Towards Voronoi-Based Backup Routing for Large-Scale Distributed Applications

Michael Grey and Markus Theil and Michael Rossberg and Guenter Schaefer
Technische Universität Ilmenau
[michael.grey, markus.theil, michael.rossberg, guenter.schaefer][at]tu-ilmenau.de

***Abstract*—The advent of critical applications that rely on Internet-based communication sheds light upon the robustness limitations of todays's Internet infrastructure. A well-known expedient are overlay applications that provide means to reroute traffic in case of network failures, but the induced overhead of existing techniques is often considered prohibitively high. Within this work we present a highly scalable backup path mechanism for large-scale distributed applications that makes use of a spherical Voronoi-based technique for peer organization and a position-dependent path selection strategy. The achieved resilience benefit of the resulting backup paths is evaluated with simulations based on real-world data, which show that connection loss in about 75% of wide-area network failures can be avoided by only two proactively selected backup peers.**

Keywords*—Voronoi, Overlay Networks, Resilience, Failover, Redirection.

I. Introduction

For many years now, the continuous pursuit of reliability improvements accompanied the evolvement of today's Internet infrastructure. However, with the advent of critical applications that rely more or less obvious on Internet-based communication, e.g., applications for automated signaling of fire, remote surgery or simply online business-logic, that pursuit becomes more relevant than ever before. Unfortunately, many of today's distributed applications often fail already in presence of shorter network glitches due to architectural limitations.

Even in case the application architecture itself does not induce pitfalls, the access network and backbone must remain reliable in presence of geographic failures. While service providers usually take good care of access structures and service providing systems, the wide-area transport network is almost always assumed to provide high reliability and sufficient failover mechanisms. In fact, not only high-impact transport network failures [17], [9] and adverse effects of political interventions [5], [3], but also the impact of sophisticated attacks on infrastructure components, e.g., attraction of traffic via compromised Border Gateway Protocol (BGP) speakers [15], confirm that this assumption does not always hold. But even if the BGP infrastructure converges, it may take up to several minutes to do so [11], which may be too late for some applications.

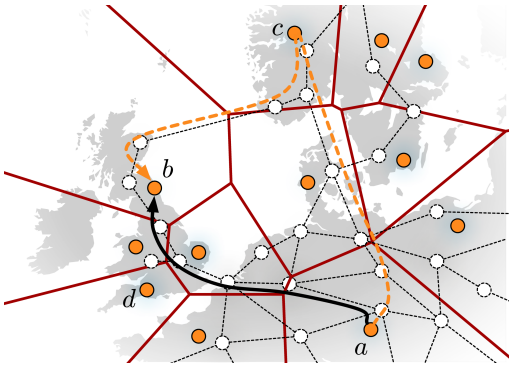

Fig. 1. As in real systems, (logical) overlay links often share physical connections. In case of infrastructural failures on the transport network path between overlay nodes a and b, an indirect backup path via node c is promising (while a path via node d is not). Structuring the overlay according to geographic locations, e.g., according to a Voronoi structure, can significantly ease the efficient search for promising backup paths.

A well-known expedient is the use of overlay applications that span an additional, manageable logical layer on top of the transport network, i.e., the Internet, and reroute traffic in case of network failures [1]. However, the overhead is often considered prohibitively high. To reduce this overhead we exploit the fact that most critical outages in wide-area networks are known to induce geographically correlated failures, e.g., as a result of disasters [4], [12], [2] or terroristic acts [16]. Thus, location-based approaches are generally promising to predetermine a possible backup routing. Unfortunately, the wide-area network architecture does not offer capabilities to make use of router or cable locations. In fact, service providers usually consider the exact geographic position of their infrastructure to be confidential.

With the increasing speed of backbone networks, non-varying parts of packet delays are, however, dominated by the signal propagation time in global networks. Thus, measured latencies between endpoints correlate with their geographic distance and geolocation information may be recovered within certain bounds, i.e., in order of a few hundred kilometers. As a result, robustness gains can be achieved by realizing dynamic overlay routing capabilities, exploiting automatically derived geographic information. Participating nodes thereby get a chance to avoid connectivity loss due to transport network failures by redirecting traffic over different parts of the network. Such a scenario is illustrated in Fig. 1.

