12 Overlay networks for name service

In order to enable communication between a dynamic collection of peers with given names such as "machine.cs.school.edu" over the Internet, a distributed name service must be implemented on top of this collection of peers that translates given names into IP addresses. The way this is currently handled in the Internet is server-based. Here, we present two approaches that provide completely decentralized name service. For this the following operations have to be implemented:

- p.JOIN(q, Name): peer p in the system receives a request to join the system from a peer q with identity Name.
- p.LEAVE(): peer p leaves the system.
- p.SEARCH(Name): peer p wants to obtain the IP address of the peer q in the system with $\mathsf{Name}(q) = \mathsf{Name}$.

Certainly, these operations can be implemented with the overlay networks presented in the previous lecture, but these overlay networks need an extra data mapping strategy in order to execute SEARCH. In this section, we will present two overlay networks that do not need an extra data mapping strategy. Every peer just represents itself. In this case, the previous overlay networks do not work any more because they require peer names to be hashed (or more precisely, uniformly distributed) to guarantee a good topology.

The two overlay networks presented in this section are called skip graphs [1, 3] and hyperrings [2]. The skip graphs are randomized overlay networks whereas the hyperring is a deterministic overlay network, i.e. JOIN and LEAVE have deterministic implementations.

12.1 Skip graphs

Given an infinite bit string $b = x_1x_2x_3\ldots$, we define $\mathrm{prefix}_0(b) = \epsilon$ (the empty word) and $\mathrm{prefix}_i(b) = x_1x_2\ldots x_i$ for every $i \geq 1$. Suppose that we have a (pseudo-)random hash function h assigning to each node an ID representing an infinite bit string. Given a set of nodes V, we define for every $v \in V$ and $i \geq 0$:

- $\mathrm{succ}_i(v) = \mathsf{argmin}\{w \in V \mid \mathsf{Name}(w) > \mathsf{Name}(v) \text{ and } \mathrm{prefix}_i(h(v)) = \mathrm{prefix}_i(h(w))\}$, i.e. $\mathrm{succ}_i(v)$ is the node w whose name is the closest successor of v's name (with respect to lexicographical ordering) with the same i first bits in $h(w)$ as $h(v)$, and
- $\mathrm{pred}_i(v) = \mathsf{argmax}\{w \in V \mid \mathsf{Name}(w) < \mathsf{Name}(v) \text{ and } \mathrm{prefix}_i(h(v)) = \mathrm{prefix}_i(h(w))\}$.

Notice that we view the name space as a ring here. This means for $\mathrm{succ}_i(v)$ that if there is no node w with $\mathsf{Name}(w) > \mathsf{Name}(v)$ that fulfills the prefix condition, then we associate $\mathrm{succ}_i(v)$ with the node w with smallest name so that $\mathrm{prefix}_i(h(v)) = \mathrm{prefix}_i(h(w))$. If there is no other node w in the network with that property, then we set $\mathrm{succ}_i(v) = v$. In skip graphs, the following invariants have to be kept at any time.

Invariant 12.1 *For any set of nodes V currently in the system, it holds for every $v \in V$ that v is connected to $\mathrm{succ}_i(v)$ and $\mathrm{pred}_i(v)$ for all $i \geq 0$.*

Invariant 12.1 requires that the nodes are organized in a hierarchy of doubly linked cycles, where the node names have to be sorted in every cycle, and every node participates in exactly one cycle for every $i \geq 0$. A cycle at level i is called i-cycle of i-ring, and an edge in an i-ring is called an i-edge. Skip graphs have the following properties, where n is the current number of nodes in the network:

Theorem 12.2 *If Invariant 12.1 is true and h assigns random bit strings to nodes, then the skip graph has a maximum degree of $O(\log n)$, a diameter of $O(\log n)$, and a node expansion of $\Omega(1/\log n)$, with high probability.*

Proof. The probability that some fixed node pair v and w fulfills $\text{prefix}_i(h(v)) = \text{prefix}_i(h(w))$ is equal to $1/2^i$. Hence, for $i \geq 3\log n$, it holds that

$$\begin{aligned}
&\Pr[\text{there is a node pair } v, w \text{ with } \text{prefix}_i(h(v)) = \text{prefix}_i(h(w))] \\
&\leq \sum_{v,w} \Pr[\text{node pair } v, w \text{ fulfills } \text{prefix}_i(h(v)) = \text{prefix}_i(h(w))] \\
&= \sum_{v,w} \frac{1}{2^{3\log n}} \leq n^2 \cdot \frac{1}{n^3} = \frac{1}{n} \,.
\end{aligned}$$

Hence, with high probability there is no ring of level $3\log n$ or higher. Thus, every node has a degree of at most $2(3\log n + 1)$.

