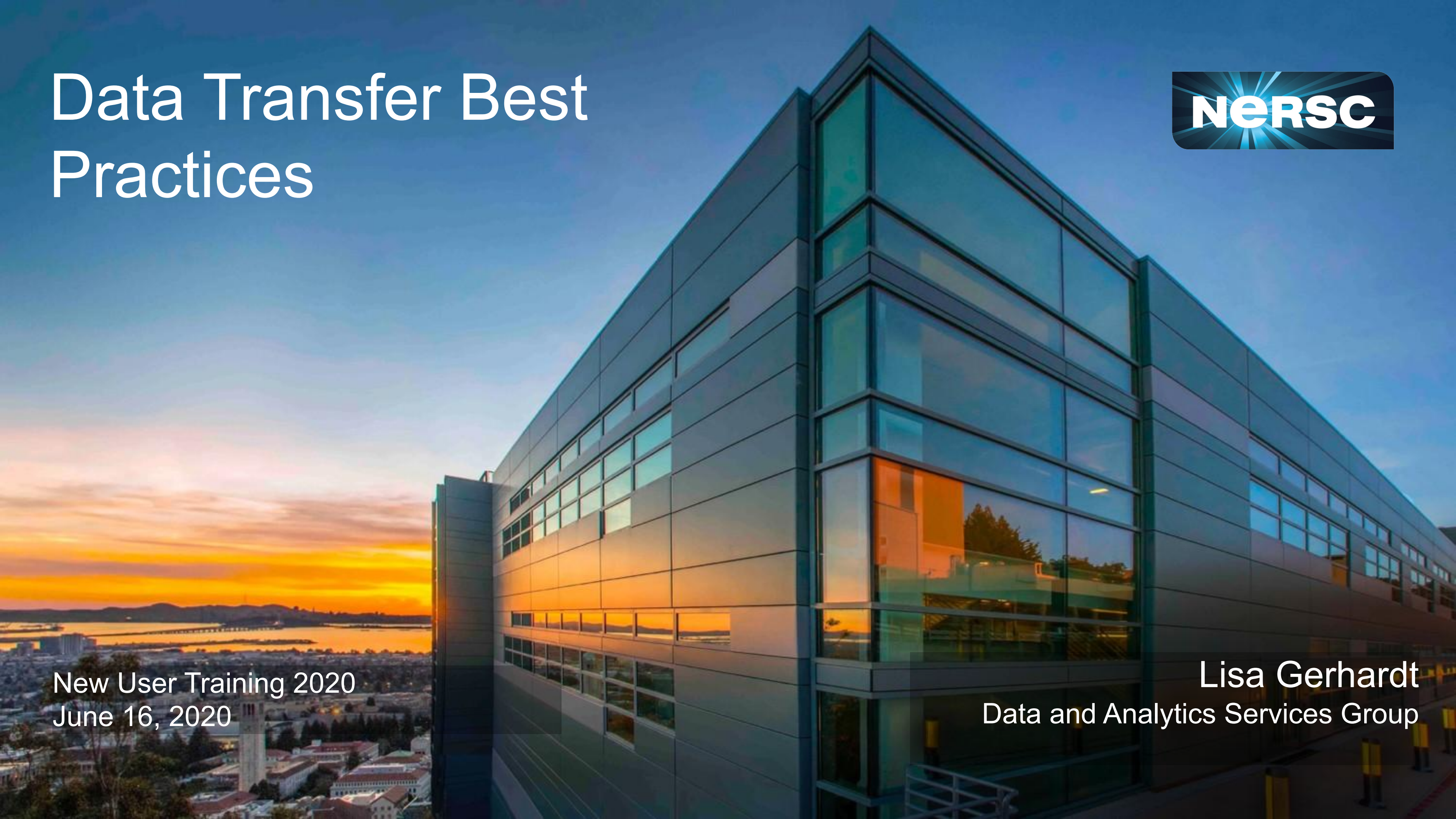

Data Transfer Best Practices

New User Training 2020
June 16, 2020

Lisa Gerhardt
Data and Analytics Services Group

Dedicated Data Transfer System: Data Transfer Nodes

- **Data Transfer Nodes (DTNs) are dedicated servers for moving data at NERSC. (dtnXX.nersc.gov)**
  - Servers include high-bandwidth network interfaces & are tuned for efficient data transfers
    - Monitored bandwidth capacity between NERSC & other major facilities such as ORNL, ANL, BNL, SLAC…
  - Direct access to global NERSC file systems & Cori cscratch1
  - Can be used to move data internally between NERSC systems and / or NERSC HPSS
  - **Use NERSC DTNs to move large volumes of data in and out of NERSC or between NERSC systems**

Globus

- **The recommended tool for moving data in & out of NERSC**
  - Reliable & easy-to-use web-based service:
    - Automatic retries
    - Email notification of success or failure
  - Accessible to all NERSC users
  - NERSC managed endpoints on DTNs for optimized data transfers
  - Web based GUI for drag and drop transfers
  - NERSC Globus scripts for command line transfers
  - REST/API for scripted interactions with service
  - Globus Connect Personal for setting up endpoints on your laptop

https://docs.nersc.gov/services/globus/

Globus Demo

General Tips for Transferring Data

- **Use Globus Online for large transfers**
  - Can use them for internal NERSC transfers e.g. between CFS and scratch
- **scp is fine for smaller, one-time transfers (<100MB)**
  - But note that Globus is also fine for small transfers
- **Don’t use DTN nodes for non-data transfer purposes**
  - Use system login nodes for more general routine tasks

Performance Considerations

- **Performance is often limited by the remote endpoint**
  - Not tuned for WAN transfers or have limited network link
  - These can lower performance < 100 MB/sec.
- **File system contention may be an issue**
  - Try the transfer at a different time or on a different FS.
- **Don't use your $HOME directory**
  - Instead use CFS, $SCRATCH …
- **If you think you are not getting the performance you expect, let us know: help.nersc.gov**

Transferring with NERSC HPSS

- **HPSS tape archive is recommended for archiving large amounts of data for long periods of time**
  - See: https://docs.nersc.gov/filesystems/archive_access/
- **Use interactive DTNs or xfer queue to transfer to / from HPSS**
  - HSI for individual files and conditional access
  - HTAR for aggregation & optimization of storage/archival of large numbers of files. Aim for bundle sizes of 200GB - 2 TB
- **User NERSC Globus Command line tools for external Globus transfers**
  - Will automatically sort files in tape order
  - However Globus does not directly support aggregation with ‘htar’ or tape-ordering
  - Preferred use is for small number of large files

https://docs.nersc.gov/services/globus/#command-line-globus-transfers-at-nersc

NERSC Globus Command Line Demo

Sharing with External Collaborators

- **Public html Access**
  - Project specific area can be created:
    - /global/cfs/cdirs/<yourproject>/www
  - These are available for public access under the URL:
    - https://portal.nersc.gov/project/<yourproject>
- **Science Gateways**
  - Web portals that allow you to interface with your data and computation at NERSC
  - Build sophisticated web applications in Spin
- **Globus Sharing**
  - Projects can set up read-only endpoints for sharing specific data with a subset of Globus users
  - Excellent way to share large volumes of data, can be incorporated into
- **Links:**
  - http://www.nersc.gov/users/data-analytics/science-gateways/
  - https://docs.nersc.gov/services/spin/getting_started/
  - https://docs.nersc.gov/services/globus/#globus-sharing

Thank You and Welcome to NERSC!