In particular, we contribute the following to provide a highly scalable backup path mechanism for large-scale distributed applications:

- We state objectives for a location-based network overlay that allows to construct suitable backup paths and use them to survey existing approaches in Sec. II.
- In Sec. III, we present a spherical Voronoi-based technique for peer organization, which is later on used for the selection of geographic backup paths (Sec. IV).
- We evaluate the proposed method and the achieved resilience benefits with simulations based on real-world data in Sec. V.

The article concludes in Sec. VI.

II. System Objectives & Related Work

To cope with correlated transport network failures or even inter-domain routing attacks, and to be yet generally usable, a rerouting scheme shall meet the following objectives:

- **Robustness:** Preserving the connectivity between arbitrary participants even in case of transport network failures is considered the primary requirement. Thus, the system itself must not rely on external services or central entities, as they may become unavailable. The involved protocols must be designed in a self-stabilizing way.
- **Scalability:** As distributed applications may involve thousands of participants, the technique shall scale with this number. A direct conclusion is, that it should depend on local (in terms of geographic and network distances) knowledge only. Also, exposed instances, e.g., coordinators, shall be avoided as they may become bottlenecks. Moreover, the induced communication overhead should be limited, i.e., transmissions for lookups and maintenance should be reasonable.
- **Universal deployment:** The proposed mechanisms must be simple and universally applicable, e.g., for IoT scenarios. Thus, extensive computational, structural complexity, or prerequisites of network-specific aspects should be avoided.

When comparing the objectives with state-of-the-art approaches, two main areas of work must be considered: approaches to increase application robustness by introducing indirect overlay routing and overlay networks on the basis of geographic node locations.

A. Overlay Networks for Indirect Routing

The foundations on indirect routing via overlay networks emerged already 20 years ago. Improving connection characteristics by an indirect overlay routing mechanism was first introduced by the Detour project [18]. The authors not only drew essential conclusions on various routing implications in wide area networking, but also provided a study on performance comparisons between direct and alternate (indirect) paths where it was shown that for almost every direct path, there is at least one indirect path that is superior in terms of latency, loss or available bandwidth. However, these results are at least partly attributable to specialties of the referred scenario.

Similar to the Detour project, the prominent Resilient Overlay Networks (RON) presented in [1] aimed at increasing end-to-end path performance by making use of indirect overlay paths. For that to happen, each peer maintains virtual connections to all overlay nodes, where connection characteristics are determined by active probing, i.e., measuring packet loss and latency. The samples are shared with other peers and are used for overlay routing decisions. Due to the maintenance of a virtual full-mesh, the size of RON is limited to about 50 peers.

To our knowledge there have been no noticeable findings in this field that are more recent.

B. Geographic Overlay Networks

Motivated by the requirements of so-called Networked Virtual Environments in Massive Multiplayer Online Games, the authors of [8] proposed the Voronoi Overlay Networks (VON) that aim at relaxing the bottleneck due to centralized server components. From a specific participants viewpoint, other participants within an area of interest are structured according to a Voronoi structure. Notably, each participant is only required to maintain knowledge on its direct neighbors, i.e., to calculate a globally correct Voronoi structure. In detail, the authors describe basic join, move and leave procedures that are necessary for an overlay organization based on local knowledge.

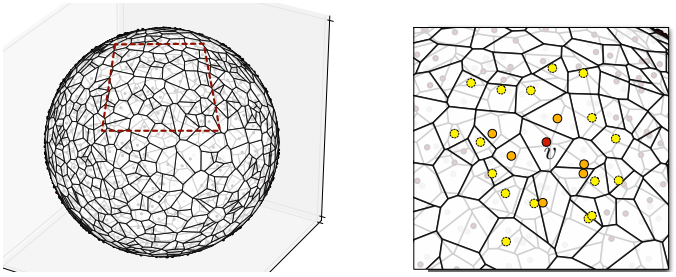

Fig. 2. A Voronoi-graph on a sphere surface with random node positions is illustrated on the left side. A magnified image section depicts the neighborhood of overlay node v. Both the 1-hop and the 2-hop neighborhood are relevant for the local view of node v.

