MOVING DATA: Getting up to speed with Globus and Science DMZ

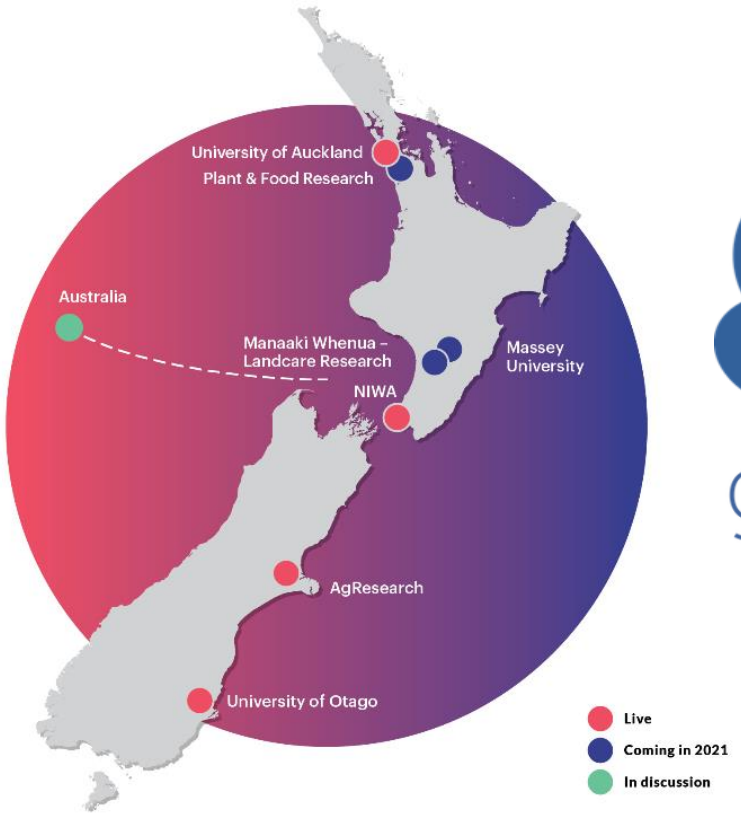
Brian Flaherty
Richard Tumaliuan

NeSI
New Zealand eScience Infrastructure

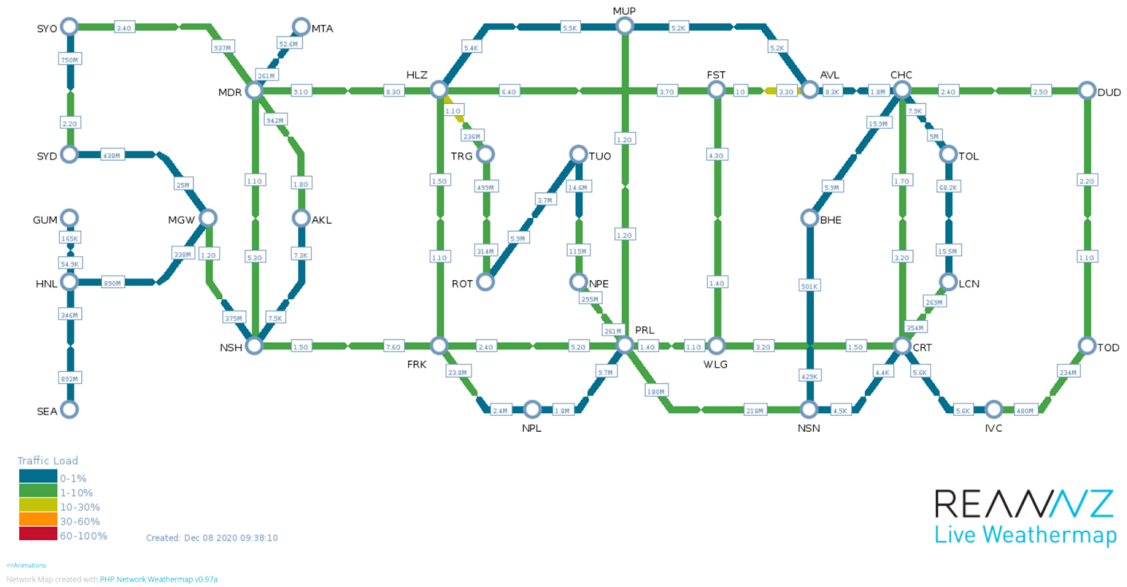
REANNZ

Overview of the tools that support and enable fast, secure, and reliable transfers with Globus and Science DMZ

How these tools fit together to improve performance across the whole data transfer journey



globus



Science DMZ

A lightweight and high performing on-ramp to the REANNZ international research and education network.

An opportunity for discussion groups to delve deeper into the user experience of Globus and the technical aspects of Science DMZ

Globus break out group - Show Globus users and potential Globus users how they can get value from the service, what is the difference between managed vs. personal endpoints? How users can engage with more advance transfer options and the steps for getting started.

globus

Science DMZ break out group – technical aspects of the Science DMZ, transfer node reference design, and the REANNZ managed network edge and Science DMZ. Examples and technical requirements.

Science DMZ

A lightweight and high performing on-ramp to the REANNZ international research and education network.

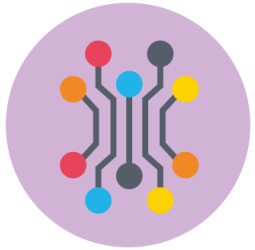
NATIONAL DATA TRANSFER PLATFORM

Globus – the world wide network of DTNs

NeSI
New Zealand eScience Infrastructure

REANNZ

NeSI Services



High performance computing (HPC) and analytics

- Fit-for-purpose national HPC platform including data analytics

Data transfer and share

- High speed, secure data transfer with end-to-end integration
- Hosting of large actively used research datasets, repositories, and archives

