Earth Lab Workflow Processes

Shelley Knuth, Timothy Dunn, Maxwell Joseph
Earth Lab, Research Computing, University of Colorado-Boulder

shelley.knuth@colorado.edu

Survey!

- http://tinyurl.com/curc16-presurvey

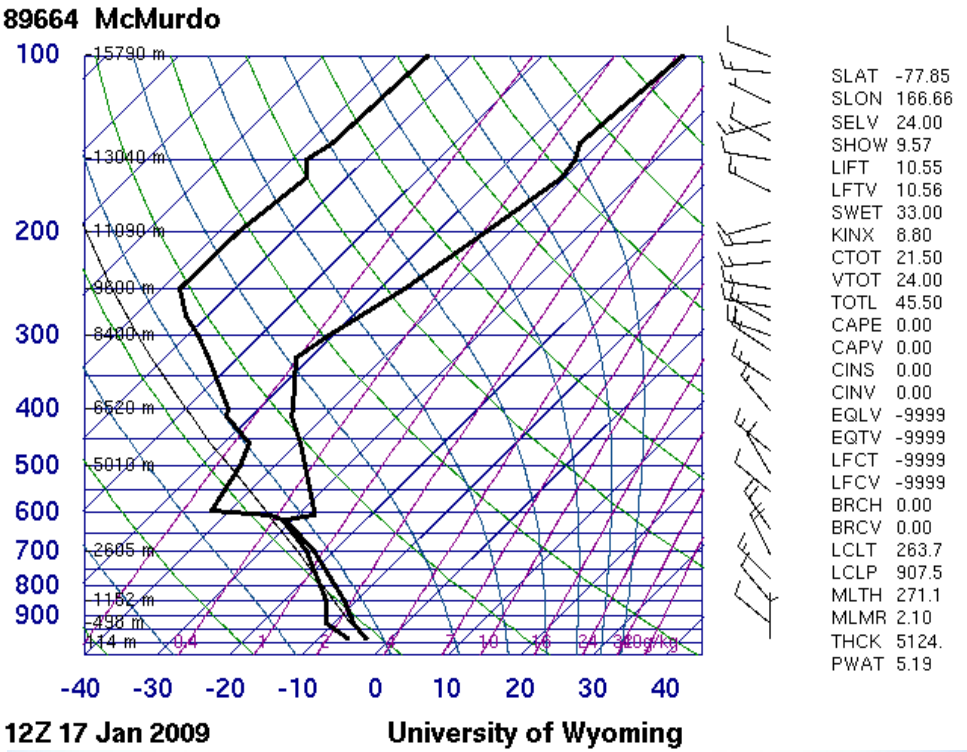
Outline

- What resources does Earth Lab have access to?
  - Infrastructure
  - Technical Team
- How does the Earth Lab team access these resources?
  - General operations and workflow

Resource Description

Overarching Issue

- Research scientists solve science problems
- Must work on technical problems
  - Data formats, transfer, integration
  - Programming
  - Data mining, data visualization
- Lack of training



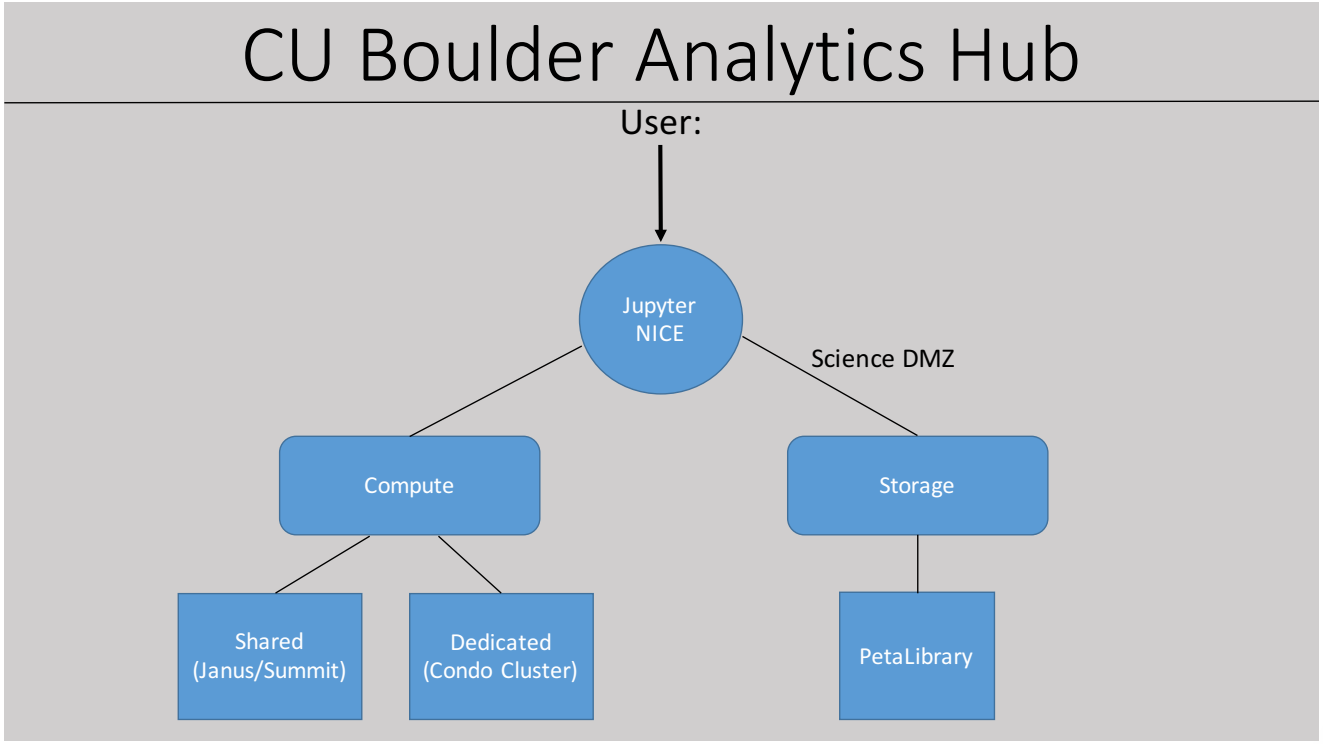
The Analytics Hub

- One mission of Earth Lab is to provide a technical platform to allow scientists better perform their research
  - The Analytics Hub
- Existing infrastructure with new tools and expertise to make this happen
- Three key components:
  - People
  - Data visualization
  - Data analytics

What Is the Analytics Hub?

- A platform to perform science easier
  - Access to large scale computing and storage
    - User-friendly interfaces
    - Don't need to understand details
    - Code recipes
  - Built by Earth Lab for Earth Lab
  - Testing phase
  - http://www.colorado.edu/earthlab/analytics-hub

Architecture of Analytics Hub



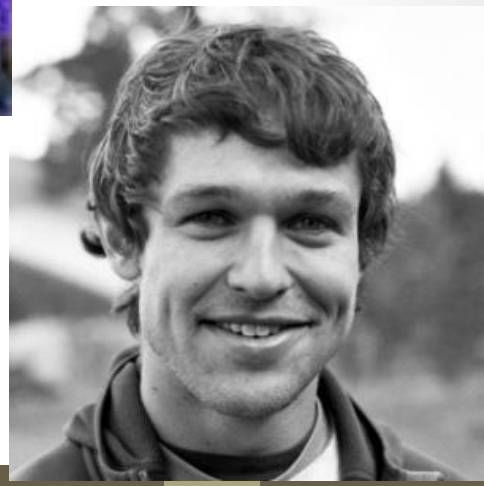
Technical Team

- Shelley Knuth – Director
- Tim Dunn – Visualization
- Max Joseph – Analytics
- Zach Schira – Intern
- Matt Oakley - Intern

Shelley Knuth

Tim Dunn

Max Joseph



Understanding the Cyberinfrastructure

- AH rests on existing infrastructure
- Managed by Research Computing
  - PetaLibrary
  - Janus supercomputer

PetaLibrary

- 960 TB data storage infrastructure
- Large scale data
- Space on Active (spinning disk) or Archive (tape)
- Mounted on Janus

Janus, etc.

- Janus
  - 1368 compute nodes (Dell C6100)
  - 16,428 total cores
  - Infiniband network
- High memory nodes
- Graphics Processing Units
- Large scale data transfer
- ScienceDMZ
  - 80 Gb/s network
  - Not competing with lower level network demands like YouTube, etc.