Another remarkable method for overlay organization named *Geodemlia* [7] basically adapts Kademlia [13] for application on a sphere surface. Each node within a Geodemlia overlay divides the surrounding space in multiple areas where each area is defined by a circular segment. Neighboring nodes are then assigned according to both the calculated angle towards a reference direction and their spherical distance towards the local node. Geodemlia demonstrates interesting means to increase the efficiency of overlay searches via sophisticated space separation. Though, the resulting space separation does not reflect node area responsibilities required by this work and the proposed overlay maintenance must be considered complex.

However, these approaches do not aim at increasing the robustness of the upper layer applications.

III. Voronoi-Based Peer Organization

Taking both, intuitive considerations and state of the art into account, spherical Voronoi structures exhibit several characteristics that make them most suitable to construct overlay networks that follow geographic properties. Given the locations of participating nodes V on the sphere surface representing the earth, a Voronoi structure partitions the surface into a set of non-overlapping surface areas such that all positions within the area of a node $v \in V$ are closer to v than to any other node.

With geographic node locations as organizational key for Voronoi overlays, we thus directly obtain distinct areas of responsibility, which are guaranteed to be consistent by construction (given a stable node state). Moreover, Voronoi structures implicitly feature basic area-balancing characteristics in reference to the size of the areas that nodes are responsible for: On the one hand, given a geographically dense group of nodes, the according areas of responsibility are considerably small. On the other hand, sparsely occurring nodes lead to related areas that are comparatively large. Please note that spanning the Voronoi overlay on a spherical surface rather than on the basis of a map projection is not only motivated by elegance. Relevant map projections lead to imbalanced Voronoi areas and require additional computational effort at the map boundaries.

A. Definition

A Voronoi overlay is formally defined by a graph $G = (V, E)$. Due to the duality between Voronoi graph and Delaunay triangulation, the set of edges $E \subseteq \binom{V}{2}$ is defined as follows:

$$e = \{v_1, v_2\} \in E \Leftrightarrow v_1 \text{ and } v_2 \text{ are Delaunay neighbors}$$

To address the key requirements of scalability and robustness we rely on a distributed calculation of the Voronoi area. Thus, each participating node $v \in V$ only maintains a partial view, i.e., it knows about a sub-graph $G_v \subseteq G$.

B. Basic Functionality

Following the definition, a calculation of local Delaunay neighborhoods can be seen as a foundation of the techniques presented in the further course of this article. The core functions of the spherical Voronoi network, which must be implemented so they only depend on local knowledge, are:

1) Bootstrapping: As already presented in [8], a joining node v may obtain at least one remote peer with the help of well-known mechanisms, e.g., predefined nodes, DNS-based or cloud services. Afterwards, a message is routed via this remote peer where the target is the Voronoi region that contains the position of v. The peer associated to that region as well as the neighborhood peers, i.e., vertices defining the 1-hop neighborhood in the Voronoi overlay, reply their positions to v. Node v can then calculate its own neighborhood and connects to the according nodes.

2) Topology Control: The local neighborhood within the overlay is periodically refreshed by each node, including a broadcast of its local neighborhood to the surrounding neighbors. To proactively mitigate erroneous states, e.g., due to node churn or failure, we rely on a simple soft-state behavior: All positioning messages contain an epoch counter. Referring to the contained counter, a receiving peer sets a timeout for the according source peer. This timeout is reset on receiving of another update originated by the according source. Implicitly, in case a timeout occurs, the assigned peer is assumed not operational or to not remain in the local neighborhood anymore, e.g., due to changes in position. The key advantage of soft-state protocols in unpredictable network environments is well-known: In contrast to most hard state schemata, they do not depend on the reliability of other peers and thus are considerably robust against malfunction, while being simple to implement. In theory, the topology control only requires the local maintenance of a 1-hop neighborhood. However, taking the robustness requirements into account, we rely on 2-hop neighborhoods in the further course of this work. Wider neighborhoods must be considered inefficient in practical arrangements and thus are out-of-scope.

