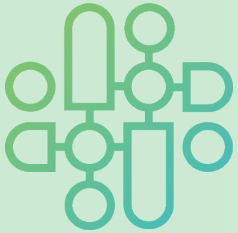


SC23
Denver, CO | i am hpc.

Democratizing Remote HPC Storage Access at Penn State

Adam Focht

Institute for Computational and Data Sciences

Penn State University

ICDS Resources

- Standard Environment
  - 25,000 cores, 100 GPUs
  - 16PB Shared VAST NFS Filesystems
- Restricted Environment
  - 24,000 cores, TBD GPUs
  - 8PB Shared GPFS Filesystems

Remote HPC Storage Access

- SSH-based
  - SCP, Filezilla, WinSCP, etc.
  - SSHFS
  - Use existing SSH services
  - Not native on Windows
- NFS
  - Firewall/security headache
  - Not native on Windows
- CIFS
  - Windows native
  - Security concerns/firewall headache

Simple HPC Storage Access

- "There's a better way" – Amit Amritkar
- Cross-platform
- Familiar Behavior
  - Similar to OneDrive, Dropbox
  - Minimal user interaction
  - File sharing interface
- Role-based Access Control
- Integration with Existing Infrastructure
  - Account data (SSO)
  - Remote shared filesystems

Unifying Storage Access

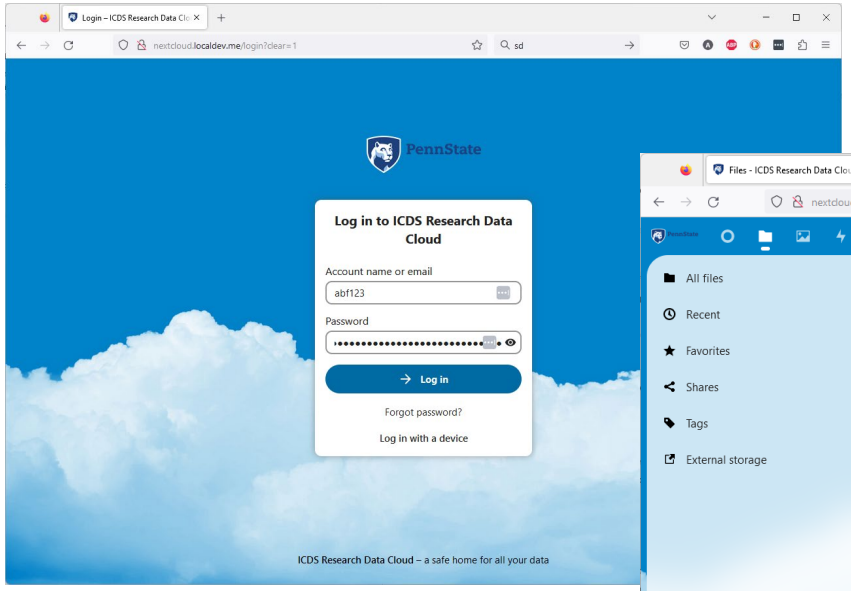
- HPC Storage
  - NFS, SMB, or SSH-based access
  - Multiple targets/systems
- University Shared Storage
  - Box – contract ended
  - OneDrive – contract ending
  - SMB mounted local storage