Next, we bound the diameter. Consider any node $v \in V$. Our aim is to move from v to a node w with largest i so that $\text{prefix}_i(w) = 00\ldots0$. To do this, we move from $u_0 = v$ to the closest successor u_1 on the 0-ring with $\text{prefix}_1(u_1) = 0$, and in general from node u_i to the closest successor u_{i+1} on the current i-ring with $\text{prefix}_{i+1}(u_{i+1}) = 00\ldots0$. For any $i \geq 0$, it holds:

$$\Pr[\text{the distance to } u_{i+1} \text{ is } \delta] = \left(\frac{1}{2}\right)^{\delta+1} \,.$$

Hence,

$$\begin{aligned}
\text{E}[\text{distance to } u_{i+1}] &= \sum_{\delta\geq 0} \delta \cdot \left(\frac{1}{2}\right)^{\delta+1} = \sum_{\delta\geq 0} (\delta+1) \cdot \left(\frac{1}{2}\right)^{\delta+2} \\
&= \frac{1}{4}\left(\sum_{\delta\geq 0} \left(\frac{1}{2}\right)^{\delta}\right) = \frac{1}{4}\cdot 4 = 1 \,.
\end{aligned}$$

Since we know from the degree proof that there are at most $3\log n$ levels with high probability, the expected number of hops we need to perform to get from v to w is $O(\log n)$. Furthermore, going from w to the node w' with smallest bit string also takes just $O(\log n)$ hops on expectation. Hence, every node in V can reach the node w' with smallest $h(w')$ in $O(\log n)$ steps on expectation, and this can also be shown to hold with high probability. Thus, the diameter is $O(\log n)$, with high probability.

The expansion proof is involved and will not be shown here. See [1] for details. □

Routing in skip graphs

Consider the following routing strategy:

Suppose that node u is the current location of a message with destination Name. As long as $\mathsf{Name} \notin (\mathsf{Name}(\mathrm{pred}_0(u)), \mathsf{Name}(u)]$ (i.e. the message has not yet reached a node u that is the closest successor of Name), u sends the message to the node $\mathrm{succ}_i(u)$ with maximum i so that $\mathsf{Name}(\mathrm{succ}_i(u)) \leq \mathsf{Name}$ (treating the name space as a ring).

Using the diameter proof in Theorem 12.2, one can show the following result:

Lemma 12.3 *For any node $v \in V$ and any name* Name*, it takes at most $O(\log n)$ hops, with high probability, to send a message from v to the node whose name is the closest successor to* Name*.*

Joining and leaving the network

Suppose that a new node v contacts node $w \in V$ to join the system. Then w will forward v's request to $\mathrm{succ}_0(v)$ using the routing strategy above with $\mathsf{Name} = \mathsf{Name}(v)$. $\mathrm{succ}_0(v)$ will then integrate v between $\mathrm{succ}_0(v)$ and $\mathrm{pred}_0(v)$. Afterwards, v sends out two requests along the 0-ring to find $\mathrm{succ}_1(v)$ and $\mathrm{pred}_1(v)$. Once they are found, v integrates itself into its 1-ring. v then uses the 1-ring to find $\mathrm{succ}_2(v)$ and $\mathrm{pred}_2(v)$, and then integrates itself into the 2-ring. This continues until v has integrated itself into the highest possible ring containing at least 2 nodes.