Training and researcher skill development

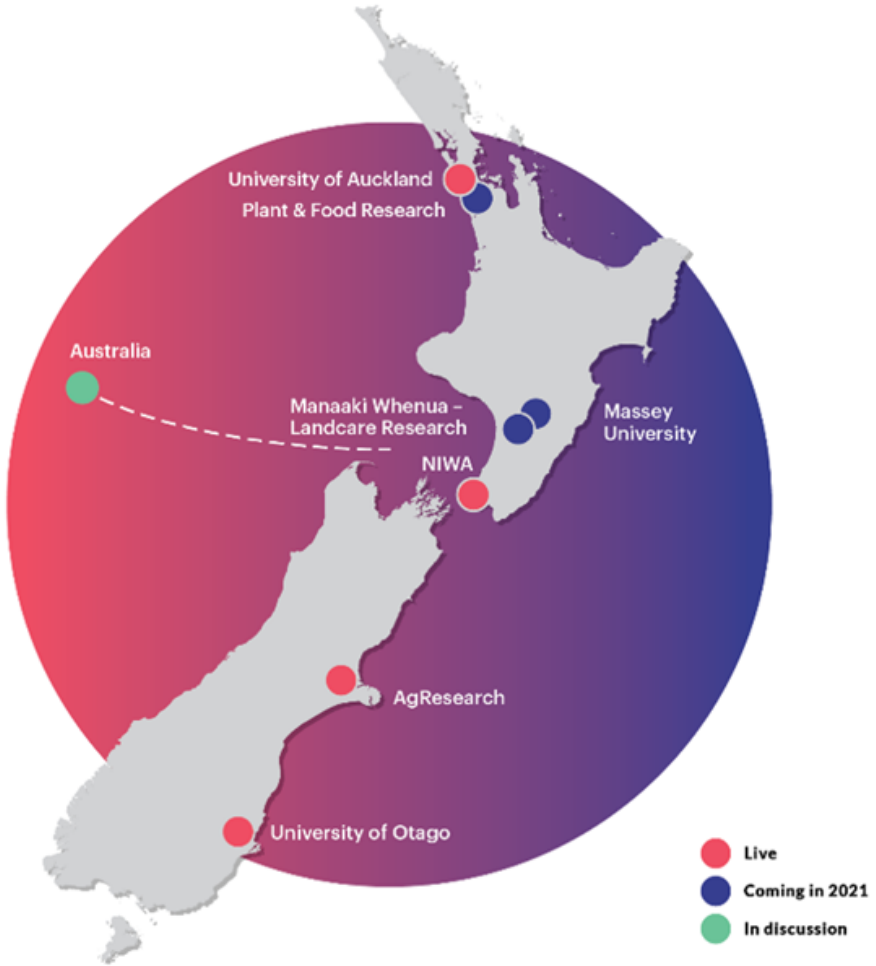
- In-person and online training to grow capabilities in NZ research sector
- Partnership with The Carpentries (global programme to teach foundational coding and data science skills to researchers)

Consultancy

- Computational science experts available to lift the computational capabilities of research teams, as well as optimise tools & workflows

National data transfer platform activities in 2020:

870 TB	5,107	182
138% increase on 2019	28% increase	42% increase
Amount of data transferred	Number of transfers made	Number of users



Globus benefits:

High-speed data transfer

- Move gigabits of data on a network 1,000 times faster than broadband Internet. Powered by REANNZ (NZ's national advanced network provider), research data transfers can be done at 10Gbps.

Secure and easy data sharing

- Share your research data with collaborators locally, nationally, and around the world. Control who has access using group management tools.

