THE PDS GEOSCIENCES NODE'S INFORMATION TECHNOLOGY INFRASTRUCTURE. Daniel M. Scholes[1], Thomas C. Stein[1], and Lars E. Arvidson[1], [1]McDonnell Center for the Space Sciences, Department of Earth and Planetary Sciences, Washington University in Saint Louis, 1 Brookings Drive, Campus Box 1169, St. Louis, Missouri, 63130, scholes@wustl.edu

Introduction: The Geosciences Node of NASA's Planetary Data System (PDS) archives and distributes science data related to the study of surfaces and interiors of terrestrial planets and their moons [1]. The data, along with tools and expert advice on their use are provided at no cost to scientists, educators, and the public. This paper focuses on the Geosciences Node information technology (IT) infrastructure developed over decades of data curation, growth of Node tools and services, improvements in technology, and experience.

Requirements: The PDS Geosciences Node IT infrastructure design is informed by the PDS level 1 requirements [2], for example, providing guidance and assistance to missions, instrument teams, and data providers in designing and documenting observational data products and archives. The Geosciences Node currently supports six active missions (Lunar Reconnaissance Orbiter, Mars 2020, Mars Express, Mars Odyssey, Mars Reconnaissance Orbiter, and Mars Science Laboratory), six developing missions (Dragonfly, Europa Clipper, Lunar Trailblazer, Veritas, VIPER, and the Mars Sample Return Program), numerous Artemis Program Commercial Lunar Payload Services missions, and some 40 individual investigators. The Node curates more than 350 TB of PDS archives online with approximately 30 TB added each year, supported by high-speed networks and processing capacity for data integrity and validation checks against PDS standards.

Archived data must be discoverable by and accessible to users. The Analyst's Notebook (AN) [3] and Orbital Data Explorer (ODE) [4], web applications created by the Geosciences Node, incorporate relevant archive product metadata from multiple Nodes and employ high-transaction database and map service engines for the best user experience.

The PDS ensures the long-term preservation of the archives using both online and offline storage options to maintain archive copies. As with the primary data store, sufficient processing capacity is needed to actively monitor and verify the integrity of secondary archive copies.

Design Considerations: Security, performance, and cost are key drivers in the Node's IT infrastructure design. Securing the Node's infrastructure and protecting the archived data are critical, as the Node is the primary storage location and online repository. Processing capacity and performance, network throughput, and storage speed and capacity must be sufficient for the Node to meet its requirements.

The Node's host institution, Washington University in St. Louis offers IT opportunities that benefit the Node, such as access to its high-speed university network and Internet 2, and a dedicated Node-specific data center space. It maintains software-licensing agreements with industry leaders, offering discounted costs for enterprise level software including operating systems, database platforms, and geographic information system (GIS) software. In addition, the university has preferred hardware retailers and commercial cloud providers who provide discounted pricing.

Implementation: The Geosciences Node operates a virtual server environment with network attached storage in a dedicated on-campus data center. The data center is physically secured with monitored door locks, door switches, security cameras, temperature monitors, motion sensors, and water detection, and it is equipped with a dedicated cooling system and on-site uninterrupted power supply (UPS). Housing storage and processing in the same data center allows more efficient scanning, cataloging, and access to the archives for internal and external use. Security considerations are prioritized throughout the IT infrastructure.

Compute Environment. A cluster of five Dell PowerEdge R940 servers hosts a virtual server environment, allowing software-level operating system images to run as independent servers that share the resources of the physical hosts. The cluster comprises a total of 526 Ghz across 160 processor cores and 10 TB of RAM, and hosts approximately 140 virtual machines (VMs) via VMware management software. A secondary VM host in a separate server room is used for disaster recovery. The cluster supports VM load balancing for optimal performance utilization and provides redundancy against hardware failures. VM storage and processing profiles are seamlessly updated as performance needs change. New virtual servers, including development environments and server clones, can be rapidly added and removed, as needed. The VMs are backed up using VEEAM software.