Using a probabilistic analysis, one can show the following result:

Theorem 12.4 *Inserting a new node requires $O(\log n)$ time and work with high probability.*

If a node wants to leave the system, it does this by simply connecting $\mathrm{pred}_i(v)$ with $\mathrm{succ}_i(v)$ for every $i \geq 0$. This gives the following result:

Theorem 12.5 *Deleting a node requires $O(\log n)$ time and work with high probability.*

Searching

When a node v searches for a node with name Name, then it simply uses the routing strategy described above. Once a node w with $\mathsf{Name}(w) = \mathsf{Name}$ has been found, w reports its IP address back to v. This strategy has the following performance:

Theorem 12.6 *Any search operation requires $O(\log n)$ time and work with high probability.*

12.2 Hyperring

Next we consider the hyperring. Like the skip graph, also the hyperring consists of a hierarchy of rings. However, here we are much more strict about how the rings are maintained.

Suppose that we have a hyperring with n nodes. Then it consists of approximately $\log n$ levels of rings, starting with level 0. Each level $i \geq 0$ consists of approximately 2^i directed cycles of approximately $n/2^i$ nodes, which we call *rings*. All rings have the same orientation, and we require the nodes in every ring to be ordered according to their names. For every ring R at level i, two rings of level $i+1$ share its nodes in an intertwined fashion. As before, a ring at level i will be called an *i-ring*, and a level i edge will be called an *i-edge*. Consider some i-ring R and let (u, v, w, x) be four consecutive nodes on R. We say that (u, v, w, x) form an *i-bridge* (or simply a *bridge* if i is clear from the context)

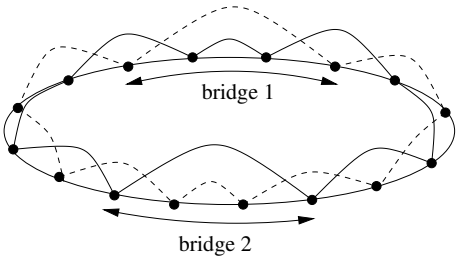

Figure 1: An example of a hyperring. The bridges have a distance of 5 from each other.

if there is an $(i+1)$-edge from u to x and an $(i+1)$-edge from v to w. An $(i+1)$-edge is called *perfect* if it bridges exactly two i-edges.

It is possible to maintain a hyperring with at most one bridge in every ring. However, in this case we would create too much update work for JOIN or LEAVE operations. Instead, we only demand that i-bridges are sufficiently far apart from each other. A hyperring is called *k-separated* if in every i-ring R the i-bridges on R are at least k nodes apart from each other, which means that there are at least $k-1$ nodes between the quadruples of nodes forming a bridge. We start with a few properties of hyperrings which are easy to prove.

Lemma 12.7 *For every $k \geq 0$, the k-separated hyperring has a maximum degree of at most $2(1 + 2/(k+1)) \log n$ and a diameter of at most $3 \log n$.*

Proof. First, we bound the maximum degree. Consider some i-ring R. In order to minimize the size of an $i+1$-ring R' on top of R without violating k-perfectness, the best one can do is using a repetitive sequence of $\lceil k/2 \rceil + 1$ edges, where one edge bridges three edges in R and the remaining $\lceil k/2 \rceil$ edges bridge two edges in R. Hence,

$$\begin{aligned} |R'| &\geq \frac{|R|}{3 + 2\lceil k/2 \rceil} \cdot (1 + \lceil k/2 \rceil) = \left(\frac{1}{2} - \frac{1}{4(\lceil k/2 \rceil + 1) + 2} \right) \cdot |R| \\ &\geq \frac{1}{2} \left(1 - \frac{1}{k+3} \right) \cdot |R| \end{aligned}$$

This also implies that $|R'|$ can be at most $\frac{1}{2}(1 + \frac{1}{k+3})|R|$. Hence, an i-ring R can have a size of at most

$$\frac{1}{2^i} \left(1 + \frac{1}{k+3} \right)^i \leq \frac{1}{2^i} \cdot e^{i/(k+3)} \leq 2^{-i(1-2/(k+3))} .$$

This is at most 1 if $i \geq (\log n)/(1 - 2/(k+3))$. Since each node has 2 edges in each level in which it participates, the maximum node degree is $2(\log n)/(1 - 2/(k+3)) = 2(1 + 2/(k+1)) \log n$.