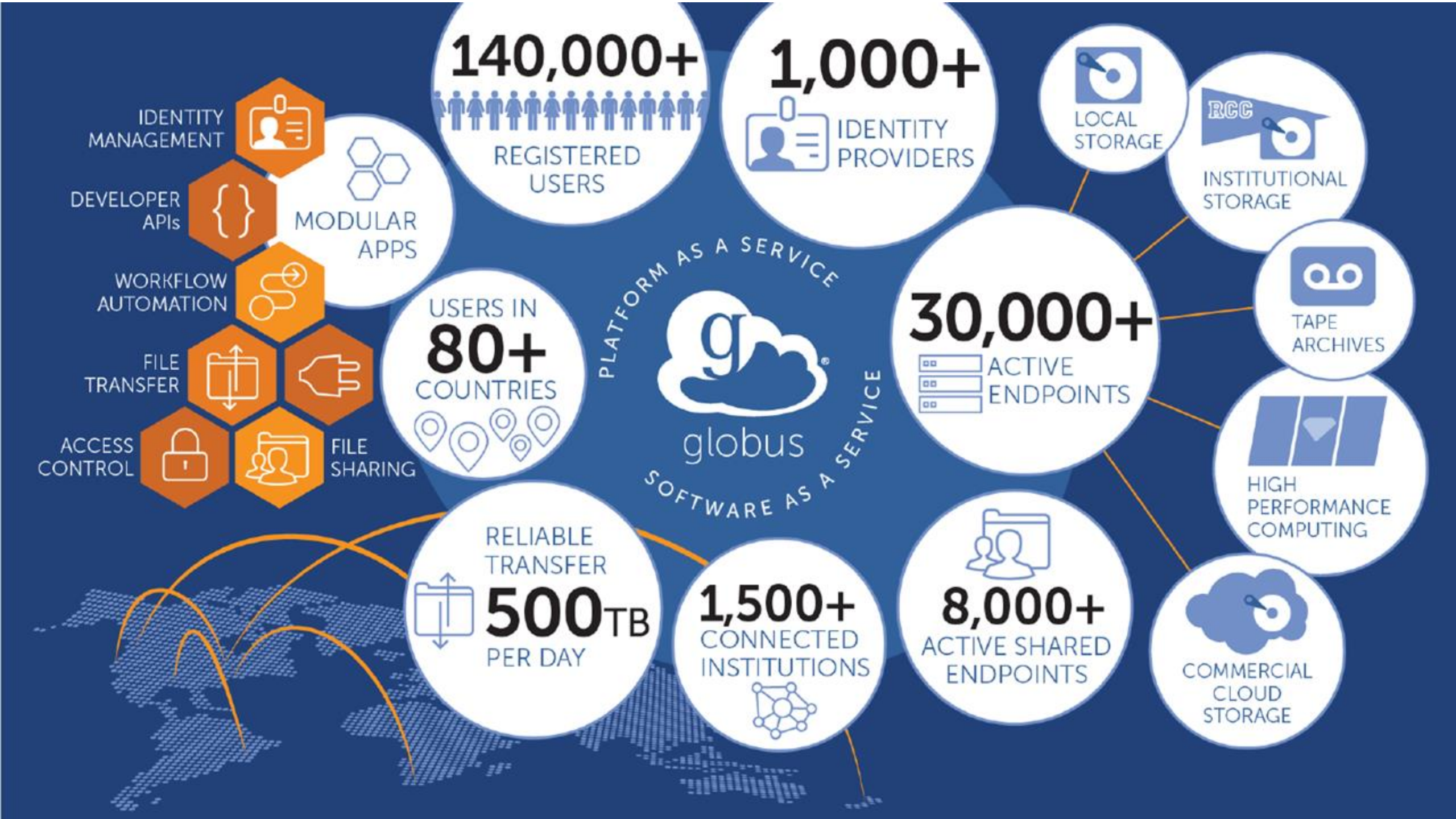
Research data delivery network

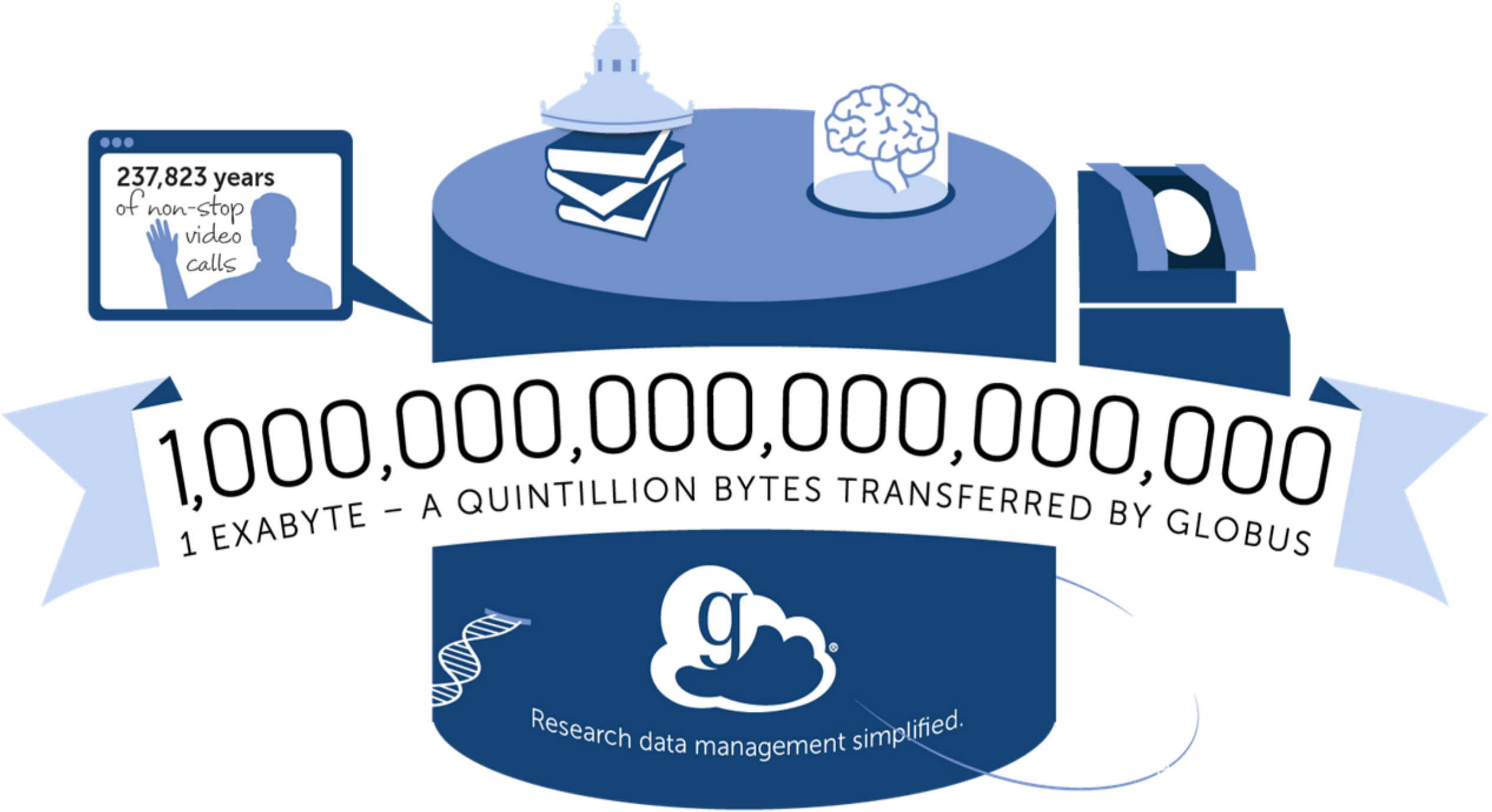
- Take advantage of our partnering institutions across NZ who have existing data delivery network nodes. Transfer to and from any laptop or server using a Globus connect endpoint.

Globus is …

a non-profit service developed and operated by

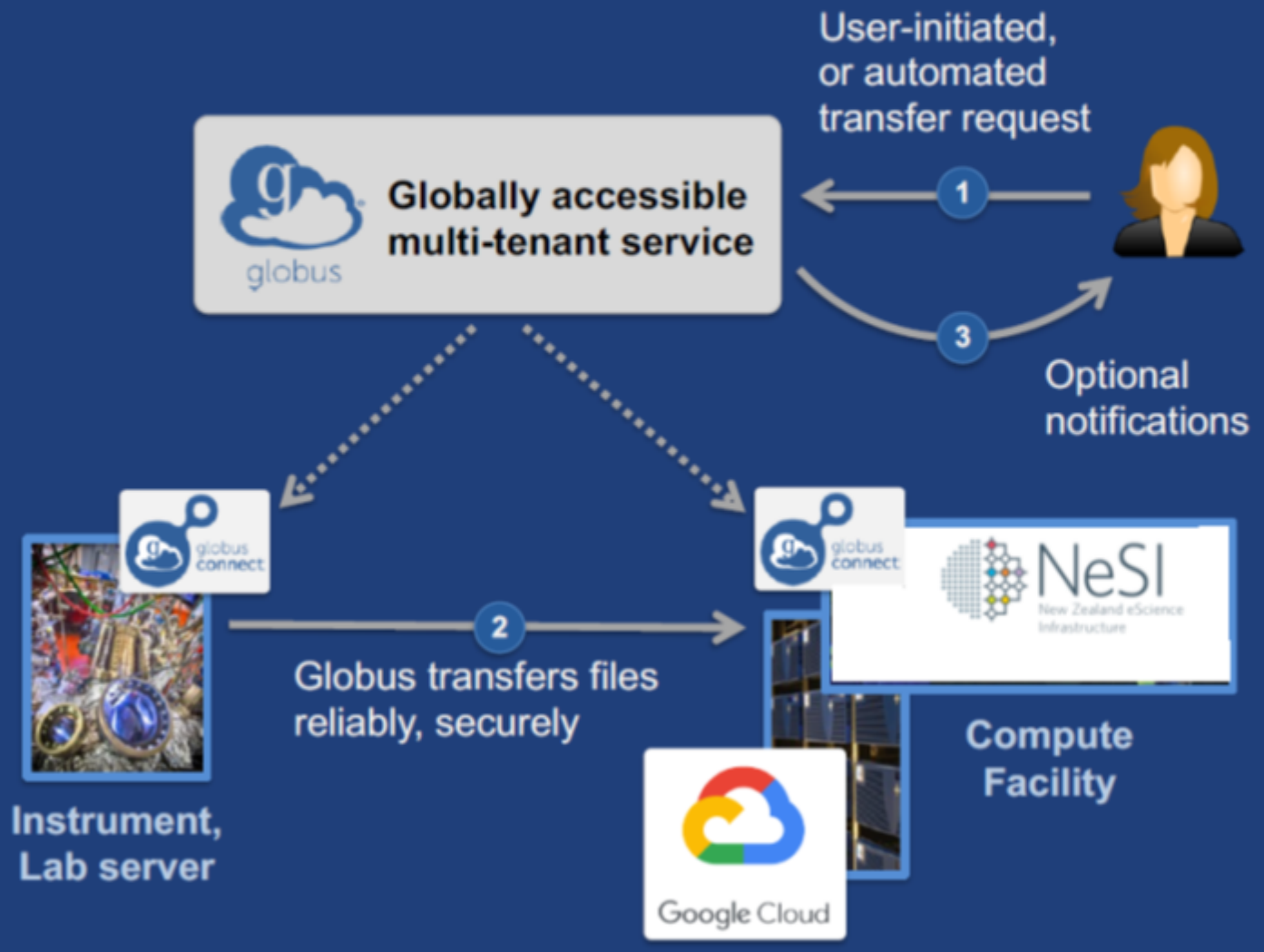
THE UNIVERSITY OF CHICAGO


140,000+
REGISTERED USERS
1,000+
IDENTITY PROVIDERS
LOCAL STORAGE
RCC
INSTITUTIONAL STORAGE
IDENTITY MANAGEMENT
DEVELOPER APIs
MODULAR APPS
WORKFLOW AUTOMATION
FILE TRANSFER
ACCESS CONTROL
FILE SHARING
USERS IN 80+ COUNTRIES
PLATFORM AS A SERVICE
globus
SOFTWARE AS A SERVICE
30,000+
ACTIVE ENDPOINTS
TAPE ARCHIVES
HIGH PERFORMANCE COMPUTING
RELIABLE TRANSFER
500TB
PER DAY
1,500+
CONNECTED INSTITUTIONS
8,000+
ACTIVE SHARED ENDPOINTS
COMMERCIAL CLOUD STORAGE