Storage Cloud Implementation

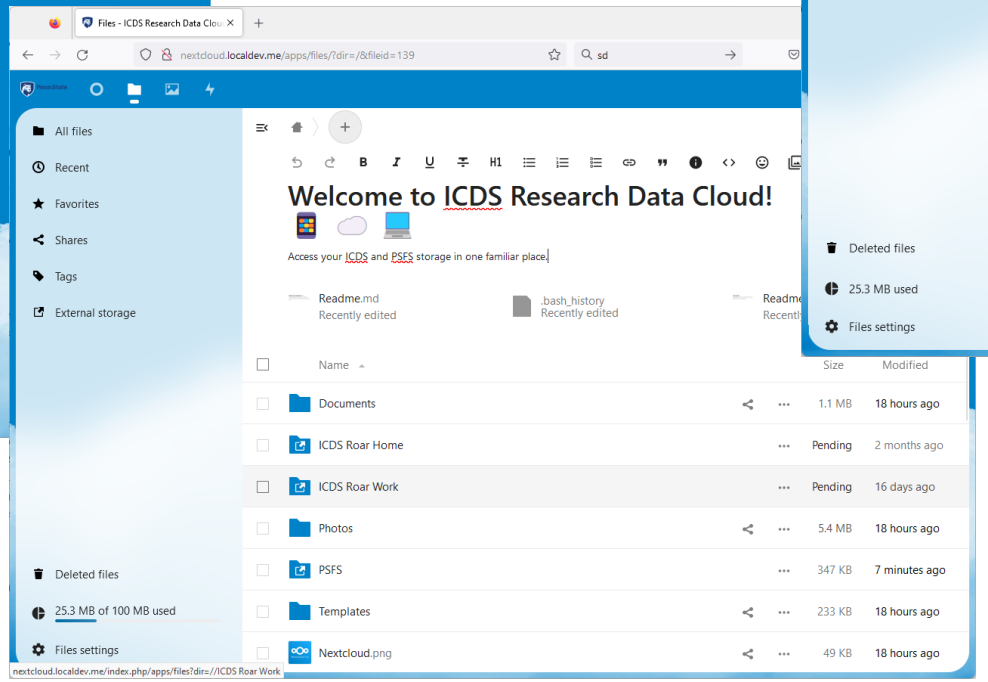
- We used Nextcloud Files
  - Meets requirements above
  - Integrates well with many providers
  - Open source, with paid option
  - Focus on security
  - Running on Kubernetes
  - Deployed via Helm
- Integrated with Penn State central AD services
- Available outside of Penn State (no VPN required)