Next, we bound the diameter. Consider any two nodes v and w on a ring R, and let R_0 and R_1 be the two intertwined rings on top of R. Furthermore, let v_0 be the node in R_0 nearest to v and w_0 be the node in R_0 nearest to w. Define v_1 and w_1 in the same way for R_1. First of all, $d_R(v, v_0)$, $d_R(v, v_1)$, $d_R(w, w_0)$, and $d_R(w, w_1)$ are all at most 1. Hence, $d_R(v_0, w_0) \leq d_R(v, w) + 2$ and $d_R(v_1, w_1) \leq d_R(v, w) + 2$. Since the nodes used by R_0 and R_1 are disjoint, it must hold that either $d_{R_0}(v_0, w_0) \leq d_R(v, w)/2 + 1$ or $d_{R_1}(v_1, w_1) \leq d_R(v, w)/2 + 1$. Hence, if we always take the

ring of lower distance in each layer, then for each layer i we obtain the recursion $d_{i+1} \leq d_i/2 + 3$ with $d_0 = d_R(v, w)$. Therefore, the total number of edges used is at most $3 \log n$. □

Unfortunately, hyperrings with constant separation can have a bad expansion.

Theorem 12.8 *For every $k \geq 0$, the k-separated hyperring has, in the worst case, an edge expansion of*

$$O(1/n^{1/(2(3(k+4))^2)}) .$$

The proof of this theorem is quite involved and can be found in [2]. Unfortunately, Theorem 12.8 implies that no k-separated hyperring with $k = O((\log n)^{1/2-\epsilon})$ for some constant $\epsilon > 0$ can guarantee an expansion of $\Omega(1/\log^c n)$ for some constant c depending on ϵ. Hence, in order to have a good expansion, we need $k = \Omega(\sqrt{\log n})$. However, notice that when k depends on the size of the hyperring, node insertions and deletions that have been performed in the past might have used a k that significantly differs from the k used by current insertions and deletions. Hence, parts of the hyperring may be out of date. So the question is whether it is necessary to revisit these parts in order to bring the hyperring up to date. Fortunately, as one of the main results in [2], it was shown that *this is not necessary*. One can simply use as the current k the degree of any node currently in the system when executing a JOIN or LEAVE operation, and old JOIN or LEAVE operations never have to be revisited, to show the following result. ($|R|$ denotes the number of nodes in a ring R, and $|e|$ denotes the number of node on the 0-ring bridged by edge e.)

Proposition 12.9 *At any time it holds:*

1. *the* ring distortion is low*, i.e. for every i-ring R, $|R| \in [\frac{1}{2} \cdot n/2^i - 1, 2 \cdot n/2^i + 1]$ and*
2. *the* edge distortion is low*, i.e. for every i-edge e, $|e| \leq 4 \cdot 2^i$.*

The proof for this is quite complicated and can be found in [2]. For simplicity, we assume for the rest of this section that k is fixed. We start with describing how to route in the hyperring.

Routing in the hyperring

We use the same routing strategy as for skip graphs:

Suppose that node u is the current location of a message with destination Name. As long as $\mathsf{Name} \notin (\mathsf{Name}(\mathrm{pred}_0(u)), \mathsf{Name}(u)]$ (i.e. the message has not yet reached a node u that is the closest successor of Name), u sends the message to the node $\mathrm{succ}_i(u)$ with maximum i so that $\mathsf{Name}(\mathrm{succ}_i(u)) \leq \mathsf{Name}$ (treating the name space as a ring).

Since this routing strategy prefers edges of higher level and every $i+1$-edge bridges at most 3 i-edges for every i, we obtain the following fact.

Fact 12.10 *Any message moves along a sequence of edges of non-increasing level and uses at most two edges in each level.*

Combining this with Lemma 12.7, which says that there are at most $(1 + 2/(k+1)) \log n$ levels, we achieve the following result.