3) Position-Based Greedy Routing: An efficient routing towards defined positions within the spanned Voronoi overlay may be simply realized using greedy strategies, as long as the churn rate does not exceed the neighborhood update rate. For our approach, we rely on a simple distance-minimizing algorithm. Thus, messages are forwarded to the neighbor with minimum distance towards the target. This is continued until the target is located within the current nodes' Voronoi region, in which case the message is handled locally.

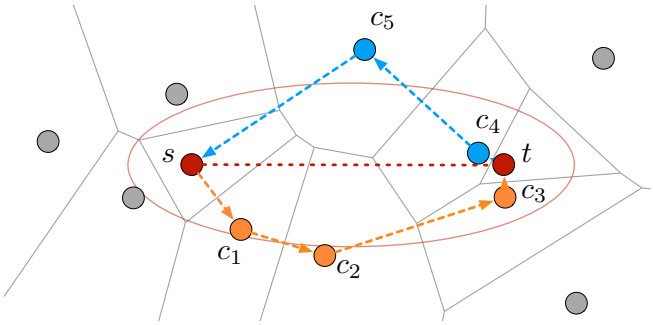

Fig. 3. Exemplary lookup based on an indication path to find a backup mediator for the s-t-connection. Candidates c_1, c_2, c_3 are traversed on the forward path, while c_4 and c_5 are found on the reverse direction.

4) Overlay Geocast: Besides message forwarding towards single locations, direct addressing of topological areas is a core functionality for the upper layer applications in scope. This is done with the help of a reverse path forwarding method: A source node v_s initiates an *area request* that is first forwarded towards the target area A via greedy routing. As soon as the request reaches an edge node within the target area, the message is forwarded towards all its neighbors within the area. Nodes in A relay the message to all neighbors also in A and discard duplicates by utilizing a bloom filter. There are alternative methods that avoid message redundancy, e.g., the authors of [10] propose a method that involves the 2-hop neighborhood and a pre-calculation of expected reverse paths, i.e., leading to a forwarding tree. However, this method – as well as comparable approaches – are not robust against node churn, which is why we rely on the simple approach outlined above.

IV. Geographically Redundant Backup Paths

Keeping the introduced overlay functionality in mind, we return to the original problem: Addressing the limited robustness in transport networks, we provide a strategy that proactively elects and maintains backup paths that protect against connectivity loss in case of outages of direct communication paths.

Finding potential mediators, i.e., peers that can possibly be used to reroute traffic in the case of network failures, is comparatively straightforward. However, predetermining a *minimal set* of mediators that are able to mitigate outages for a reference connection with high probability is more complex.

Several approaches for this task could be envisaged. *Rule-based knowledge* can be introduced to select mediators according to the specific regions of nodes, e.g., defined by country boundaries. These rule sets require either a manual definition or at least a predefined, static formulation using some semi-automatic derivation. Due to means for a specific handling of geographic and topological properties, a well-modeled rule set is the most promising approach for specific scenarios with static node sets. The approach implies problems to universal deployability, though. Using generic geometric shape definitions, *preferential zones*, i.e., geographic regions that a mediator should lie in, may be automatically derived to reflect areas of maximum expected potential. Notably, we evaluated *elliptic zones* to be best suited. Searching mediators with the help of lookup zones is feasible from a functional viewpoint. However, the communication overhead due to addressing zones, broadcasting in the zones and collecting replies is not negligible

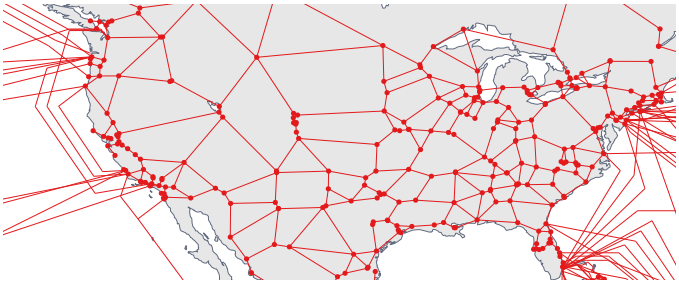

Fig. 4. Synthetically generated topology instance [6], Excerpt: North America

in larger networks. A slightly different approach involves finding mediators along an *indication path*, i.e., a path that is divergent to the direct geodesic line between source and target.