Data Storage. The Node uses multiple tiers of storage systems to meet the needs of specific processes. A high-performance 66 TB Dell Unity solid-state storage array houses data requiring the fastest read and write capability: VM virtual disks, SQL databases, and

GIS mapping data. More than 700 TB of PDS archives, pre-processed and support files for web applications, and general workspace are maintained on a high-capacity Dell Isilon network storage system that incorporates multi-tiered storage and includes built-in functionality to migrate frequently accessed data to faster storage pools. Data can be allocated to specific tiers within the Isilon system, such as keeping website support files in a higher-performance pool to support faster website response. Both storage systems are expandable without taking services offline and include layers of redundancy to sustain operations through hardware or network failures.

The Node maintains multiple copies of the data archives to ensure restoration capability in the case of corruption or disaster. The primary archive version is stored on-site on the Dell Isilon network storage system and is used by Node compute and web services. A second copy is held in Azure online blob storage, and it is synced through file-level updates from the primary copy. A third copy is written to AWS Glacier using commercial backup software (CommVault). In addition, archives are transferred to the NASA Space Science Data Coordinated Archive (NSSDCA) as required by PDS.

Internal and External Network Connectivity. The Geosciences Node infrastructure comprises multiple network segments. A 25 Gb network within the primary server room handles internal traffic between virtual server hosts that support web services, databases, processing servers, and workstations. Data transfers to users on the Internet2 network are carried over the Washington University Research Network, which allows our inbound and outbound traffic to by-pass the university's user traffic on a dedicated 10 Gb network. Other external data transfers are relegated to commodity networks.

Two onsite IBM Aspera [5] servers host high-speed external data transfers, each capped at 1 Gb/s. Most data providers deliver their files to the Node via Aspera, and ODE offers an Aspera web browser plug-in for single-click download of data orders.

Security Controls. The Geosciences Node follows PDS security requirements and industry-best practices (NIST 800-53 [6]) for securing its systems. Staff members remotely access internal resources via a virtual private network (VPN) with two-factor authentication. Redundant 10 Gb hardware firewalls protect the Node's local network with advanced features such as real-time scanning for malware, viruses, intuition detection, denial of service, and Geo-IP blocking. Servers and workstations are protected by active virus protection and centrally managed operating system updates.

Additional external monitoring and protection tools are provided by the university. OpenDNS software blocks malicious web traffic from penetrating the network, and monthly vulnerability scans are run against external facing surfaces. Node staff are alerted to potential issues identified by the university's scans. Monitoring tools query Node web services to ensure that applications are accessible outside of our network, provide expected results, and produce satisfactory response times.

Data Discovery & Distribution Services. The Geosciences Node website provides HTTP and FTP access to our archives, hosted on a dedicated server secured behind the Node's firewall. Node web application (ODE, AN, and Spectral Library) operations, including cataloging metadata, front-end services, on-demand processing, and order request fulfillment, are supported by a suite of high-performance web, database, map, and compute servers.

Cloud Adoption: We adopt new and alternative technologies as they make sense. We reduced operation costs beginning in 2019 by migrating system backups from an on-premises tape library to the cloud. At present, hosting core operations within our infrastructure is more cost-effective than hosting these functions in the commercial cloud. Resources such as data center space, cooling, and network usage are provided by the university at no additional cost. Cloud processing costs are twice as expensive as our on-site virtual server environment, even with discounted rates available through the university.

Outcome: The Geosciences Node IT infrastructure is a cost-effective solution that enables the Node to successfully complete its requirements. The VM environment, tiered storage, and network redundancy serve all aspects of Node operations with a 99.99% scheduled uptime.

Acknowledgments: The PDS Geosciences Node operates with funding from NASA. Cooperation from mission science and operation teams, the PDS Project Office and other Nodes, and Washington University IT is greatly appreciated.

References: [1] Slavney, S. et al. (2019), 50th LPS, Abstract #1685. [2] PDS Management Council (2017, April 20), https://pds-engineering.jpl.nasa.gov/sites/default/files/pds_level123_requirements_20170420.pdf. [3] Stein, T.C. et al. (2010), 41st LPS, Abstract #1414. [4] Bennett, K. et al. (2008), 39th LPS, Abstract #1379. [5] Scholes D. et al. (2018), 49th LPS, Abstract #1235. [6] National Institute of Standards and Technology (2020), doi:10.6028/NIST.SP.800-53r5.