237,823 years
of non-stop
video
calls
1,000,000,000,000,000,000
1 EXABYTE – A QUINTILLION BYTES TRANSFERRED BY GLOBUS
Research data management simplified.

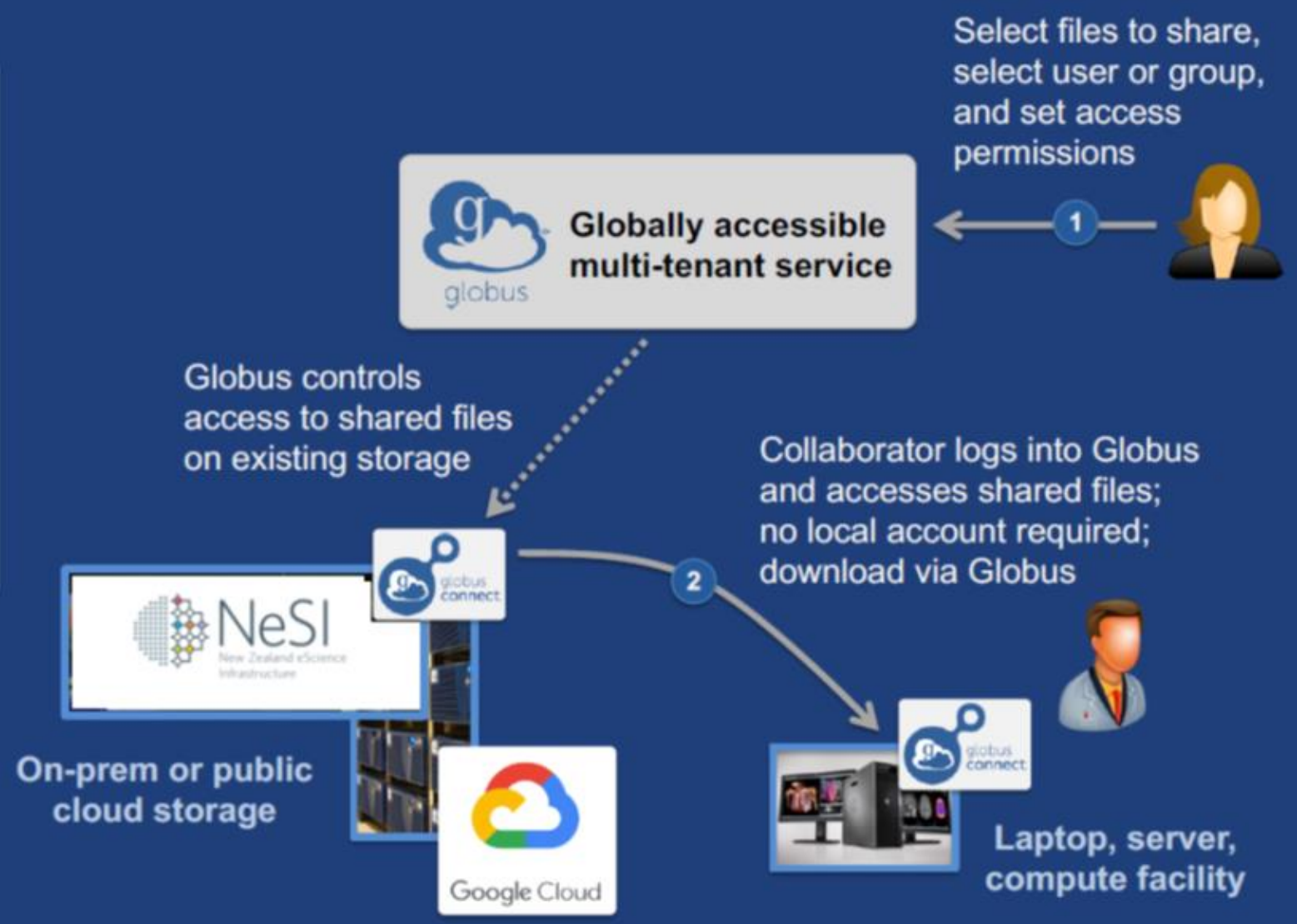
Fast, reliable file transfer …from any to any system

- Fire-and-forget transfers
- Optimized speed
- Assured reliability
- Unified view of storage
- Browser, REST API, CLI



Secure data sharing …from any storage

- **Fine-grained access control "overlay" on storage system**
- **Share with any identity, email, group**
- **No need to stage data just for sharing**



