SC29/WG11, "Coding of Moving Pictures and Audio, N4801: MPEG-21 Overview, version 4," 2002.
3. M. Day, B. Cain, G. Tomlinson, and P. Rzewski, "Internet RFC 3466: A Model for Content Internetworking (CDI)." http://www.faqs.org/rfcs/rfc3466.html, February 2003.
4. A. Oram, *Harnessing the Power of Disuptive Technologies*, O'Reilly, Sebastopol, CA, 2001.
5. "Gnutella." http://www.gnutella.com, 2005.
6. "eDonkey2000." http://www.edonkey2000.com, 2005.
7. I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A scalable Peer-to-Peer Lookup Service for Internet Applications," *IEEE Transactions on Networking* **11**, pp. 17–32, February 2003.
8. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, "A scalable Content Addressable Network," in *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 161–172, ACM Press, 2001.
9. Y. Cui and K. Nahrstedt, "Layered Peer-to-Peer Streaming," in *Proceedings of the International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV03)*, pp. 162–171, June 2003.
10. X. Jiang, Y. Dong, D. Xu, and B. Bhargava, "Gnustream: a P2P media streaming system prototype," in *Proceedings of the International Conference on Multimedia and Expo (ICME) 2*, July 2003.
11. M. Hefeeda, A. Habib, B. Botev, D. Xu, and B. Bhargava, "PROMISE: Peer-to-Peer Media Streaming Using CollectCast," in *Proceedings of the ACM Multimedia*, November 2003.
12. M. Zink and A. Mauthe, "P2P Streaming using Multiple Description Coded Video," in *Proceedings of the 30th EUROMICRO Conference*, September 2004.
13. H. Deshpande, M. Bawa, and H. Garcia-Molina, "Efficient Topology-Aware Overlay Network," in *Proceedings of the 1st Workshop on Hot Topics in Networks*, October 2002.
14. H. Deshpande, M. Bawa, and H. Garcia-Molina, "Streaming live media over a peer-to-peer network," Technical Report 2001-31, Stanford University, 2001.
15. A. Nicolosi and S. Annapureddy, "P2PCast: A Peer-to-Peer Multicast Scheme for Streaming Data," in *Proceedings of IRIS Student Workshop, MIT*, October 2003.

16. M. Castro, P. Druschel, A. Kermarrec, A. Nandi, A. Rowstron, and A. Singh, "SplitStream: High-Bandwidth Multicast in Cooperative Environments," in *Proceedings of the ACM SOSP*, October 2003.
17. G. On, *Quality of Availability for Widely Distributed and Replicated Content Stores.* PhD thesis, Technische Universität Darmstadt, Germany, June 2004.
18. M. Izal, G. Urvoy-Keller, E. Biersack, P. Felber, A. A. Hamra, and L. Garces-Erice, "Dissecting BitTorrent: Five months in a Torrent's Lifetime," in *Proceedings of Passive and Active Measurements (PAM) 2004*, April 2004.
19. L. Cherkasova and J. Lee, "FastReplica: Efficient Large File Distribution Within Content Delivery Networks," in *Proceedings of the Fourth USENIX Symposium on Internet Technologies and Systems*, March 2003.
20. F. E. Bustamante and Y. Qiao, "Friendships that last: Peer lifespan and its role in P2P protocols," in *Proceedings of the International Workshop on Web Content Caching and Distribution*, October 2003.
21. S. Saroiu, P. K. Gummadi, and S. D. Gribble, "A Measurement Study of Peer-to-Peer File Sharing Systems," in *Proceedings of Multimedia Computing and Networking 2002 (MMCN '02)*, 2002.
22. V. Darlagiannis, A. Mauthe, and R. Steinmetz, "Overlay Design Mechanisms for Heterogeneous, Large Scale, Dynamic P2P Systems," *Journal of Networks and System Management* **12**(3), pp. 371–395, 2004.
23. N. G. de Bruijn, "A combinatorial problem," in *Proceedings of the Koninklije Nederlandse Academie van Wetenshapen*, pp. 758–764, 1946.
24. F. Hsu and D. Wei, "Efficient Routing and Sorting Schemes for de Bruijn Networks," *IEEE Transactions on Parallel and Distributed Systems* **8**(11), pp. 1157–1170, 1997.
25. Z. Liu and T.-Y. Sung, "Routing and Transmitting Problems in de Bruijn Networks," *IEEE Transactions on Computers* **45**(9), pp. 1056–1062, 1996.
26. M. Fiol, L. A. Yebra, and I. A. de Miquel, "Line Digraph Iterations and the (d,k) Digraph Problem," *IEEE Transactions on Computers* **33**(5), pp. 400–403, 1984.
27. W. Bridges and S. Toueg, "On the impossibility of directed Moore graphs," *Journal of Combinatorial Theory Series B* **29**, pp. 339–341, 1980.
28. M. Fiol and A. Llado, "The Partial Line Digraph Technique in the Design of Large Interconnection Networks," *IEEE Transactions on Computers* **41**(7), pp. 848–857, 1992.
29. K. Sivarajan and R. Ramaswami, "Lightwave networks based on de Bruijn graphs," *IEEE/ACM Transactions on Networking (TON)* **2**(1), pp. 70–79, 1994.
30. D. Loguinov, A. Kumar, V. Rai, and S. Ganesh, "Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routing Distances and Fault Resilience," in *Proceedings of ACM SIGCOMM'03*, pp. 395–406, August 2003.
31. F. Bernabei, V. D. Simone, L. Gratta, and M. Listanti, "Shuffle vs. Kautz/De Bruijn Logical Topologies for Multihop Networks: a Throughput Comparison," in *Proceedings of the International Broadband Communications*, pp. 271–282, 1996.
32. F. Kaashoek and D. R. Karger, "Koorde: A Simple Degree-optimal Hash Table," in *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS03)*, February 2003.
33. P. Fraigniaud and P. Gauron, "An Overview of the Content-Addressable Network D2B," in *Annual ACM Symposium on Principles of Distributed Computing*, July 2003.
34. L. Gong, "Project JXTA: A technology overview," October 2002.
35. J. Mischke and B. Stiller, "Rich and Scalable Peer-to-Peer Search with SHARK," in *5th International Workshop on Active Middleware Services (AMS 2003)*, June 2003.
36. J. Considine, "Cluster-based Optimizations for Distributed Hash Tables," tech. rep., 2003-031, CS Department, Boston University, November 2002.
37. V. Darlagiannis, *Overlay Network Mechanisms for Peer-to-Peer Systems.* PhD thesis, Department of Computer Science, Technische Universität Darmstadt, Germany, June 2005.
38. V. Darlagiannis, A. Mauthe, N. Liebau, and R. Steinmetz, "An Adaptable, Role-based Simulator for P2P Networks," in *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods*, pp. 52–59, June 2004.