Summit: CU-Boulder's Next Supercomputer

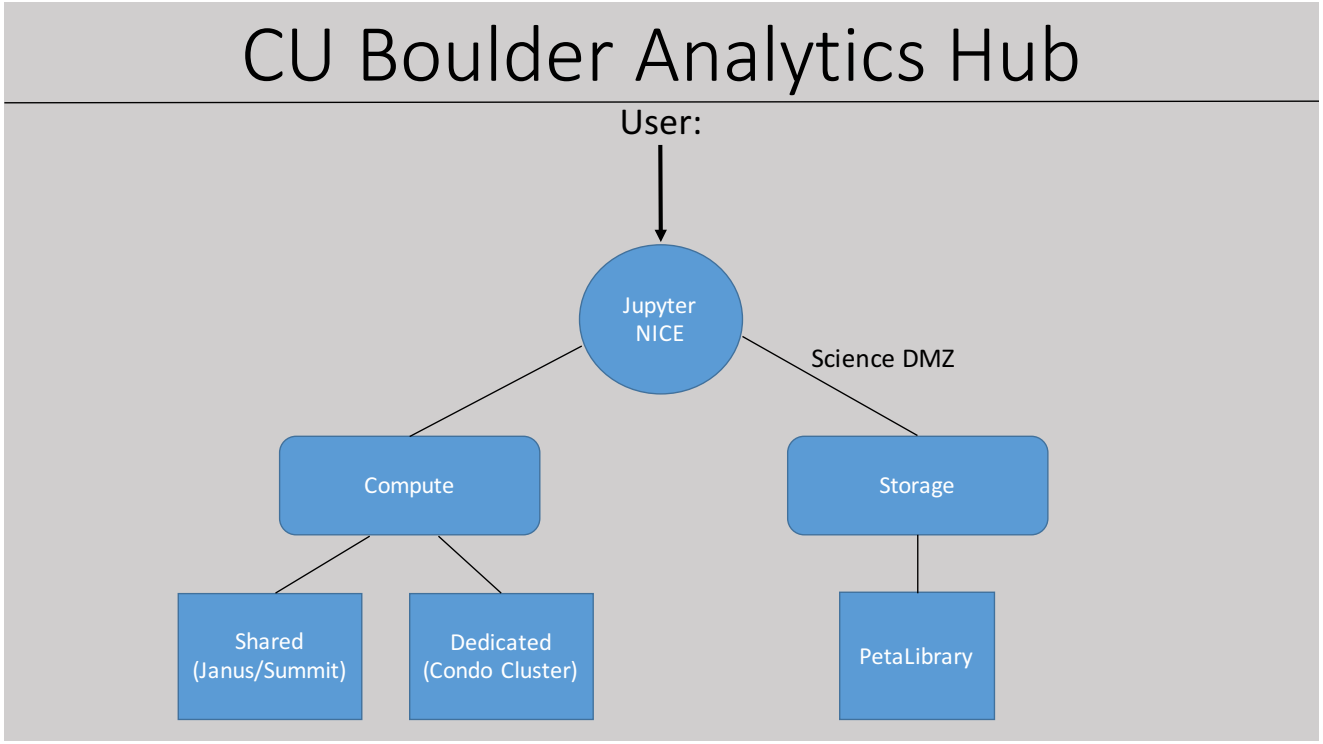
- Joint between CU-Boulder and CSU
- Installed and running by fall semester 2016
- Any CU-Boulder researcher
- Peak performance ~450 TFLOPS
  - JANUS – 170 TFLOPS
- 5-year lifetime
- Dell principal vendor

Credit: Tim Dunn

Dedicated Compute - Condo Clusters

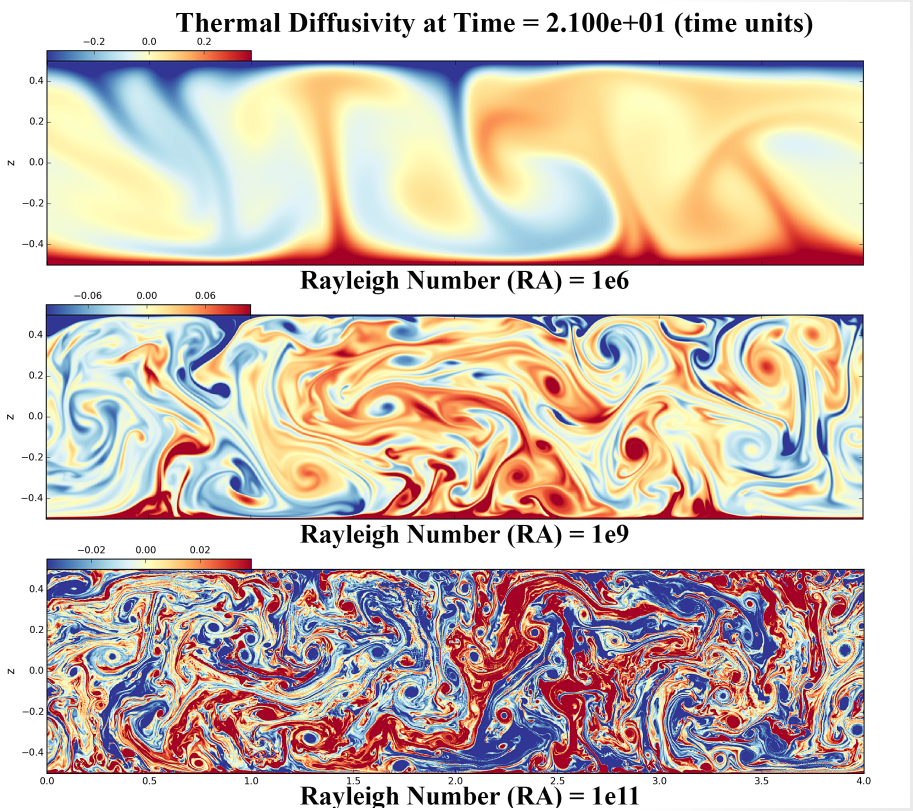
- Groups can purchase compute nodes as part of a cluster called “Blanca”
- Priority for owner
- All other nodes available
- Maximum 4 hour wait
- Earth Lab purchased two nodes

Architecture of Analytics Hub



Data Visualization - NICE

- Remote visualization platform
  - Software on GPUs that are part of the supercomputer
- Platform allows:
  - Easy login - website
  - Produce visualizations near data
  - Don’t have to know how to submit job



Credit: Tim Dunn

Jupyter

- Jupyter notebooks are web applications that allow you to create and share documents that contain live code, equations, visualizations and explanatory text
- Can support 40 programming languages including
  - Python, R, Julia, Scala
  - Jupyter notebooks are really great for teaching
- Earth Lab is utilizing notebooks for our code recipes

Jupyter Hub

- Another available resource
- Used at
  - Berkeley, Software carpentry, SDSC, Pacific Research Platform
  - Recently implemented at Research Computing
- Run a Jupyter notebook easily and seamlessly off the supercomputer
  - No port forwarding
- Can run code recipes as part of the Hub

Earth Lab Participants

- Can access the compute and data resources
  - Directly through the command line
  - Via interfaces
    - Jupyter Hub
    - NICE
- Store data on the PetaLibrary
- Do heavy computational work using this data on Janus/Summit
- Use the Jupyter Hub to run canned code recipes
- Use NICE to visualize the data
- Share data with other researchers using Globus
- Receive technical assistance from staff

Accessing Resources

Earth Lab PetaLibrary

- Purchased 5 TB of space
- To access, you need a Research Computing account

- https://github.com/earthlab/tutorials/blob/JANUS_documentation/documentation/Getting_Started_with_JANUS.md

PetaLibrary Access

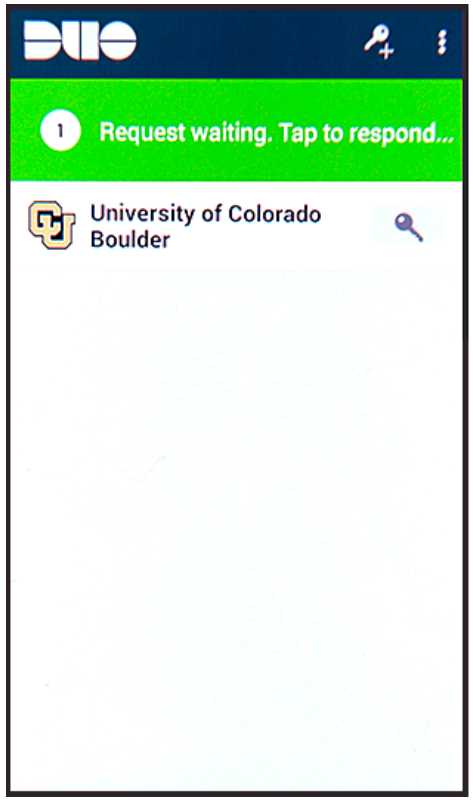
- To access Earth Lab's PetaLibrary space, please do the following:

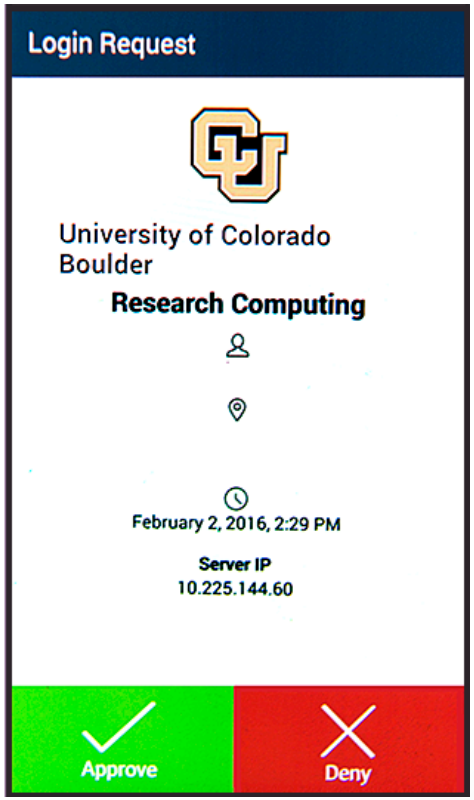
```
ssh login.rc.colorado.edu –l <username>
Password:  duo:<identikey>
cd /work/earthlab/EarthLab_shared
```

- Need the Duo application
- From there you can cd into your project directory

Duo Authentication

- After hitting enter your phone will alert you to an incoming Duo authentication.
- Open your Duo app and click on ‘Request waiting. Tap to Respond…’.
- Next click on ‘Approve’.
- You will receive a message informing you authentication was approved and the EnginFrame VDI page will appear (see next slide).
- **Note**: App graphics will vary by phone type.