Globus Connectors

amazon web services™ S3

Google Cloud

SPECTRA

box

Caringo Swarm™

Quantum ActiveScale Object Storage

IBM Cloud Object Storage

IBM Spectrum Scale

Google Drive

SCALITY

hadoop HDFS

ceph

HPSS

wasabi hot cloud storage

lustre™

Planned

OneDrive

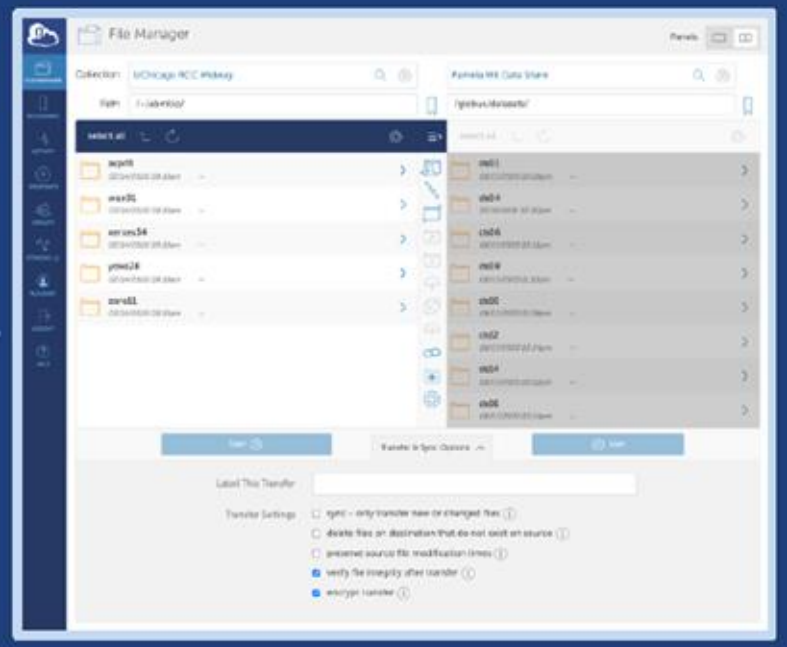
Microsoft Azure Blob Storage

Dropbox

Use(r)-appropriate interfaces

Globus service

Web



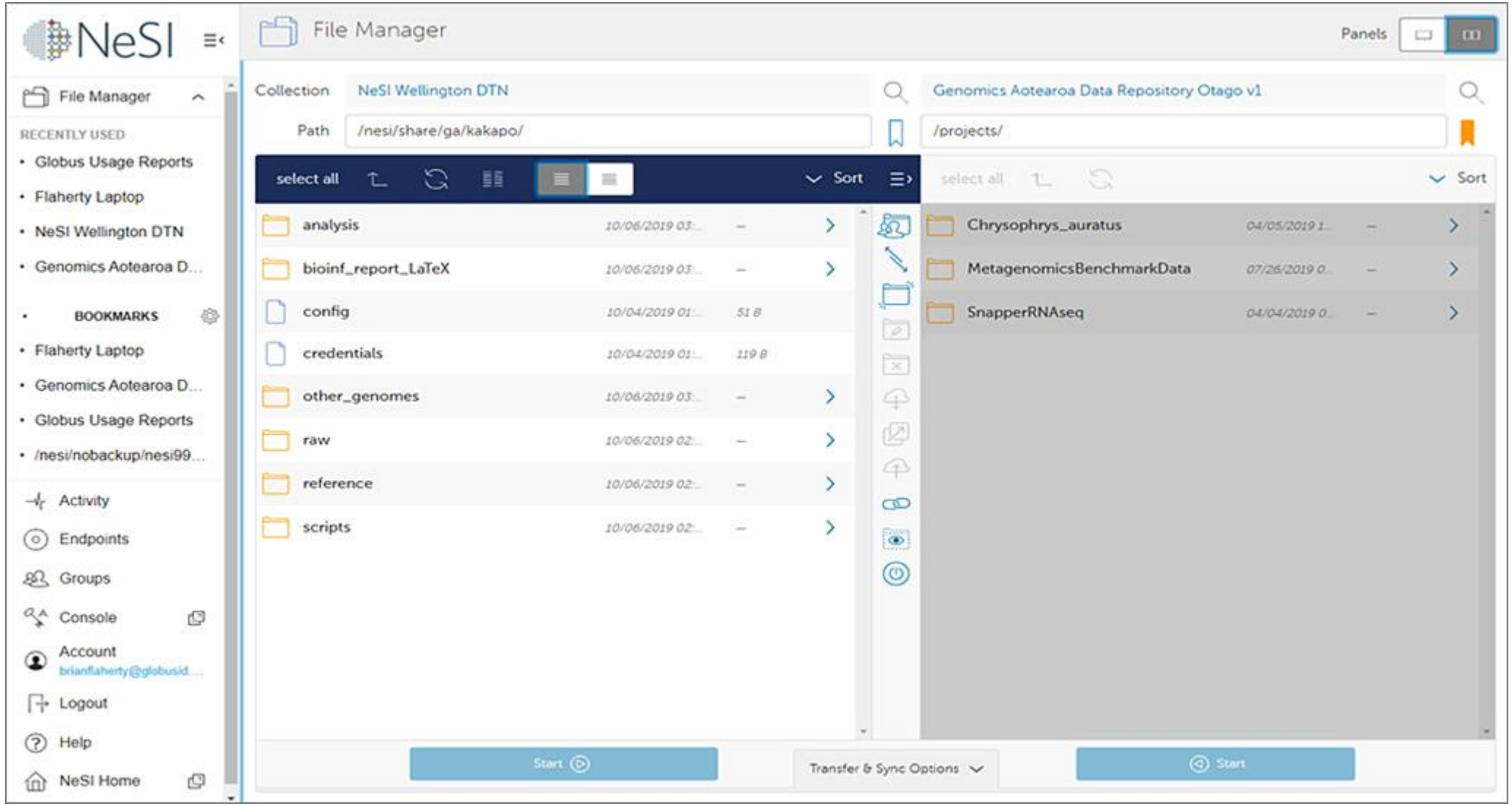
CLI

```
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose                   Control level of output
  -h, --help                      Show this message and exit.
  -F, --format [unix|json|text]   Output format for stdout. Defaults to text
  --jmespath, --jq TEXT           A JMESPath expression to apply to json
                                  output. Takes precedence over any specified '
                                  --format' and forces the format to be json
                                  processed by this expression
  --map-http-status TEXT          Map HTTP statuses to any of these exit codes:
                                  0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark        Manage endpoint bookmarks
  config          Manage your Globus config file. (Advanced Users)
  delete          Submit a delete task (asynchronous)
  endpoint        Manage Globus endpoint definitions
  get-identities  Lookup Globus Auth Identities
  list-commands   List all CLI Commands
  login           Log into Globus to get credentials for the Globus CLI
  logout          Logout of the Globus CLI
  ls              List endpoint directory contents
  mkdir           Make a directory on an endpoint
  rename          Rename a file or directory on an endpoint
  rm              Delete a single path; wait for it to complete
  session         Manage your CLI auth session
  task            Manage asynchronous tasks
  transfer        Submit a transfer task (asynchronous)
  update          Update the Globus CLI to its latest version
  version         Show the version and exit
  whoami          Show the currently logged-in primary identity.
```

Platform (RESTful APIs)

```
GET /endpoint/go%23ep1
PUT /endpoint/demodoc#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
```

transfer.nesi.org.nz



Automation Examples

- **Syncing a directory**
  - bash script; calls the Globus CLI
  - Python module; run as script or import as module
- **Staging data for distribution**
  - bash and Python variants
- **Removing directories after files are transferred**
  - Python script

github.com/globus/automation-examples

Globus Automate

A platform service for defining, applying, and sharing distributed research automation **flows**

- **Triggers** start flows based on subscribed events
- Flows call **Action Providers** to perform tasks

Automation Action Providers

Transfer	Delete	ACLs	funcX	DLHub
Identifier	User Form	Notification	Xtract	Web Form
Ingest	Expression Evaluation	Search	Describe	

