# IT 2035C Network Infrastructure Management

## Design Scenario 4: Genome-4U

Group: Samuel Hazlett, Beka Martin

Genome-4U is a scientific research project at a large university in the United States. Genome-4U has recently started a large-scale project to sequence the genomes of 100,000 volunteers with a goal of creating a set of publicly accessible databases with human genomic, trait, and medical data.

The project’s founder, a brilliant man with many talents and interests, tells you that the public databases will provide information to the world’s scientific community in general, not just those interested in medical research. Genome-4U is trying not to prejudge how the data will be used because there may be opportunities for interconnections and correlations that computers can find that people might have missed. The founder envisions clusters of servers that will be accessible by researchers all over the world. The databases will be used by end users to study their own genetic heritage, with the help of their doctors and genetic counselors. In addition, the data will be used by computer scientists, mathematicians, physicians, social scientists, and other researchers.

The genome for a single human consists of complementary DNA strands wound together in a double helix. The strands hold 6 billion base pairs of nucleotides connected by hydrogen bonds. To store the research data, 1 byte of capacity is used for each base pair. As a result, 6 GB of data capacity is needed to store the genetic information of just one person. Storage will be required for these data sets and the raw nucleotide data so the project plans to use network-attached storage (NAS) cluster.

In addition to genetic information, the project will ask volunteers to provide detailed information about their traits so the researchers can find correlations between traits and genes. Volunteers will also provide their medical records. No information that identifies the volunteers will be made available to the public.

You have been brought in as a network design consultant to help the Genome-4U project.

1. List five major user communities, four from the scenario and one user community not mentioned in the scenario. Describe how the new user community can benefit from access to these data stores.

* Medical Researchers:

Benefit: They can identify genetic markers associated with diseases, leading to improved diagnostics, early detection, and targeted treatments.

Impact: This enables the development of gene-based therapies, personalized treatment plans, and a deeper understanding of hereditary conditions. Physicians and genetic counselors can also use this data to guide patient care and risk assessments.

* Computer Scientists:

Benefit: They can apply machine learning and artificial intelligence to analyze massive genomic datasets, uncover hidden patterns, and optimize data processing.

Impact: Their work can improve predictive models for genetic diseases, create more efficient ways to store and retrieve genomic data, and contribute to advancements in bioinformatics and computational biology.

* Mathematicians:

Benefit: They can perform statistical analysis to detect trends and relationships between genetic variations and traits.

Impact: Their findings help validate genetic research, improve accuracy in identifying risk factors for diseases, and contribute to population genetics studies by mapping inheritance patterns across generations.

* Social Scientists:

Benefit: They can study the role of genetics in human behavior, cultural evolution, and societal diversity.

Impact: This research can lead to a better understanding of how genetics influence intelligence, personality traits, and predispositions to social behaviors, contributing to fields like psychology, anthropology, and ethics.

* Pharmaceutical Companies:

Benefit: They can use genomic data to develop personalized medicine, creating drugs that work better for specific genetic profiles.

Impact: This reduces trial-and-error prescribing, minimizes adverse drug reactions, and increases the effectiveness of treatments for conditions like cancer, autoimmune diseases, and rare genetic disorders. Additionally, pharmaceutical companies can use this data to accelerate drug discovery and design clinical trials tailored to genetic subgroups.

1. Use a chart to list the major data stores and the user communities you named in question #1. Determine datastore are will used by which user communities.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| User Group | Medical Researchers | Computer Scientists | Mathematicians | Social Scientists | Pharmaceutical Companies |
| Medical Data | x | x |  |  | x |
| Traits | x | x | x | x | x |
| Genome | x | x | x | x | x |

1. Quality of service (QoS) requirements. You will not be able to precisely characterize the traffic but evaluate expected traffic.

**Flow:**

The network traffic is expected to be bi-directional. Researchers and end-users will be both uploading large genomic datasets to the storage systems (especially raw nucleotide data) and downloading them for analysis. Volunteers' data (medical records and traits) may also need to be uploaded into the system from multiple remote locations. Therefore, the flow will involve data moving in and out of the NAS clusters, as well as between various global research teams accessing the data.

**Type:**

Since the dataset contains large genomic files (6 GB per person), most of the traffic will involve large, bulk data transfers. This includes uploads of raw genomic sequences, trait data, and medical information from volunteers, as well as downloads by researchers for analysis.

While the bulk of the traffic will be large file transfers, there will also be real-time or low-latency traffic for tasks like querying specific genetic data, running statistical analysis, or using computational tools that rely on real-time data. These might require a more responsive network to ensure that users can interact with the data without noticeable delays.

**Quality of Service**

1.) Bandwidth**:** The network will require sufficient bandwidth to handle the large data volumes, particularly for the uploads and downloads of genomic data. Given that each person’s genetic data is about 6 GB, with 100,000 volunteers, the network needs to support a substantial amount of traffic both inbound and outbound.

2.) Latency: Some users may require low-latency communication for accessing or analyzing data, especially if conducting real-time collaborative analysis. For those accessing the data from remote locations (globally), low latency will be essential to ensure a seamless user experience when querying or interacting with the data.

3.) Jitter: Network jitter (variability in latency) should be minimized, especially for research tasks that require the transfer of multiple data packets in a continuous stream or real-time queries for specific genetic data sets.

4.) Packet Loss: Since the data involved is highly sensitive, minimizing packet loss is crucial. Any packet loss could lead to data corruption, which could severely impact research and lead to erroneous results. Ensuring that the network can handle these large data transfers without losing information will be critical.

5.) Prioritization: Some data flows, such as real-time analysis requests or communications between researchers, may need to be prioritized over bulk transfers. The network will need to implement traffic prioritization, ensuring that high-priority tasks such as real-time queries or database updates are given precedence over less time-sensitive data transfers.

1. What questions would you ask Genome-4U’s founder about this project? What questions would you ask the network administer concerning the network design? Provide two questions for each person interviewed.

Note: What additional information will help you in efficiently designing this network?

**Questions for Genome-4U's founder:**

1.) What are the key performance metrics and scalability goals for the network, given that the project may scale to include even more volunteers beyond the initial 100,000?

This question helps to understand the founder’s expectations for future growth and how the infrastructure should scale to handle increasing data volumes and users.

2.) What is the budget and timeline for implementing the infrastructure, and are there any restrictions or constraints on the technologies we can use (e.g., cloud services, on-prem servers, etc.)?

This question clarifies financial and time constraints, which are crucial for designing a cost-effective and feasible network solution.

**Questions for the Network Admin:**

1.) What current network architecture and hardware do we have in place, and are there any limitations or bottlenecks we need to be aware of when handling large datasets like the genomic information?

This helps assess the starting point and whether the existing network infrastructure can handle the expected traffic, or if upgrades are needed.

2.) What is the expected number of concurrent users accessing the database and how do we handle different types of data requests (e.g., large file downloads vs real-time data queries)?

This question will provide insight into the expected traffic patterns and guide the design of the network to ensure smooth operation for both data-heavy and real-time access requirements.

Some additional information we could acquire for an efficient network design would be data storage requirements, security requirements, geographic distribution of users, and backup and recovery plans. This will help provide a clear view of the project’s infrastructure needs. Therefore, helping in designing a network that is robust, scalable, and secure for their long-term success.

Teamwork

This submission was completed as a team effort by Samuel Hazlett and Beka Martain, our team member Sidi Mohammed has yet to respond to any attempt of contact by both team members since the first week of the semester. We have notified the professor.