PetaLibrary Guidelines

- Each project should store their data within their project folders
- There is no specific limit for data uploads per project
  - However, we only have 5 TB to share as a group
  - If you need to store data that is greater than that or a significant portion of that (> 1 TB) please talk to Shelley first
- Will discuss proper data organization in our Data Standards training in late July

Moving Data to the PetaLibrary

- Globus is our preferred method of data transfer
- Designed with researchers in mind
- End points between computers make for easy data transfer with an easy to use interface
  - Endpoints are different locations that data can be moved to/from
  - Personal or multi-user
- Scripting in use also if don't want to use GUI

www.globus.org

Globus

- Preserves the integrity of data
  - Compares checksums
  - Resumes data transfer if interrupted
- Fast transfer of large data sets
- Globus can be set up to easily share data among collaborators
- Log in using Identikey:
  - https://docs.globus.org/how-to/get-started/
- Set up an endpoint: https://docs.globus.org/how-to/globus-connect-personal-mac/

Condo Cluster

- To access:

```
ssh login.rc.colorado.edu –l <username>
Password:  duo:<identikey password>
```

```
ml slurm/blanca
```

- To submit to Earth Lab nodes:

```
sbatch –qos=blanca-el
```

- To submit to all of blanca:

```
sbatch –qos=blanca
```

Analytics Hub Code Recipes

Built based on common needs of Earth Lab participants

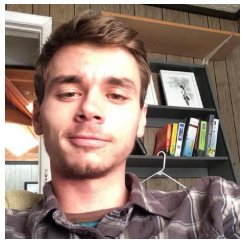
Jupyter notebooks:

https://github.com/earthlab/tutorials

Website:

earthlab.github.io

Zach Schira



Matt Oakley

Remote Visualization

Using the NICE EnginFrame Remote Visualization Platform

Practice

- Log into our tutorial nodes

`tutorial-login.rc.colorado.edu`

- Grab the data 8947.txt in /lustre/janus_scratch/knuths and move it into your directory
- Using the meetup reservation, use Matlab to run the script matlab_test_plot.m

- Can we do better?

Going Native

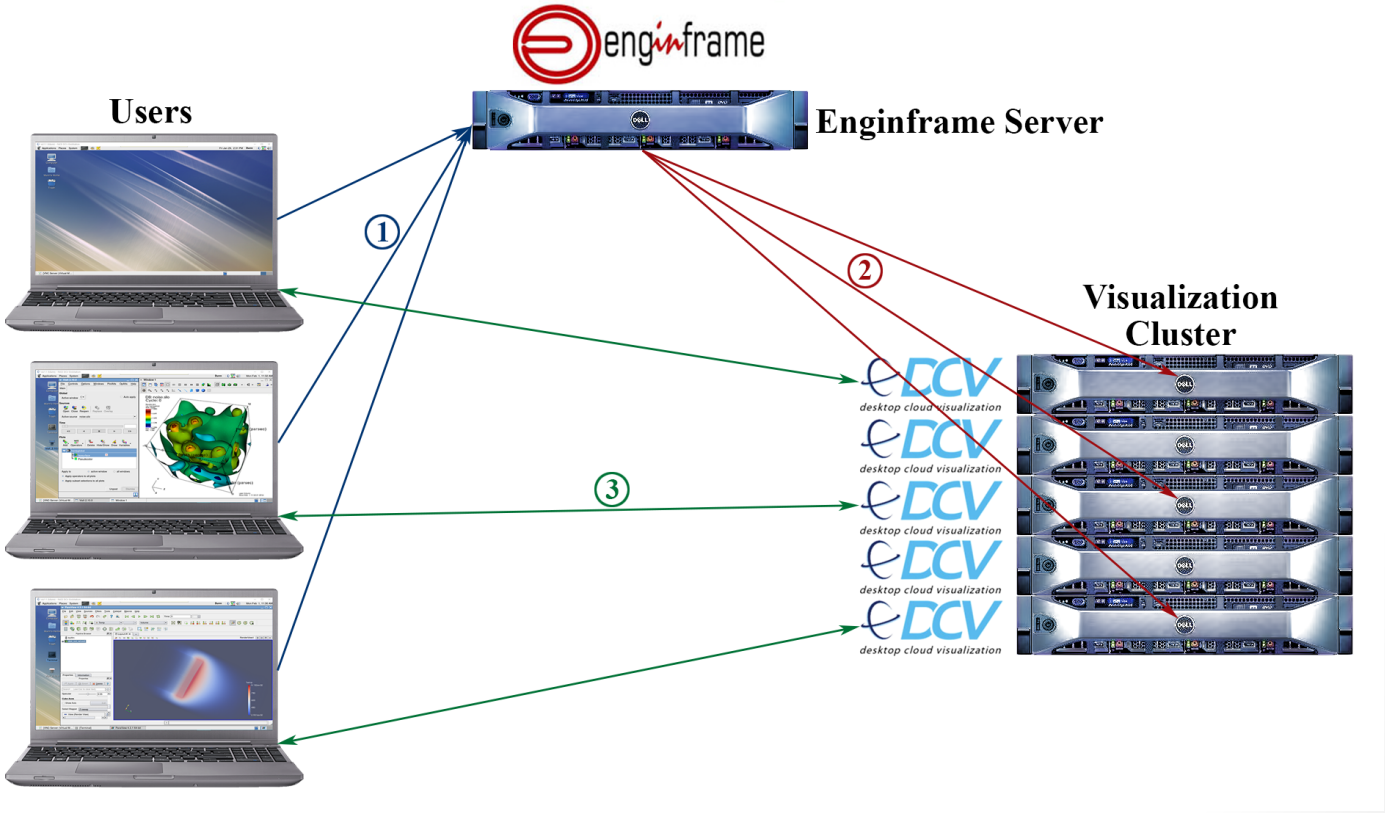
- Using native X-tunneling is hard and very slow.
- Using native VNC is less-hard, slightly faster but not always available.
- Using ‘native’ methods means you are transferring all of your data and updates back and forth so if you are playing with large datasets it could take a very long time for updates.
- If working only on a local machine you may not have decent or even required graphic capabilities.
- If working on a local machine you probably do not have a copy of your data and if you do it may be to large to run.
- Almost impossible to collaborate your work with others.

Remote HPC Visualization

- Using remote HPC visualization eliminates the X-tunneling connectivity issues!
- Only keystroke and mouse commands travel to the remote desktop and only updated screenshots travels back to the user so vastly less bandwidth issues!
- Remote visualization clusters utilize the latest and greatest graphic hardware!
- You work right beside your datasets!
- You can easily share your work, live, with other users.

EnginFrame Remote Desktop Architecture

1. User requests a remote desktop job via EnginFrame.
2. EnginFrame creates a new job and starts a DCV-VNC session and launches a Remote Desktop.
3. The user connects to the DCV Remote Desktop and does their desired work. Only keystroke and mouse commands travel to the Remote Desktop and only updated screen shots from the Remote Desktop travels to the user.