Globus action providers

Custom action providers

NATIONAL RESEARCH AND EDUCATION NETWORK

Research Education Advanced Network NZ

NeSI
New Zealand eScience Infrastructure

REANNZ

REANNZ

Research and Education, Advanced Network New Zealand

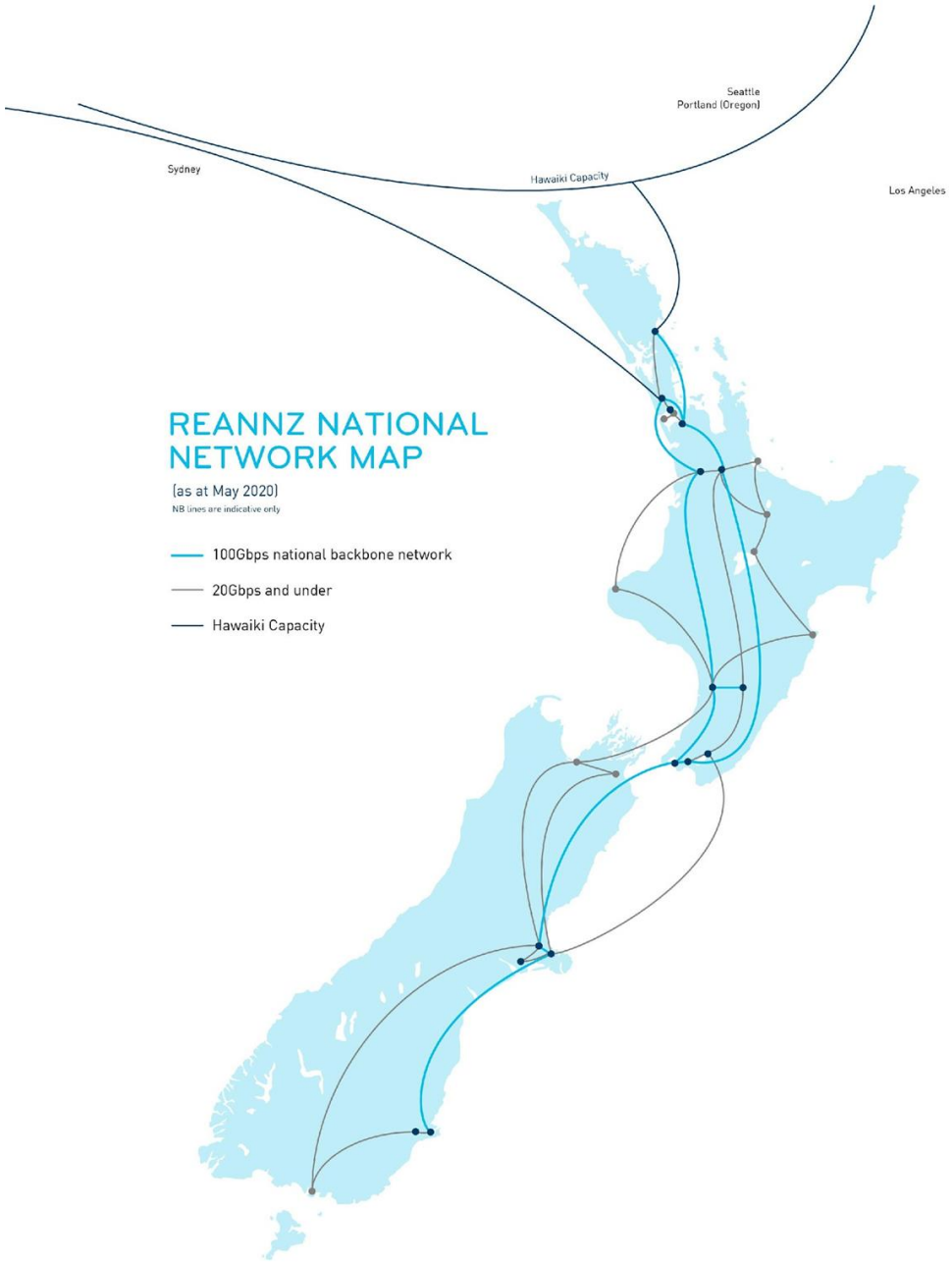
NREN

National, Research and Education Network

120 global NRENs

Supporting and enabling research outcomes

Crown Owned entity, not for profit



Enabling research in NZ

Operating the network and underlying infrastructure

Core services

domestic network (26 POPs)

International network (5 POPs)

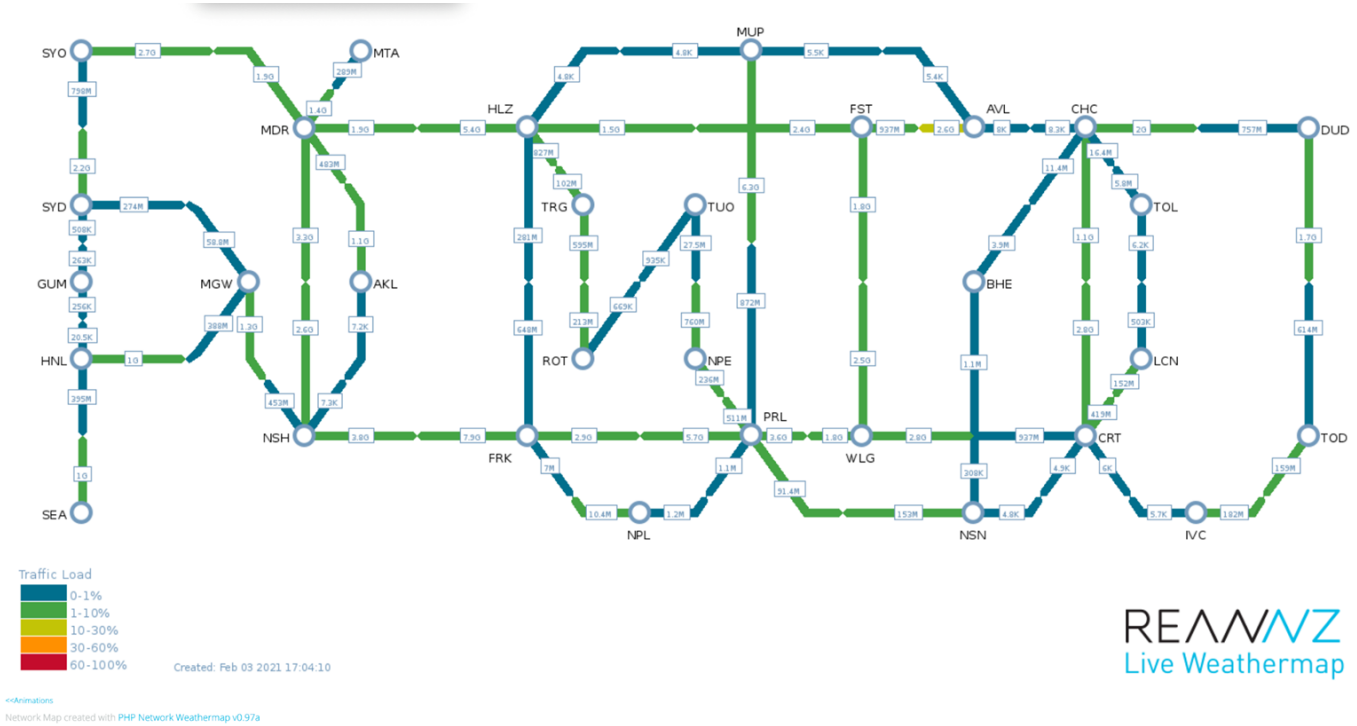
Connectivity solutions

(managed access/edge, manage firewall, science DMZ, cloud connect, eduroam, tuakiri (trust ident), professional services)