As illustrated in Fig. 3, the indication path from source s towards target t is found step-by-step where each node v forwards the lookup request to the neighbor that both:

1) Reduces the distance towards the target t and
2) Minimizes the deviation $w(v)$ from the reference distance metric of a reference path, i.e., in case an ellipse with focal points s and t; for each point p on the ellipse the sum $d(s,p) + d(p,t)$ equals the major axis length:

$$w(v) = |d(s,v) + d(v,t) - d(s,t) - d_{ref}|$$

This method is most efficient if the s-t direction and the vice versa direction enforce paths with differing orientation to the reference connection. Using elliptical reference paths is beneficial due to their distance to both nodes and the geographical corridor between them. After returning to the originator, a mediator can be chosen from the set of candidates that is defined by the previously traversed overlay nodes. Please note, that forwarding decisions involve the calculation of the distance towards the geodesic line between s and t. Using the 3-dimensional plane equation on the according great circle, the distance can be calculated by first projecting the point in question on the great circle. Given the endpoints $s = (\varphi_s, \lambda_s)$, $t = (\varphi_t, \lambda_t)$ and the candidate node $c = (\varphi_c, \lambda_c)$ as well as their projections on the unit sphere $v_s = proj(s)$, $v_t = proj(t)$ and $v_c = proj(c)$, the normal vector n on the plane is defined by $n = v_s \times v_t$. The great circle projection c_p of candidate node c is derived as follows:

$$v_c^p = v_c - \langle v_c, n \rangle \cdot n$$
$$c_p = \frac{v_c^p}{\|v_c^p\|}$$

V. Evaluation

To assess the proposed approach, key characteristics with respect to the objectives from Sec. II are discussed first. Afterwards, the achieved failure resistance is subject to a simulative evaluation based on real-world PlanetLab data.

Robustness: The proposed approach does not introduce exposed entities and does not rely on external services. Furthermore, the system tolerates fail-stop errors as well as correlated failures. This is achieved by relying on local knowledge only and implementing soft-state protocols, which offer high tolerance against unreliability of peers and networks.
Scalability: Both the induced computational effort and the communication overhead basically depend on the sizes of neighborhoods to be managed. Due to the Voronoi-based

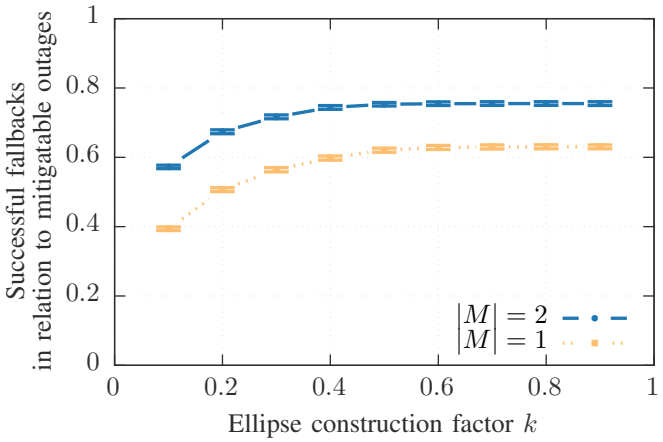

Fig. 5. Amount of successfully mitigated path outages, i.e., due to an unaffected fallback path, in relation to path outages that are potentially mitigatable by an indirect overlay path. Mean values over 32 simulation runs are shown where error bars denote 99% confidence intervals.

organization, the number of each nodes' neighbors depends on the spatial distribution of nodes, but may be assumed constant in practical scenarios, i.e., does not depend on the overall number of nodes. Also, the absolute effort for managing overlay connections is considerably small due to the use of lightweight protocols. Thus, the system scales over the number of nodes.
Universal Deployment: The Voronoi-based peer organization is comparatively lightweight. Due to the distributed organization scheme that solely relies on local knowledge, the computational and structural complexity is negligible.