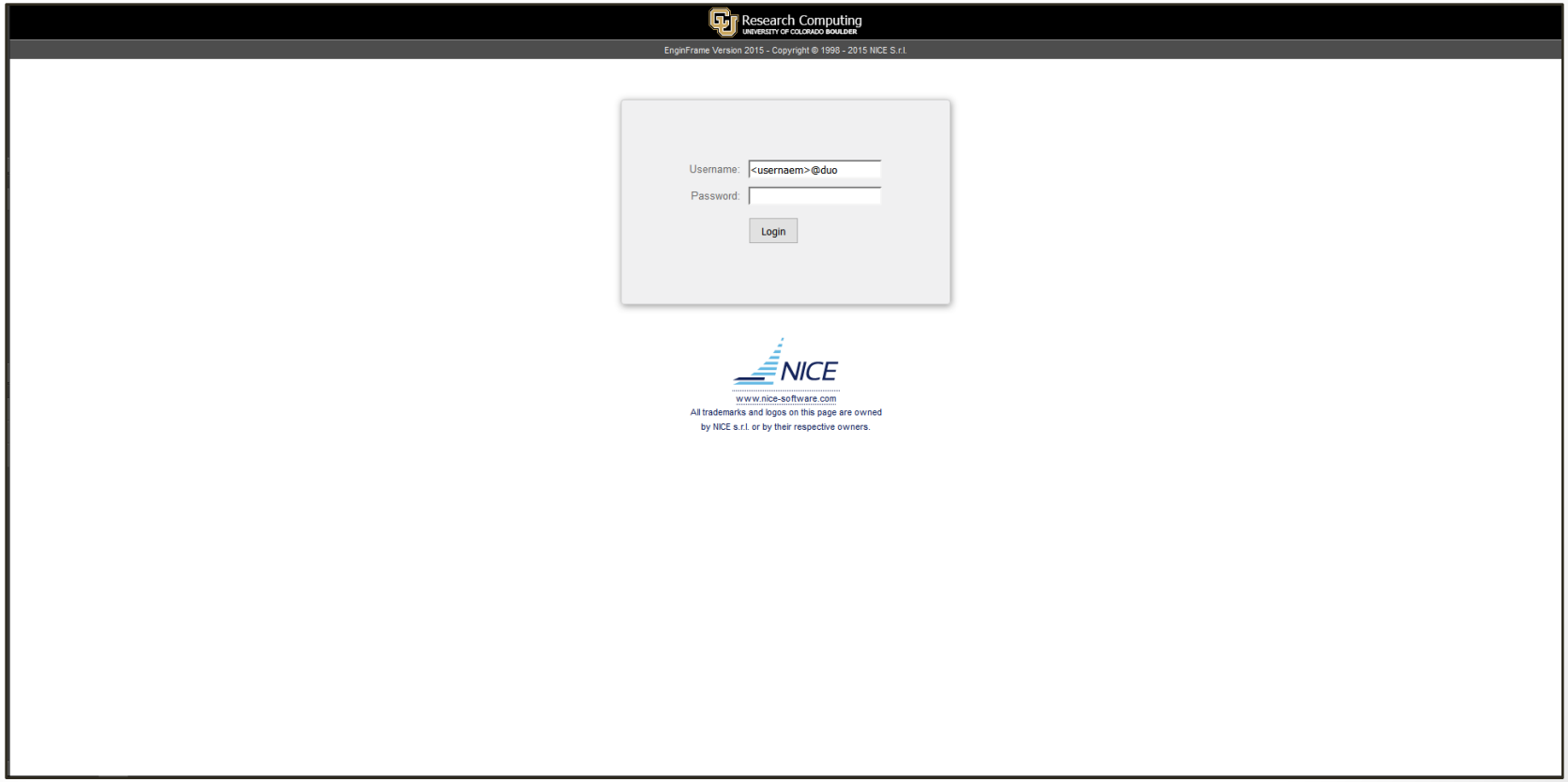
Requirements to Use EnginFrame

In order to access the visualization cluster, users must meet the following requirements:

- You must have an account with CU Research Computing.
- You must have a ‘Duo’ dual authentication account through CU Research Computing.
- You must have access to the internet and an internet browser.
- You must install the NICE DCV Endstation for your operating system. This can be obtained from;
  http://www.nice-software.com/download/nice-dcv-2016

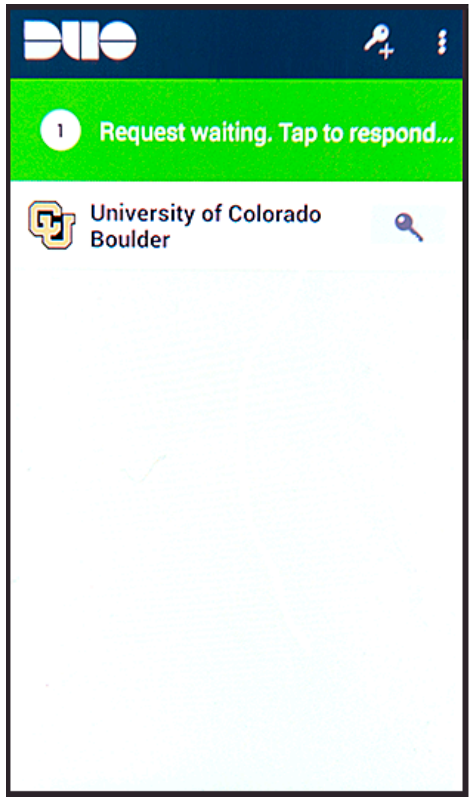
Logging into EnginFrame

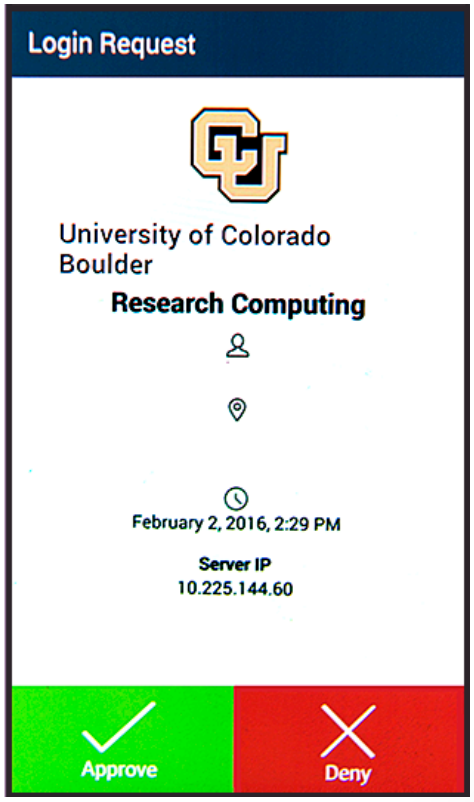
- From a browser navigate to : https://viz1.rc.colorado.edu/enginframe.
- For 'Username' enter: *<Identikey username>*
- For Password enter: duo:*<Identikey password>*.
- Click 'Login' to begin Duo authentication.



Duo Authentication

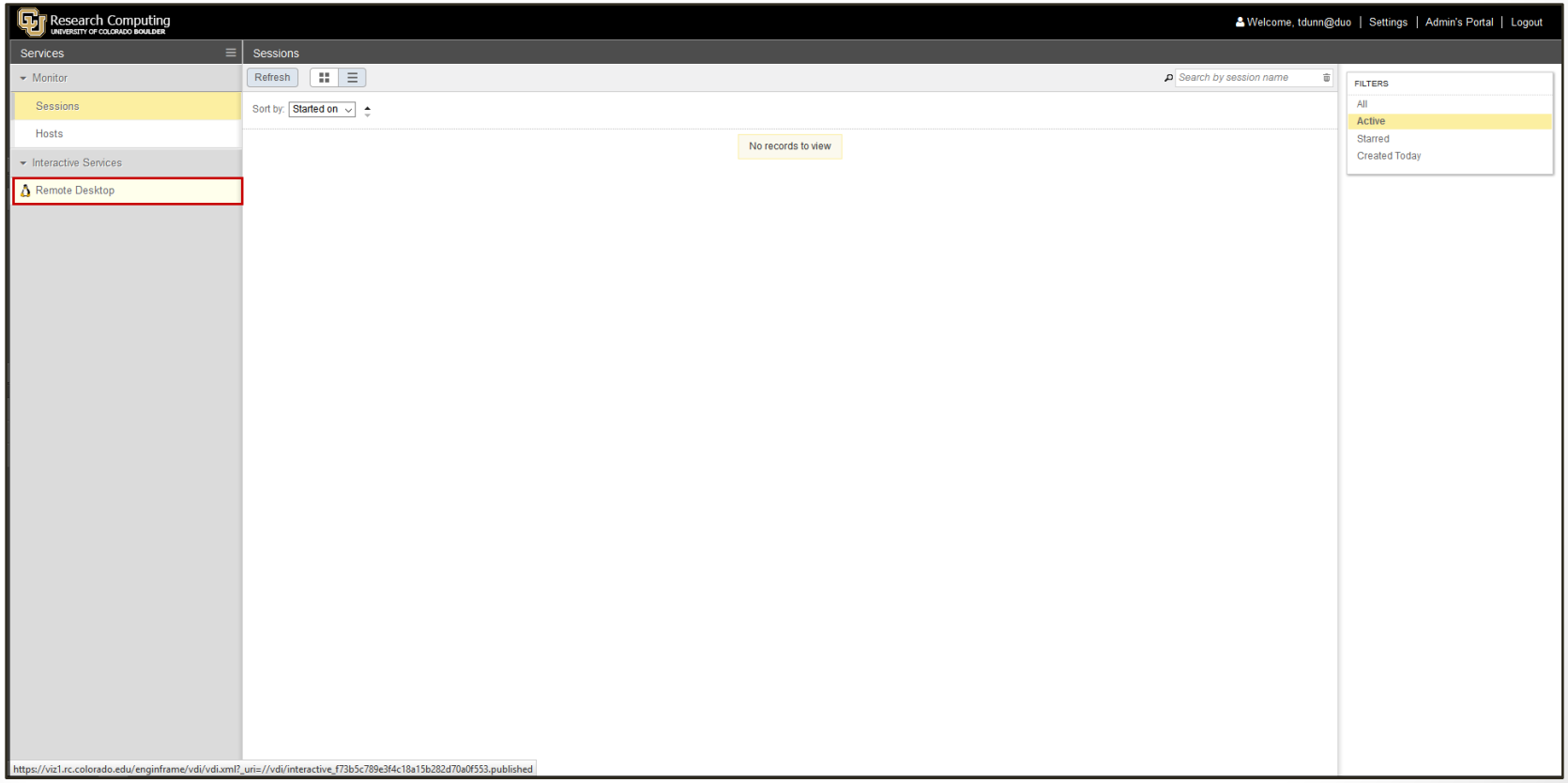
- After clicking on ‘Login’ your phone will alert you to an incoming Duo authentication.
- Open your Duo app and click on ‘Request waiting. Tap to Respond…’.
- Next click on ‘Approve’.
- You will receive a message informing you authentication was approved and the EnginFrame VDI page will appear (see next slide).
- **Note**: App graphics will vary by phone type.