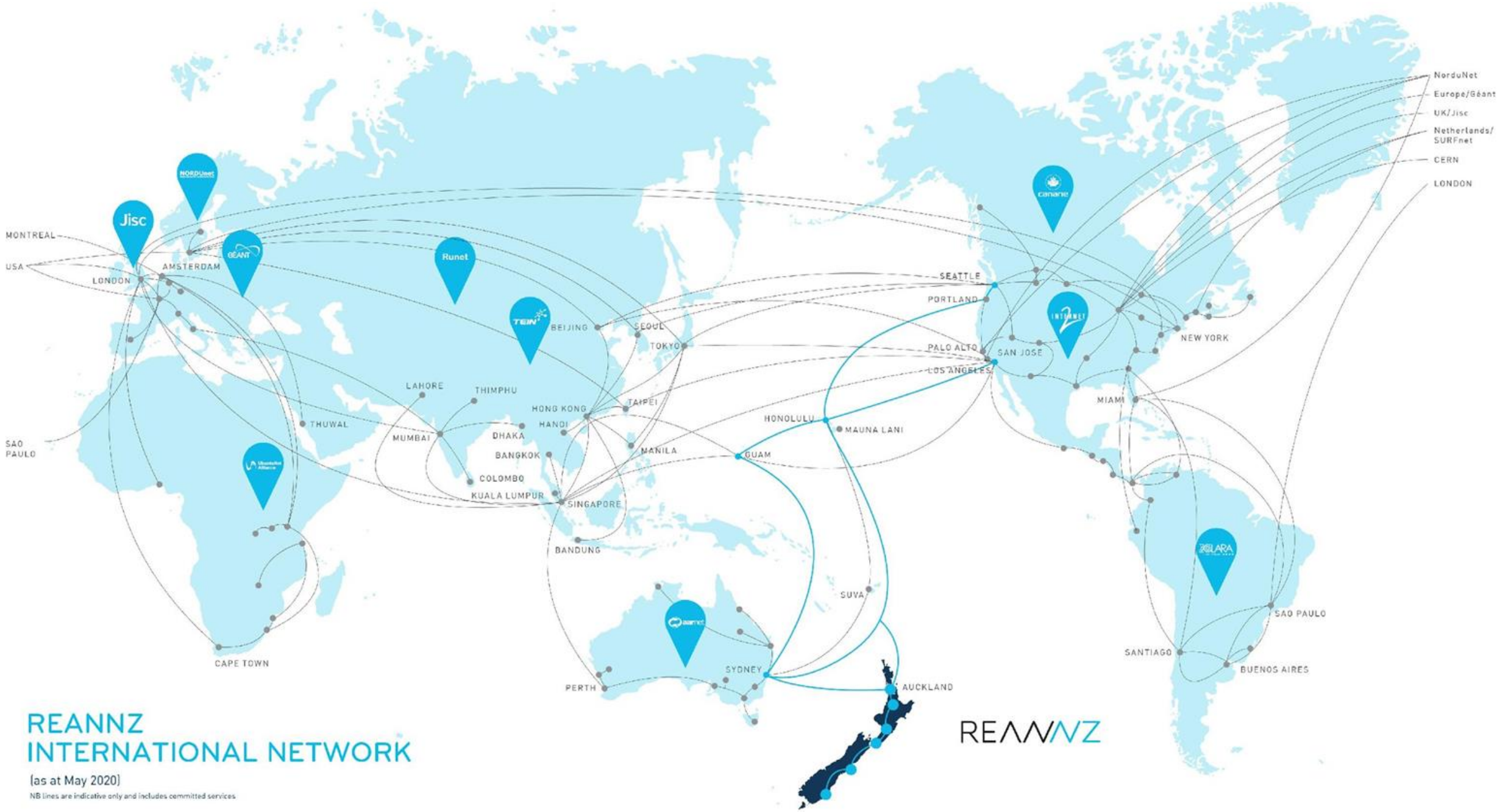
On call 24/7

end to end visibility

Partnership with member technology teams



Global National Research and Education Networks



Global National Research and Education Networks

NRENs operate nationally, but connect globally
Seamless connectivity and tailored services
Sharing information and collaboration, driving research

National R&E networks leverage the global community to support the bespoke and demanding needs of the research and education sector.

By sharing information and developing collaborations of best practice, the global R&E network community creates efficiencies and avoids reinventing the wheel.

In many countries, R&E networks provide connectivity for universities and institutions and organisations with a research and education mission at a cost and capacity that commercial networks cannot.

Network Challenges for a Data Transfer Service

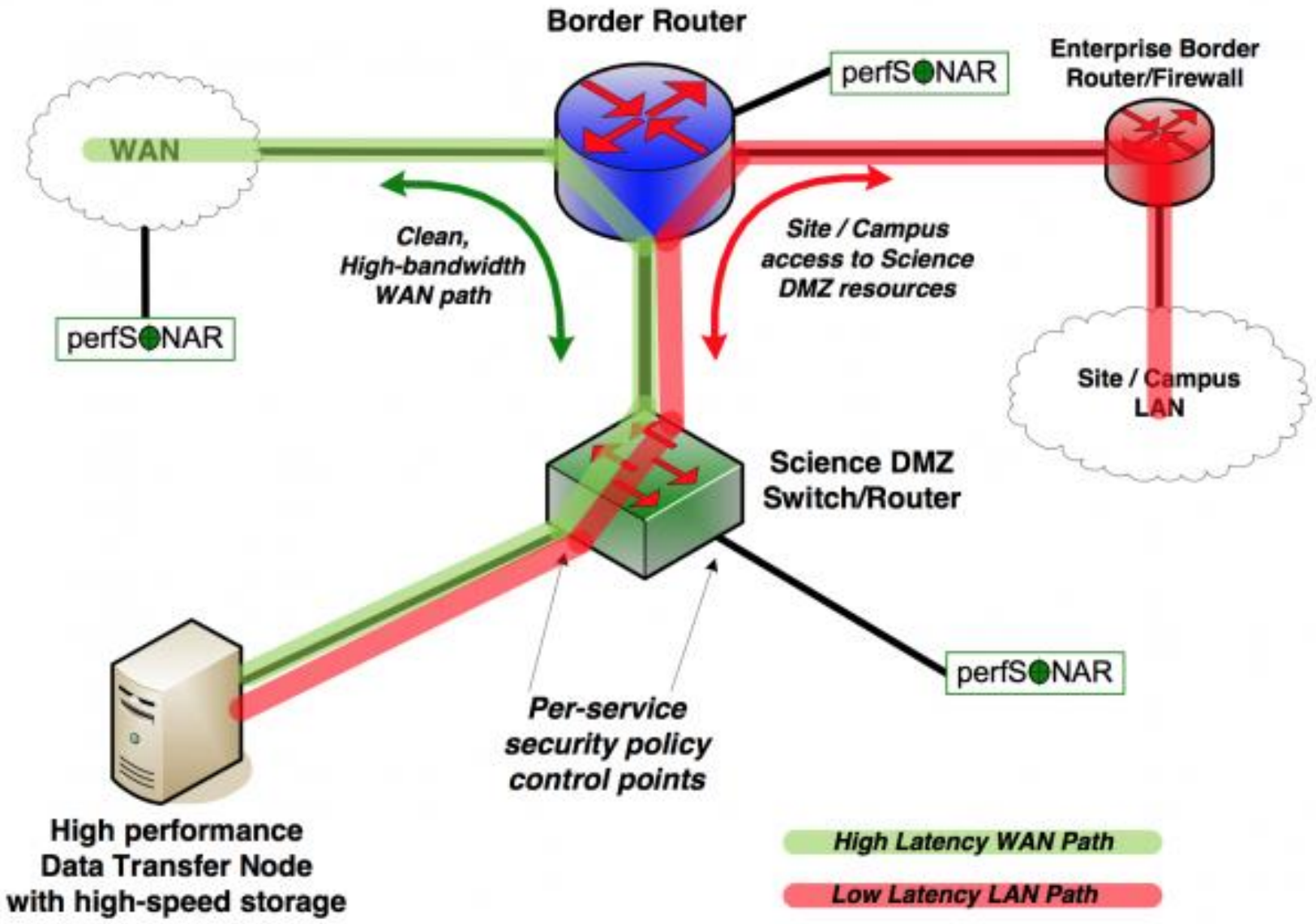
- Think of a network ‘pipe’ – how to get the most data flowing through it per second, every second.
- Make sure your network ‘pipes’ are large – everywhere and at every junction.
- Make sure you have enough data ready to ‘pour’ into the network pipe – so you need fast multiplexed data storage at each end.
- How do you get data into the network pipe fast enough – use multiple high speed data ‘pumps’ - each uses many large ‘data buckets’. They all pump at once.
- Keep the network data pipes below capacity and minimise the data flow controls.

