# *GENI Topology Design*

*GDD-06-27*

## *GENI: Global Environment for Network Innovations*

September 24, 2006

Status: Draft (Version 0.1)

Note to the reader: this document is a work in progress and continues to evolve rapidly. Certain aspects of the GENI architecture are not yet addressed at all, and, for those aspects that are addressed here, a number of unresolved issues are identified in the text. Further, due to the active development and editing process, some portions of the document may be logically inconsistent with others.

This document is prepared by the Backbone Working Group.

Editors:

Jennifer Rexford, *Princeton University*

Contributing workgroup members:

Jennifer Rexford, *Princeton University*

We'd also like to acknowledge comments and suggestions from the Backbone Working Group and Deep Medhi.

In this short note, we discuss the design parameters of the GENI network, including the motivation for having around 25 backbone nodes and 200 edge sites connected to the backbone by tail circuits.

Subjecting experimental network architectures to realistic traffic and network conditions is one of the main goals of the GENI facility. Certainly, researchers may start evaluating their new architectures via controlled experiments with synthetic traffic, emulated network topologies, and artificial network events. Yet, the purpose of the GENI facility is to allow them to gradually subject their architectures to increasing realism, such as carrying real user traffic, handling unexpected network events, and operating at scale. This enables the experiments run in GENI to uncover unanticipated interactions that would not arise in controlled experiments in a simulator or testbed. In fact, some GENI experiments may evolve into long-running deployment studies that offer new services to end users. Running realistic experiments and attracting real users requires a facility that can offer the kind of performance the users would expect on the Internet.

The desire for realism is major factor in the design and sizing of the GENI facility, though this goal must be tempered by the need to limit cost. As an illustrative example, consider a straw-man design that places all GENI components (such as backbone nodes, wireless/sensor subnets, and compute clusters) at a single location, with connections to end users and legacy services via dedicated tail circuits and upstream connections to the Internet. In this model, artificial delay could be added to emulate the propagation delays in a realistic backbone network. The appeal of this approach is the reduction in deployment and management costs by placing all of the equipment in a single location. However, a centralized facility would have several serious shortcomings:

- High latency: Connections to end users and legacy sites would experience unusually high latency, due to propagation delay to and from the central site.

- High backhaul costs: Nearly all of the dedicated tail circuits would traverse large distances, making them extremely expensive.

- Limited spectrum: The wireless and sensor subnets may not have sufficient spectrum, and would likely interfere with each other.

- Poor robustness: A failure that disconnects the central site from the rest of the Internet would render the entire GENI facility unusable.

These issues lead us to design a distributed facility. To reduce propagation delay and backhaul cost, we envision having a distributed collection of backbone nodes with tail circuits to nearby GENI edge sites and upstream connections to the legacy Internet. In selecting a backbone topology for the facility, we look to the rules of thumb that drive the design of commercial Internet Service Provider backbones, and the large research and education networks like the National Lambda Rail (NLR) and the Abilene Internet2 backbone, including:

- Bounding delay for interactive applications: Backbone networks typically a sufficiently rich topology such that the end-to-end propagation delay between each pair of sites is small enough

to support interactive applications, auch as telephone calls and video conferencing. This limits end-to-end paths to around 100 msec of delay.

- Keeping delays within a small factor of physical distance: Commercial ISPs typically try to limit the end-to-end propagation delay for each pair of backbone sites to some small multiplier (e.g., 2X) of the "air miles" between the sites. Otherwise, a competing ISP with a direct link between the same two locations could offer much lower latency. For most transport protocols, achieving high throughput requires low propagation delay, making propagation delay an important consideration even for elastic applications like Web browsing.

- Path diversity: Many experiments with new network architectures capitalize on the presence of multiple paths between a pair of sites; some architectures even need multiple link-disjoint or node-disjoint paths. For example, some architectures perform load balancing by splitting traffic over multiple paths, whereas others switch from one path to another in response to congestion or equipment failures. In addition, the ability of the GENI facility itself to survive node and link failures depends on the underlying diversity of the backbone network.

- Underlying fiber paths: The existing fiber-optic map in the United States imposes limits on the specific backbone sites that can have a direct fiber-optic connection between them. Placing backbone nodes in the key cities where multiple fiber-optic connections are available is extremely important to reduce the cost and deployment time of GENI. In addition, though it is possible to provide the illusion of dedicated links between any pair of backbone sites, providing links that match the underlying fiber map reduces cost and offers a more realistic deployment scenario.