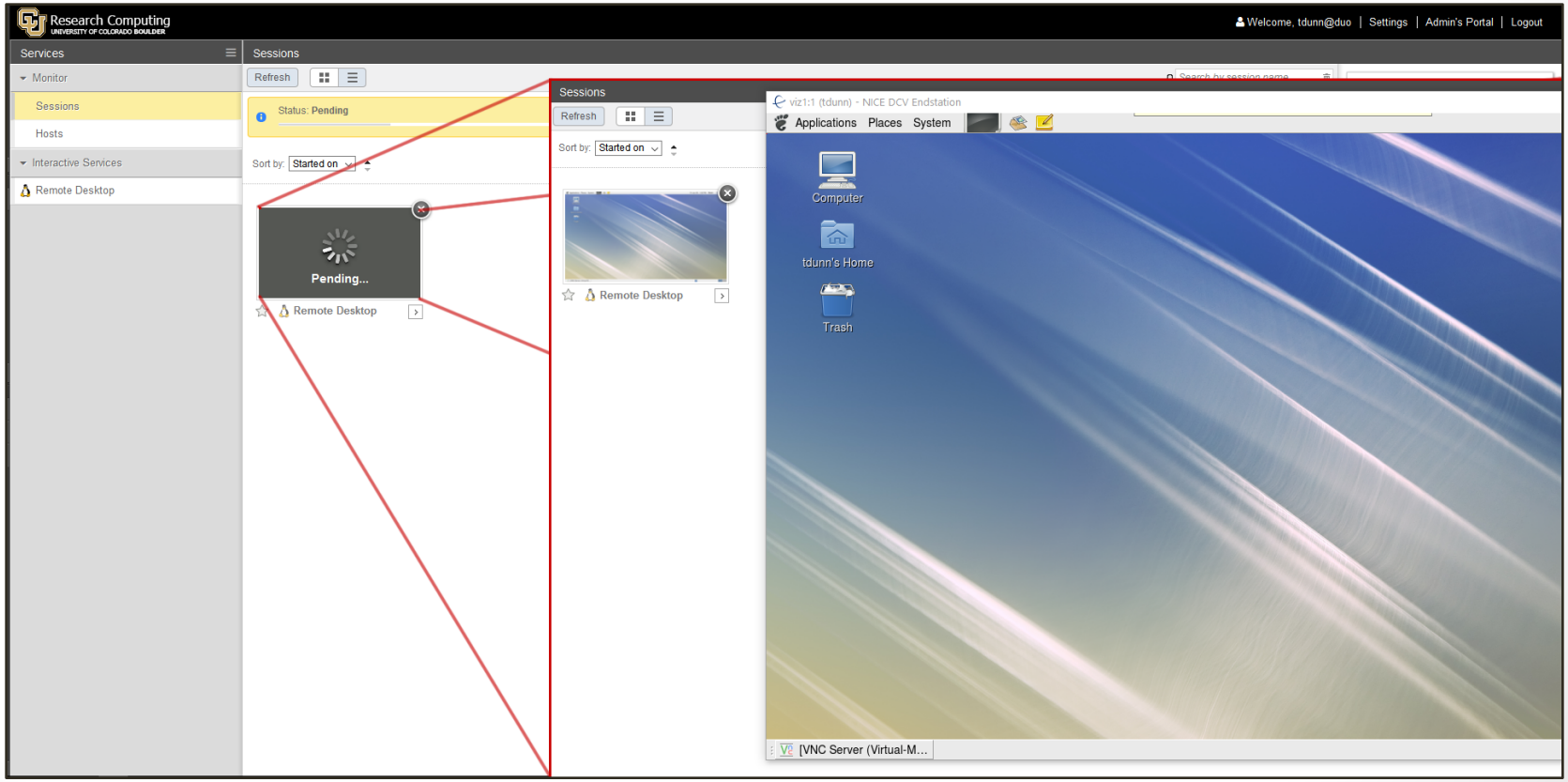
Starting a Remote Desktop Session

- To create a Remote Desktop click on the ‘Remote Desktop’ link under ‘Interactive Services’.



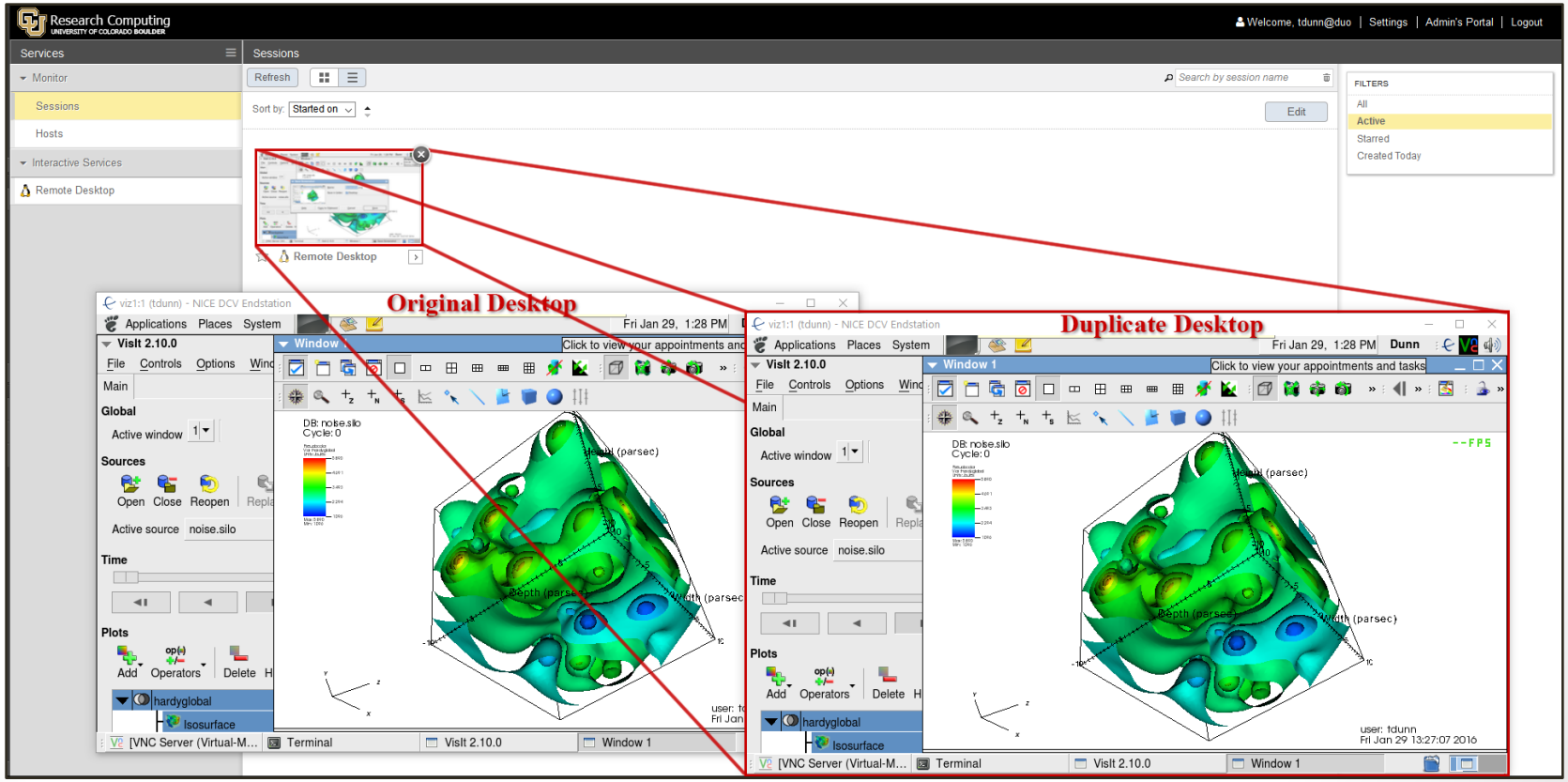
Connecting to a Remote Desktop Session

- You will be in a ‘Pending’ status state as you wait for the job scheduler, SLURM, to start a new job on the visualization cluster.
- Once a job has started the remote desktop will start and launch your VNC Linux Remote Desktop.
- A thumbnail of the remote desktop will be displayed in the Sessions browser