What is a Science DMZ?

- A Science DMZ is a high-performing on-ramp at or near the site network perimeter, dedicated to supporting data-intensive science resources.
- The Science DMZ is a portion of the network, built at or near the campus or laboratory's local network perimeter. It is designed so that the equipment, configuration, and security policies are optimised for high-performance scientific applications, rather than for general-purpose business systems or “enterprise” traffic.
- Can be used as dedicated systems for data transfer (DTNs)
- Integrated performance management and security
- Built with high performance components
- Equipment, configuration and policies are optimized for high performance scientific applications

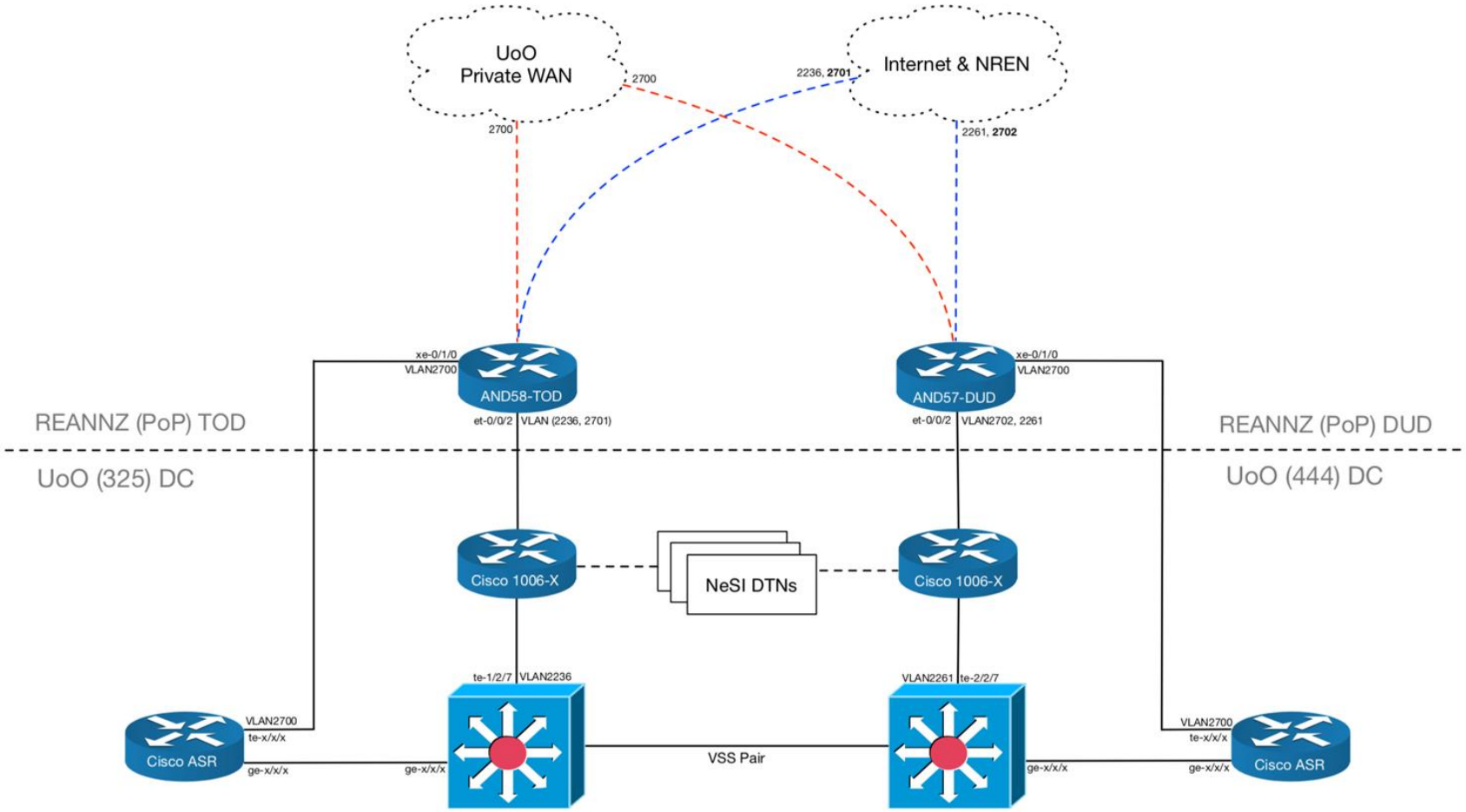
How does it work?

- Addresses common network performance problems encountered at research institution
- Science applications will go through clean high bandwidth path
- Equipment, configuration and policies are optimized for high performance scientific applications
- Managed edge switch, so that everyday internet traffic could be directed down a normal path protected by enterprise firewalls etc. while research data would go down a different, faster path, which bypassed all of the gates it didn't need to go through.
- Creating a clear, high-speed research data path to and from the REANNZ network.
- REANNZ can manage the switch on behalf of the member, so if changes or upgrades need to be made in the future REANNZ can make them, meaning less work for the member.
- Talk about the following example of a Science DMZ topology



Source: https://fasterdata.es.net/science-dmz/science-dmz-architecture/

University of Otago – NeSI Science DMZ



Thank you!

Questions?

Brian Flaherty
Richard Tumaliuan

NeSI REANNZ
New Zealand eScience Infrastructure