- Major interconnection points: Deploying GENI backbone elements at existing interconnection points where other ISPs have their backbone sites would allow GENI to amortize the costs of space, power, and "hands and eyes" support. Locating GENI backbone nodes at major exchange points would be useful for acquiring upstream connectivity to the legacy Internet; similarly, having GENI backbone nodes at major aggregation points (such as the GigaPoPs) would facilitate efficient, low-cost connectivity to edge sites, such as university campuses.

- Tail circuit cost: The tail circuits connect to edge sites that house compute clusters, wireless / sensor subnets, and end users. The cost of these tail circuits depends, in large part, on the length of the circuit and the presence of existing fiber. Most major campus and enterprise sites already have tail circuits to the legacy Internet and perhaps also to research / education networks like the Abilene Internet2 backbone and the National Lambda Rail (NLR). Many of the university campuses connect via 15-20 GigaPoPs that provide connectivity, making it attractive to locate GENI backbone sites at or near these locations.

All of these issues point to having a backbone design that is similar to existing commercial ISPs, which have around 20-30 major sites spread throughout the country, to provide low latency, path diversity, a good match with the underlying fiber infrastructure, sufficient peering points with other providers, and economical tail circuits. In addition, some GENI experiments may want the illusion of a less "backbone centric" kind of network architecture, with direct links between edge sites. Although in practice the fiber map may not have direct connectivity between pairs of edge sites, a 20-30 node GENI backbone can easily support the embedding of virtual links that connect pairs of edge sites. Having a rich topology that closely matches the

fiber map, with short tail circuits to edge sites, would make that illusion as close to a reality as possible.

The number and location of edge sites determines how well GENI can support realistic experiments with new distributed services and backhaul end-user traffic to/from the GENI facility. We also plan to have several deployments of wireless and sensor subnets, but not at every site. Hence, the need to support distributed services and the backhaul of user traffic are the major drivers for the number of edge sites, as follows:

- Realistic distributed services: Many distributed services, such as content distribution networks (e.g., Akamai), are deployed over several hundred locations. This enables users to communicate with a nearby server for better performance. Having a relatively large number of edge sites with compute clusters is important for accurately representing the kinds of environments distributed services expect. Having a large number of sites helps attract real users to the services deployed on GENI; with only a small deployment, new services would not be able to offer sufficient performance to compete with legacy services.

- Existing research in distributed services: During the past few years, distributed systems has become an increasingly empirical research discipline. This is due, in large part, to the availability of shared, distributed facilities such as PlanetLab. PlanetLab currently has several hundred nodes spread throughout the globe. GENI can offer these researchers a higher degree of both realism and control than PlanetLab does today. However, providing a similar level of geographic distribution of the sites is important, too, for attracting these researchers and, in turn, attracting real users to the new services they create.

- Hierarchy of compute clusters: Having compute clusters at multiple locations creates a hierarchy of clusters that an individual user may experience. For example, students at a university may access a service at a GENI compute cluster on their campus, at another site in the same city (e.g., connected to the same GENI backbone site), and so on. We envision new distributed services would offer good performance by exploiting the hierarchy of available resources, and invoke clever algorithms for storing and generating content and for directing end users to the appropriate site.

In summary, to make GENI an attractive facility for deploying and evaluating new distributed services, GENI needs to have a few hundred sites spread throughout the country. Otherwise, services running on GENI would not be able to achieve the kinds of performance available by existing distributed services, such as content distribution networks and peer-to-peer systems. Having around 200 edges sites and 25 backbone sites leads to an average of eight tail circuits terminating at each backbone site, which is a reasonable number to support.

Another important issue in the design of the GENI facility is the sizing of the compute clusters and the links; the GENI planning group is still working out these sizing issues. Sizing the resources for each compute cluster, tail circuit, and backbone link ultimately devolves into a trade-off between cost and the desire to support a large number of simultaneous experiments (each with sufficient resources). For example, we initially envision a backbone built of 10 Gbps links, which would offer enable (say) a thousand simultaneous experiments that consume 10 Mbps on every link. Fortunately, the provisioning of bandwidth for GENI follows a nice "virtuous cycle": if successful, GENI would attract further deployment of additional bandwidth

by various entities (e.g., government agencies, companies, etc.) to enable their experiments and long-running services. Still, we must be cautious not to under-provision GENI at the beginning, to prevent a "success disaster" where GENI has the flexibility to support many novel architectures and services, but not sufficient resources to run them at reasonable performance.