Lemma 12.11 *For any node $v \in V$ and any name* Name*, it takes at most $O(\log n)$ hops to send a message from v to the node whose name is the closest successor to* Name*.*

12.3 Joining and leaving the network

First, we introduce some notation. Let $\mathrm{succ}_i(v)$ be the successor of v in its i-ring and $\mathrm{pred}_i(v)$ be the predecessor of v in its i-ring. For every node v on R, its $> i$-endpoints represent all endpoints of edges in v with level more than i. Notice that each node has two endpoints in each level. By "moving" the i-endpoints from u to v, we mean that we replace the i-edges $(\mathrm{pred}_i(u), u)$ and $(u, \mathrm{succ}_i(u))$ by the i-edges $(\mathrm{pred}_i(u), v)$ and $(v, \mathrm{succ}_i(u))$. By "permuting" the i-endpoints of u and v, we mean that we move the i-endpoints of u to v and the i-endpoints of v to u.

Suppose now that a new node u contacts some node $v \in V$ to join the system. Then v will forward u's request to $\mathrm{succ}_0(u)$ using the routing strategy above with $\mathsf{Name} = \mathsf{Name}(u)$. $\mathrm{succ}_0(u)$ will then integrate u between $\mathrm{succ}_0(u)$ and $\mathrm{pred}_0(u)$. Afterwards, u is integrated into the hyperring level by level, starting with level 0. In each level i, we integrate the node by either removing an already existing bridge in its $k+2$-neighborhood or by creating a new bridge. A bridge is removed by first dragging it over to u by permuting $> i$-endpoints (see Figure 3). Then case (b) or (c) in Figure 2 is applied. Otherwise, we just apply case (a). JOIN terminates once we reach a ring of size in $\{4, \ldots, 7\}$ (for larger rings, two new subrings are created).

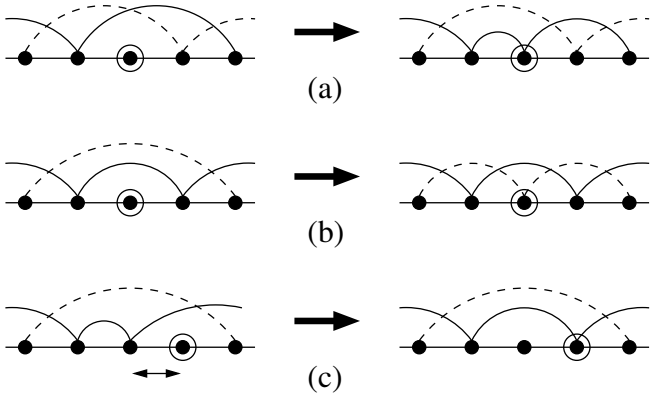

Figure 2: The three cases when adding a node. Case (c) reduces to case (b).

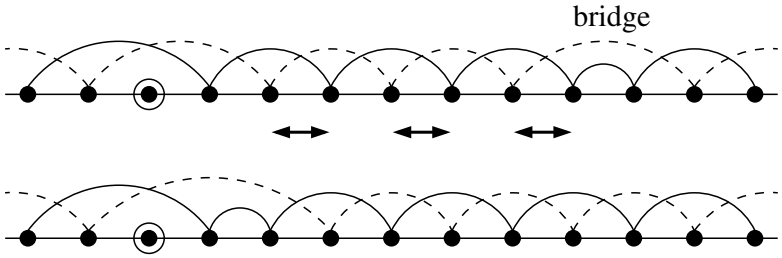

Figure 3: Permuting $> i$-endpoints drags the bridge over to obtain, e.g., case (c) in Figure 2.

Theorem 12.12 JOIN *locally preserves the k-separation of the hyperring and requires $O(k \log^2 n)$ work and $O(\log k \cdot \log n)$ time.*

Proof. JOIN locally preserves the k-separation property because it only creates a bridge if there is no other bridge in the $k+2$-neighborhood. Otherwise, it removes a bridge. Thus, it remains to prove the work and time bounds.

In each level, only a $O(k)$-neighborhood is investigated. In the worst case, a bridge has to be moved to u (resp. to the node to be integrated into that level in place of u). This requires $O(k \log n)$ message transmissions. Since the hyperring has $O(\log n)$ levels, the total work is $O(k \log^2 n)$.