Given the proposed backup path scheme, the achieved outage resistance as well as the induced load are the most interesting metrics. Both are evaluated with the help of quantitative measurements. To obtain reproducible yet significant results, a simulative study was performed making use of 530 real *PlanetLab* site locations. To evaluate the resilience against (large-scale) geographic outages, we refer to a previously developed backbone model [6], which is capable to generate network models with the characteristics of real-world physical topologies. To illustrate these topologies, Fig. V shows an exemplary network in North America. For the simulation, overlay nodes, whose position is derived from PlanetLab sites, are directly connected to the closest point of the infrastructure. For comprehensibility, we make use of distance-weighted shortest path routing in the transport network topologies.

In addition to the model of the backbone infrastructure, an evaluation of the backup path scheme requires a failure model. As we aim at understanding the behavior in presence of disasters with wide-area impact rather than the resistance against simple and easily mitigatable link failures, the outage model must represent that type of geographically correlated failures. Hence, outages are modeled by random circular-shaped areas (cf. [14]) on the sphere with radius $r_o \sim exp(\lambda)$, where $\lambda = \overline{r_o}^{-1}$. For the actual evaluation case we chose $\overline{r_o} = 200\,km$ where the results are restricted to an interval $r_o \in [r_{min} = 25\,km, r_{max} = 500\,km]$.

In view of the mediator selection, the reference distance d_{ref} of the ellipsis is the most relevant parameter. As a static parameterization is not considered reasonable, we choose d_{ref} in reference to the spherical distance between the original path's endpoints where $d_{ref} = k \cdot d(s,t)$. The construction factor $k \in$

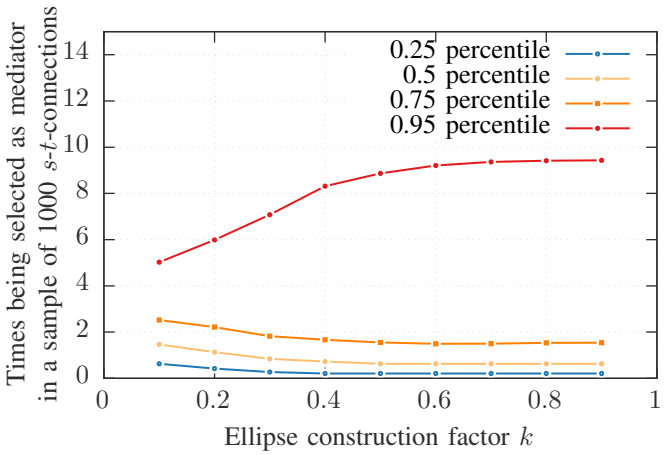

Fig. 6. Times a node is selected as mediator in a sample of 1000 s-t-connections as a function of k. Mean (0.25, 0.5, 0.75, 0.95)-percentiles over 32 simulation runs are shown.

$(0, 1]$ is a configuration parameter. Moreover, we evaluate the selection of one backup path mediator per reference connection ($|M| = 1$) versus the selection of two mediators (i.e., $|M| = 2$, where we try to select the mediators with different orientations towards the bee-line between s and t). Accordingly, Fig. 5 illustrates the amount of successfully mitigated path outages as fraction of path outages that are potentially mitigatable by an indirect overlay path as a function of k. The graph indicates that too small values for k (and d_{ref} as a result) lead to unsatisfying results. However, apart from the number of mediators, the achieved quantity of successful fallbacks does not significantly vary for values of $k \geq 0.5$. The maximum mean value (at $k = 0.5$) for a single path mediator is approx. 62% while the maximum value for two mediators is slightly above 75%. In comparison, randomly choosing a single mediator leads to a mean value of 13%, while choosing two random mediators we obtain success rate of 23%.

To evaluate the overhead being induced for each peer we consider the additional load that is generated by the scheme. The overall number of to-be-managed backup connections is easily estimated, i.e., in case of $|M| = 1$ one backup path to be managed for each reference connection, but the distribution over dedicated participants depends on the mediator selection strategy. Fig. 6 illustrates the times a node is selected in a sample of 1000 randomly selected reference connections that are protected by a backup path ($|M| = 1$). The graph depicts the mean 25%-, 50%- (median), 75%- and 95%-percentiles dependent on k. Notably, even for large k, 95% percent of nodes are selected only 10 times. For small values of k, nodes are more equally burdened. This is attributable to long-range reference connections: With increasing parameter k the resulting backup paths are increasingly concentrated on mediators in sparsely populated areas in northern and southern latitudes. However, an intolerable burden for the small remaining number of exposed nodes can be easily avoided during path setups, e.g., by managing a performance-dependent maximum.