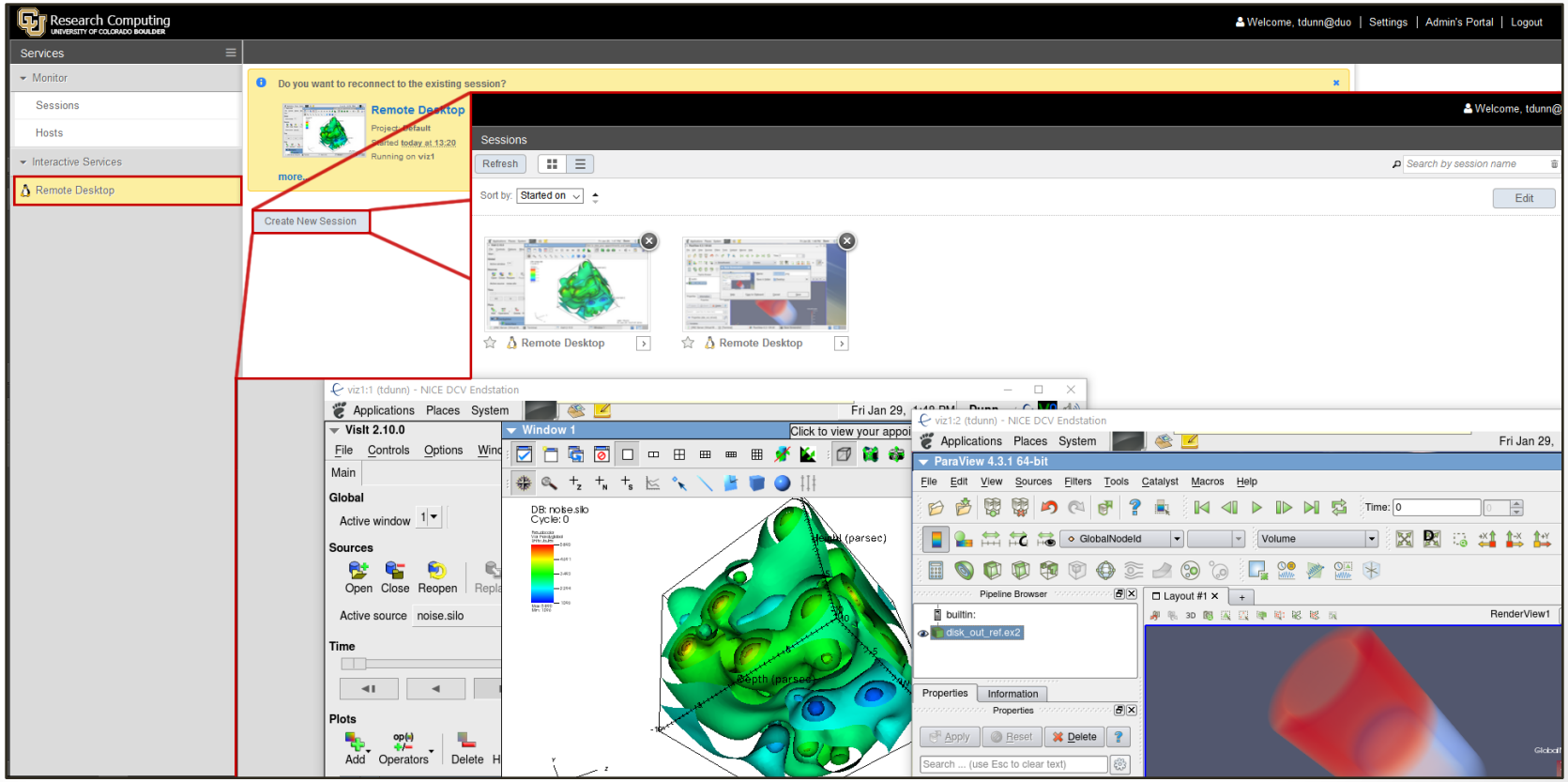
Creating Multiple Concurrent Sessions

- By clicking on your ‘Remote Desktop’ in the Sessions browser you can duplicate your ‘Remote Desktop’ session.
- Changes you make in one Remote Desktop is reflected in the duplicate session(s).



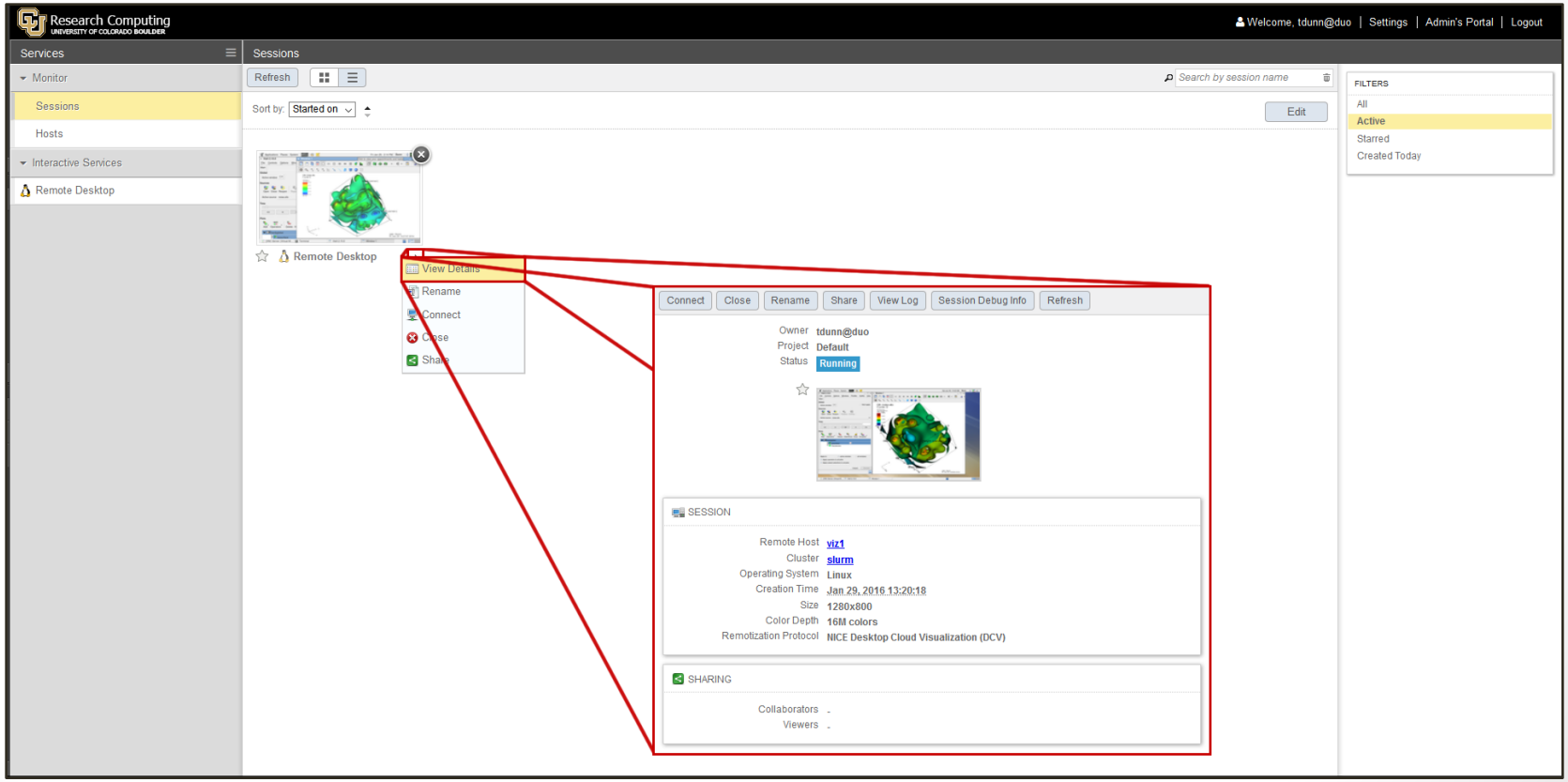
Creating Multiple Separate Sessions

- By clicking on your ‘Remote Desktop’ under ‘Interactive Services’ you can create a new, separate ‘Remote Desktop’ session.
- Changes you make in one Remote Desktop are NOT reflected in the new session(s).