Web Client

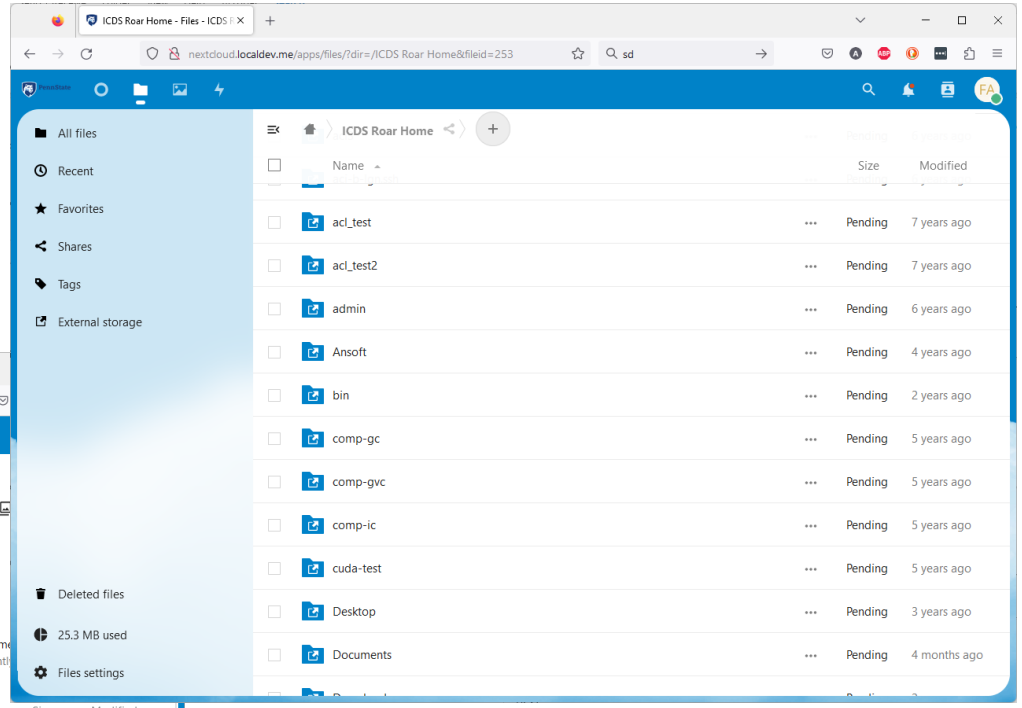
Login with University Credentials



Nextcloud Base Directory

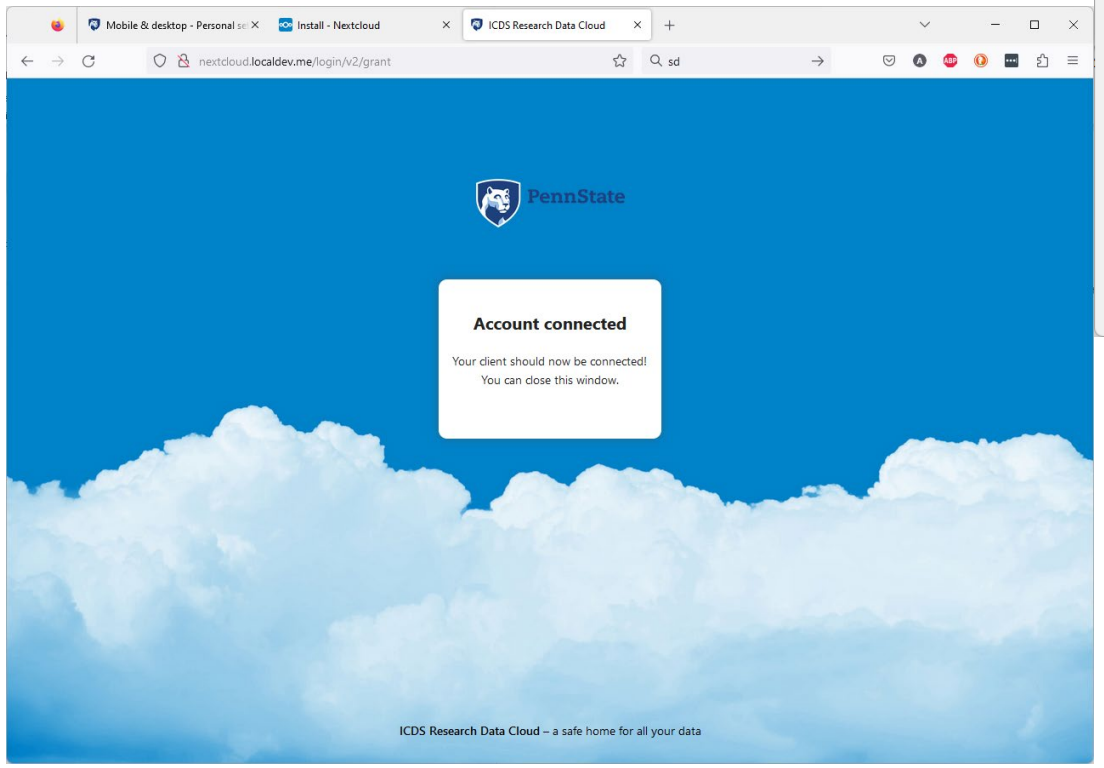


Remote Cluster Home

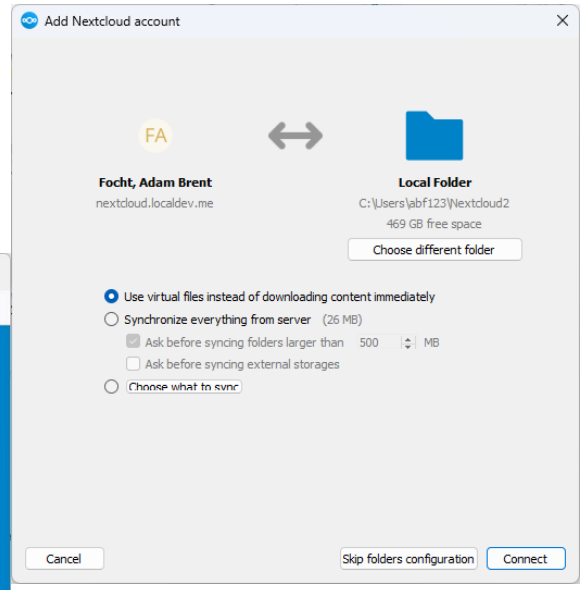


Desktop Client

Authentication via Web Browser

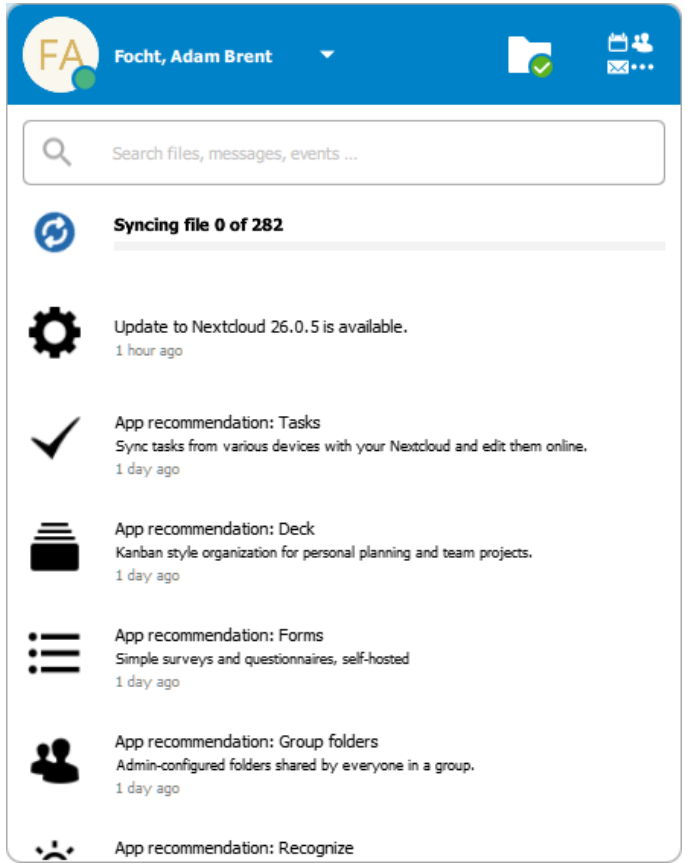


Initial Client Setup



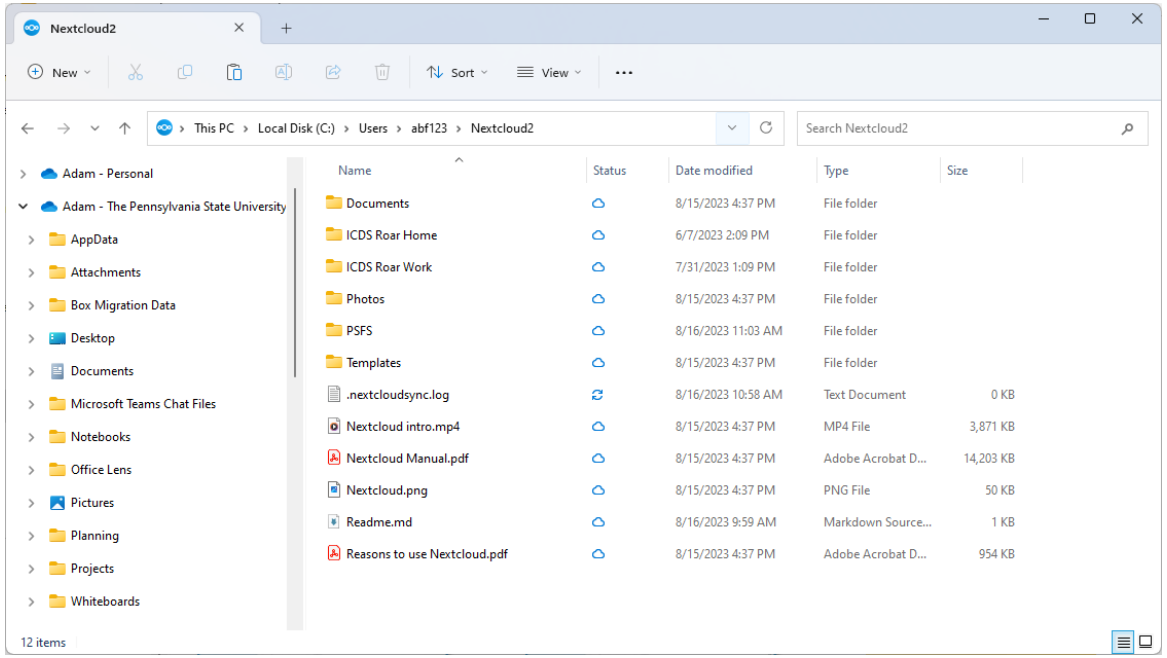
Taskbar Status

Client Status Pop-up

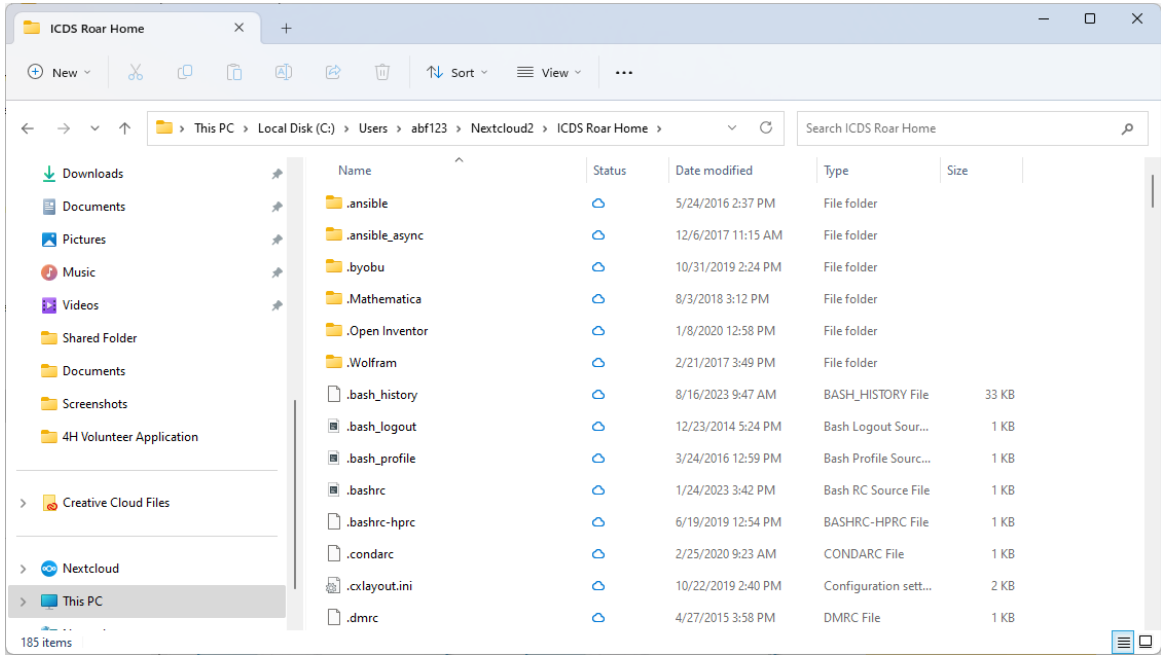


Desktop File Explorer Integration

Local Nextcloud Directory



Local Cluster Home Synced Directory



Storage Cloud Lessons Learned

- Large (LDAP/AD) Directories Significantly Affect Performance
- Backing Database Impact on Performance
- Multi-Factor Auth Can Be Incredibly Annoying
  - Opens multiple SSH sessions for multi-threaded interface
  - MFA prompt per session if no MFA caching
  - Internal SSH w/o MFA using Web SSO w/ MFA
- Kubernetes Persistent Storage Multi-access Challenges
- Helm Ease-of-Deployment

Need for Further Testing

- Large File Transfers
  - Timeouts
  - Performance
- General Performance of SSH-based Filesystems
- File Locking Across Access Vectors
- Redis Caching Effect on Performance
- Horizontal Pod Auto-Scaling (Kubernetes)

Enabling HIPAA-Aligned Workflow

- … or any managed restricted data store …
- Data Manager Role
  - Transition data through lifecycle
  - Verify access needs
- Restrict Access via (Auto-)Tagging
  - Enable download, view, etc. via well-formed names or attributes
- … or work with SELinux MLS? …