When using edges in higher levels, we can investigate the $O(k)$-neighborhood of a node in $O(\log k)$ steps. Thus, in $O(\log k)$ steps we can update the endpoints necessary to proceed with the next higher level. Since there are $O(\log n)$ levels, this results in $O(\log k \cdot \log n)$ time. □

We also remove a node u from the hyperring level by level, starting with level 0. In each level, we remove the node by either removing an already existing bridge in its $k+2$-neighborhood or by creating a new bridge. A bridge is removed by first dragging it over (Figure 3) and then applying case (b) or (c) in Figure 4. Otherwise, we just apply case (a). LEAVE terminates once we reach a ring of size in $\{4, \ldots, 7\}$ (rings smaller than 4 are removed).

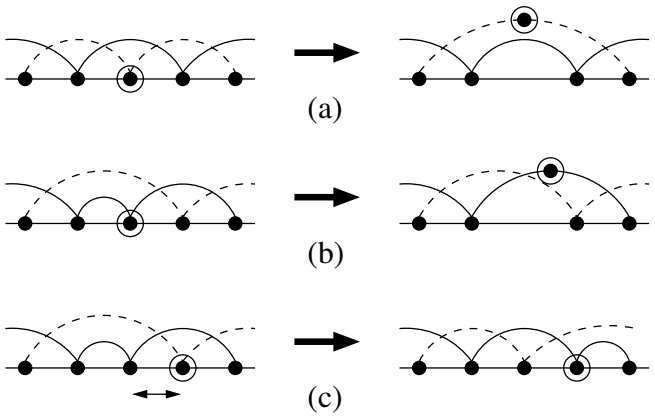

Figure 4: The three cases when removing a node. Case (c) reduces to case (b).

Theorem 12.13 LEAVE *locally preserves the k-separation of the hyperring and requires $O(k \log^2 n)$ work and $O(\log k \cdot \log n)$ time.*

The proof is similar to the proof of Theorem 12.12.

Searching

When a node v searches for a node with name Name, then it simply uses the routing strategy described above. Once a node w with $\mathsf{Name}(w) = \mathsf{Name}$ has been found, w reports its IP address back to v. This strategy has the following performance:

Theorem 12.14 *Any search operation requires $O(\log n)$ time and work with high probability.*

Furthermore, we can show the following result, demonstrating that not only the dilation but also the congestion of search requests can be kept low in the hyperring.

Theorem 12.15 *The congestion caused by n SEARCH requests, one per node, with random destinations is $O(\log n)$, with high probability.*

Proof. Fact 12.10 implies that every i-ring R can only receive requests from rings on top of it. Thus, it can only receive requests from its own nodes. Consider now an arbitrary node v in R. It is easy to check that only those requests will be sent to v whose destination is bridged by the i-edge e leaving v in R. From Proposition 12.9 we know that e bridges at most $4 \cdot 2^i$ nodes and that R consists of at most $3 \cdot n/2^i$ nodes. Since every node is the starting point of one request and every request has a random destination, the expected number of requests that want to reach v in R is at most $(4 \cdot 2^i/n) \cdot (3 \cdot n/2^i) = 12$. Combining this with the fact that every request only uses at most 2 edges in R (see Fact 12.10), the expected number of requests that traverse v in R is at most 24. Because every node participates in at most $\log n + O(1)$ levels, the overall expected number of search requests passing through v is $O(\log n)$. Using the fact that every request picks a random destination independently from other requests, one can also show that the congestion caused by SEARCH is $O(\log n)$ with high probability. □

References

[1] J. Aspnes and G. Shah. Skip graphs. In *Proc. of the 14th ACM/SIAM Symp. on Discrete Algorithms (SODA)*, pages 384–393, 2003.

[2] B. Awerbuch and C. Scheideler. The Hyperring: A low-congestion deterministic data structure for distributed environments. In *Proc. of the 15th ACM/SIAM Symp. on Discrete Algorithms (SODA)*, 2004.

[3] N. J. Harvey, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman. Skipnet: A scalable overlay network with practical locality properties. In *4th USENIX Symposium on Internet Technologies and Systems*, 2003.