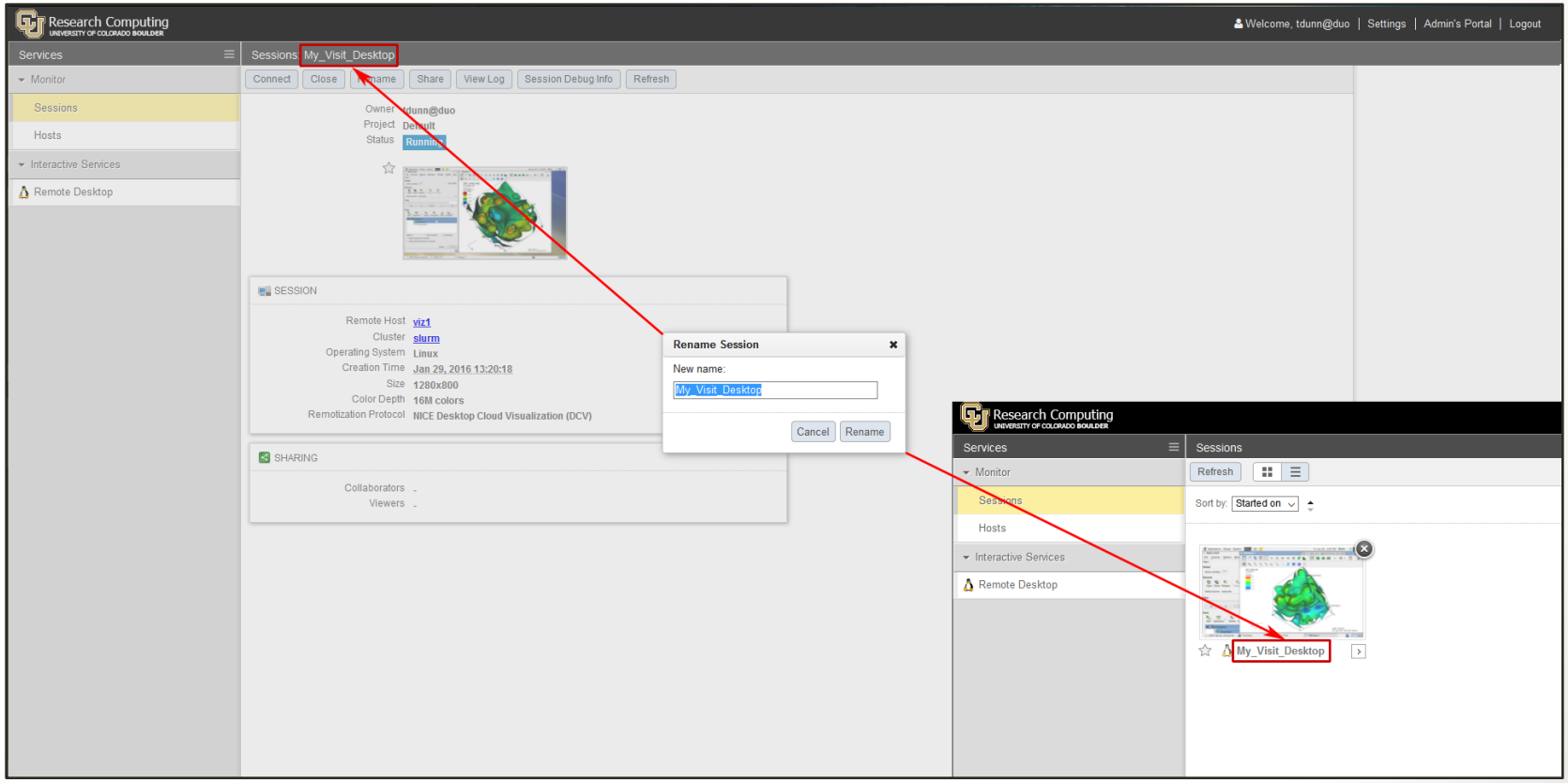
Session Details

- Clicking on the arrow next to your ‘Remote Desktop’ will open a hotkey menu.
- Selecting ‘View Details’ will send you to the ‘View Details’ page.
- The ‘View Detail’ hotkey menu and browser page allows you to ‘Connect’, ‘Close’, ‘Rename’, ‘Share’, ‘View’ logs, and ‘Refresh’ your remote desktop session(s).
- You can return to the Sessions browser by clicking on ‘Sessions’ under the ‘Monitor’ service.



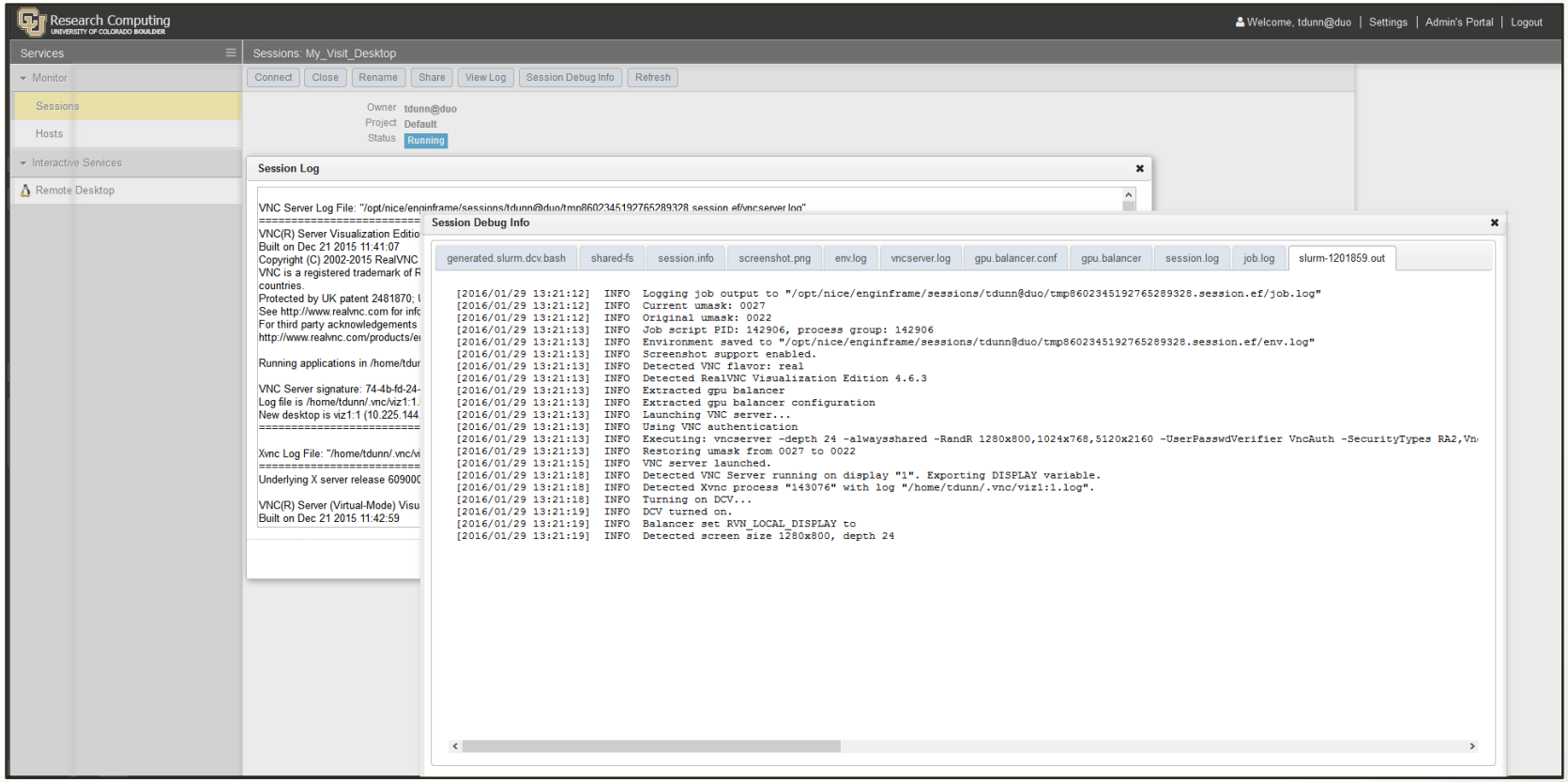
Renaming Sessions

- Renaming allows you to change the name of your session so if you have multiple sessions and/or sharing your work with others they an be easily identified.