VI. Conclusion & Future Work

Within this work, we present a backup mechanism for large-scale distributed applications that is based on a spherical Voronoi-based overlay network for peer organization. The evaluation illustrated that the lightweight approach achieves significant resilience benefits in presence of geographically correlated outages, e.g., induced by disasters. Notably, by solely relying on local knowledge the method is highly scalable and lightweight. However, the proposed backup path selection can be further improved, e.g., by tuning the lookup strategy or by augmenting the decision process with additional path information that are obtained by sophisticated path probing.

In future, we plan to evaluate advanced backup path selection strategies and load-dependent peer organization by using more versatile structures, e.g., power diagrams on the sphere surface. Furthermore, the proposed technique will be combined with a distributed approach for geographic position estimation, where the resulting system is expected to significantly increase the reliability of wide-area overlay applications without relying on external position information.

References

[1] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient Overlay Networks. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*, 2001.

[2] D. Belson. Internet impacts of hurricanes harvey, irma, and maria. *Dyn Research Blog, https://dyn.com/blog/internet-impacts-of-hurricanes-harvey-irma-and-maria, September*, 2017.

[3] D. Belson. The migration of political internet shutdowns. *Dyn Research Blog, https://dyn.com/blog/the-migration-of-political-internet-shutdowns, January*, 2017.

[4] J. Cowie, A. Popescu, and T. Underwood. Impact of hurricane katrina on internet infrastructure. *Report, Renesys*, 2005.

[5] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescapé. Analysis of country-wide internet outages caused by censorship. In *ACM SIGCOMM conference on Internet measurement*. ACM, 2011.

[6] M. Grey, M. Theil, M. Rossberg, and G. Schaefer. Towards a model for global-scale backbone networks. In *ICC*. IEEE, 2015.

[7] C. Gross, D. Stingl, B. Richerzhagen, A. Hemel, R. Steinmetz, and D. Hausheer. Geodemlia: A robust peer-to-peer overlay supporting location-based search. In *Conference on Peer-to-Peer Computing (P2P)*. IEEE, 2012.

[8] S.-Y. Hu, J.-F. Chen, and T.-H. Chen. Von: a scalable peer-to-peer network for virtual environments. *IEEE Network*, 20(4), 2006.

[9] G. Iannaccone, C.-n. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot. Analysis of link failures in an ip backbone. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. ACM, 2002.

[10] J.-R. Jiang, Y.-L. Huang, and S.-Y. Hu. Scalable aoi-cast for peer-to-peer networked virtual environments. In *Distributed Computing Systems Workshops, 2008. ICDCS'08.* IEEE, 2008.

[11] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence. *ACM SIGCOMM CCR*, 30(4), 2000.

[12] S. LaPerrière. Taiwan earthquake fiber cuts: a service provider view. *NANOG39, Febraury*, 5, 2007.

[13] P. Maymounkov and D. Mazieres. Kademlia: A peer-to-peer information system based on the xor metric. In *International Workshop on Peer-to-Peer Systems*, pages 53–65. Springer, 2002.

[14] S. Neumayer, A. Efrat, and E. Modiano. Geographic max-flow and min-cut under a circular disk failure model. *Computer Netw.*, 77, 2015.

[15] O. Nordström and C. Dovrolis. Beware of BGP attacks. *ACM SIGCOMM CCR*, 34(2):1–8, 2004.

[16] A. Ogielski and J. Cowie. Internet routing behavior on 9/11 and in the following weeks, 2002.

[17] A. Popescu. Deja vu all over again: Cables cut in the mediterranean. *Renesys Blog*, 19, 2008.

[18] S. Savage, T. Anderson, A. Aggarwal, D. Becker, N. Cardwell, A. Collins, E. Hoffman, J. Snell, et al. Detour: Informed Internet Routing and Transport. *IEEE Micro*, 19(1), 1999.