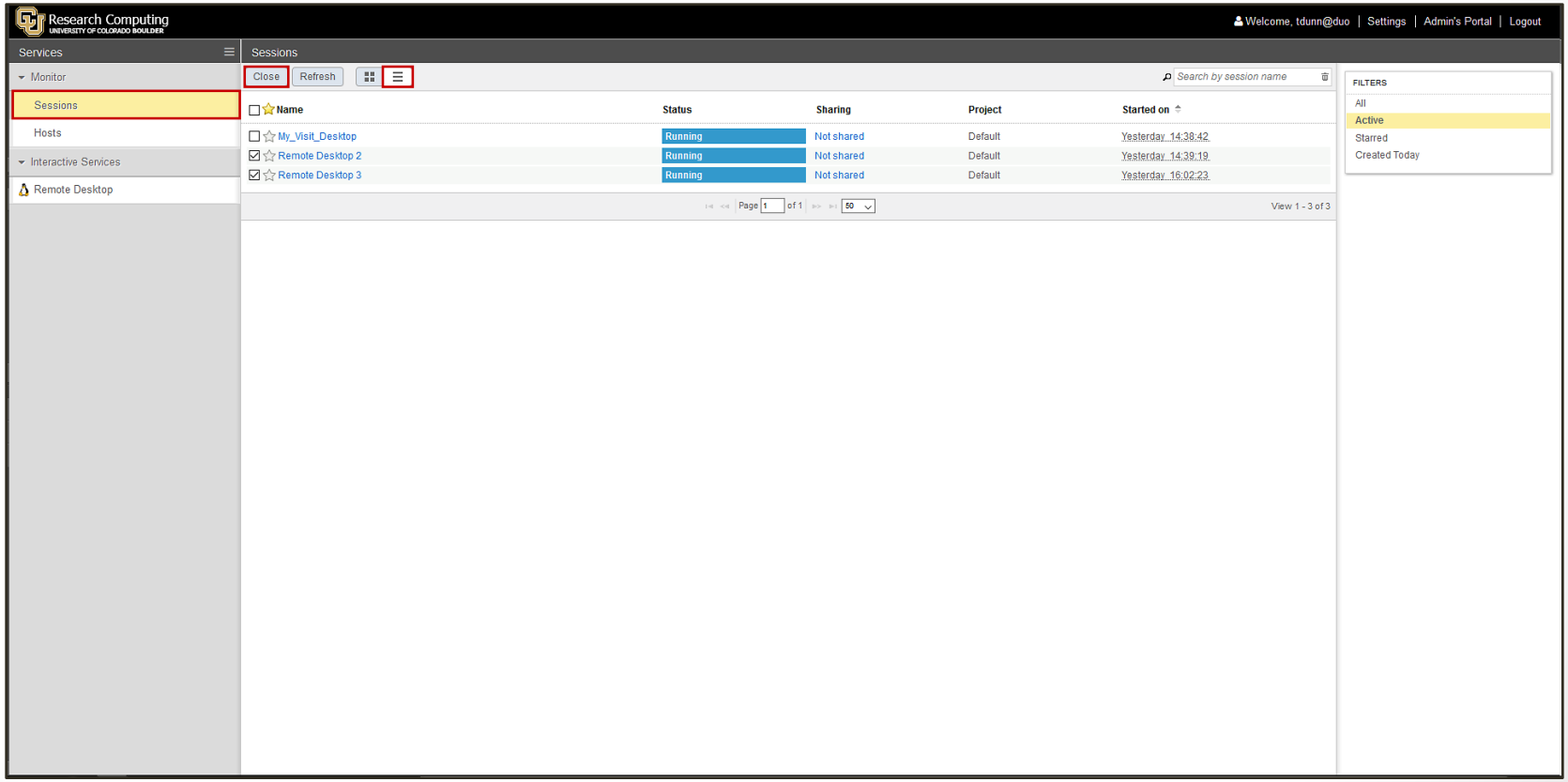
View Logs

- The ‘View Log’ and ‘Sessions Debug Info’ selections allow you to look at some of the various logs that EnginFrame generates.
- If you run into problems with your sessions and you need to submit a ticket to rc-help@colorado.edu, you may be asked to provide some of this information to help resolve your issues.



Closing a Session

- You can close a session by clicking on ‘Close’ or clicking the ‘X’ button at the top right of a the sessions thumbnail image.
- Alternatively, to close one or more sessions at once from the Sessions browser, click on the ‘List’ view icon, choose which session(s) you desire to close and click on ‘Close’.

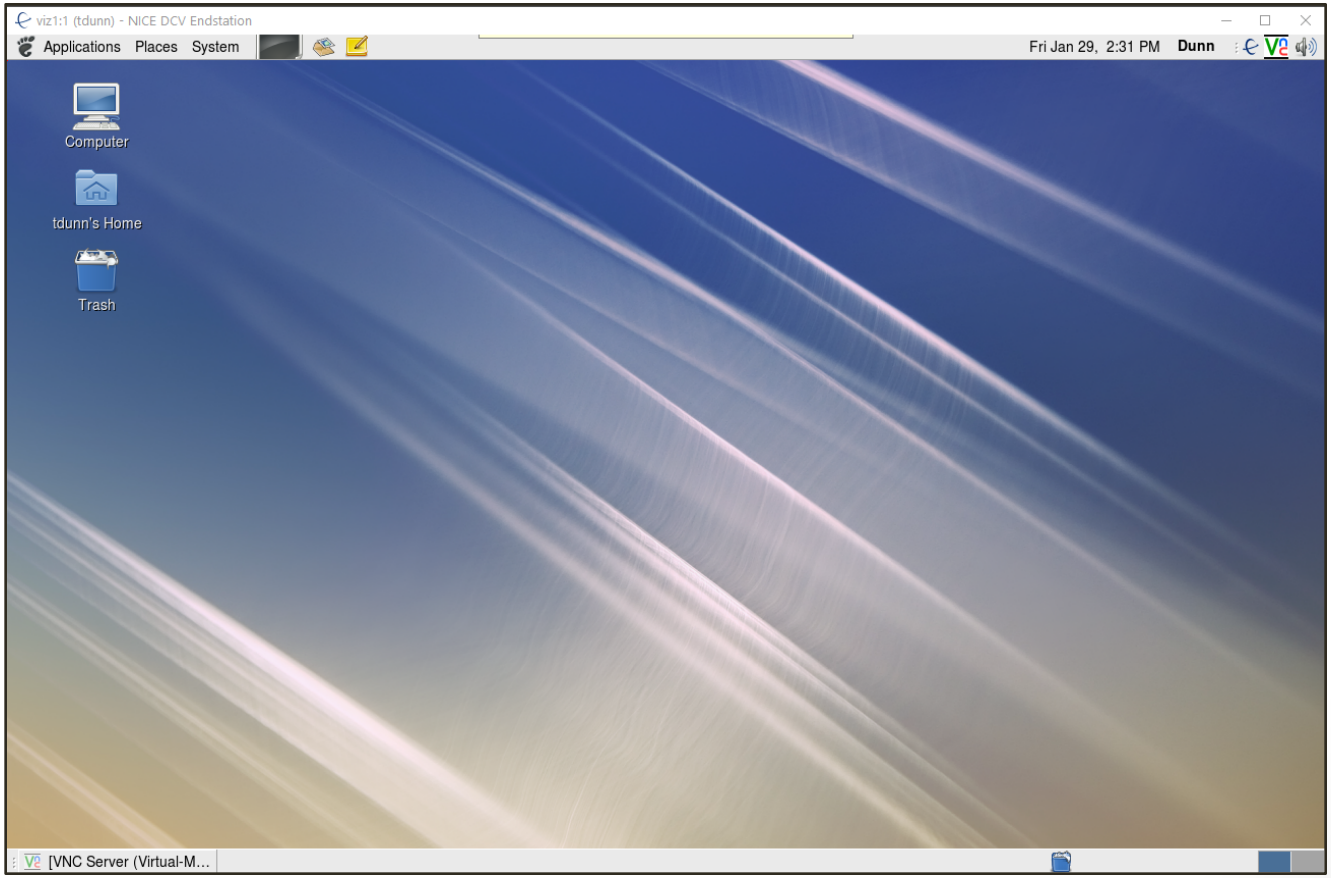


Session Lifetime

- Any remote desktop session you create will be available for a maximum of 24 hours.
- During that 24 hour lifetime your session resides fully on the visualization cluster. Thus;
  - If you close your browser tab or even the browser itself, you can open a new browser/browser tab, log back into EnginFrame and reconnect to your session(s).
  - If you turn of your machine off or loose power to it you can restart it, log back into EnginFrame, and reconnect to your session(s).
  - After 24 hours you will loose anything not saved either by you or by an application specific backup file system.
  - If the visualization cluster looses power or if SLURM issues arise you may loose your session and any work not saved will be lost.

The Remote Desktop

- When your Remote Desktop starts you will have a new DCV desktop running Redhat Linux with a slim Gnome desktop.
- You will automatically be set to your default CURC home directory running what ever environment you have set up.



Running a Visualization Application

- To run a visualization application on the visualization cluster you may use either the ones installed on Janus (or Summit in a few months) by loading the appropriate module and then launching it.

- It is important to note that only the new LMOD module system will be accessible from the visualization cluster. If you already have an account on Janus then you just need to open a terminal window, making sure you are in your home directory, and typing;

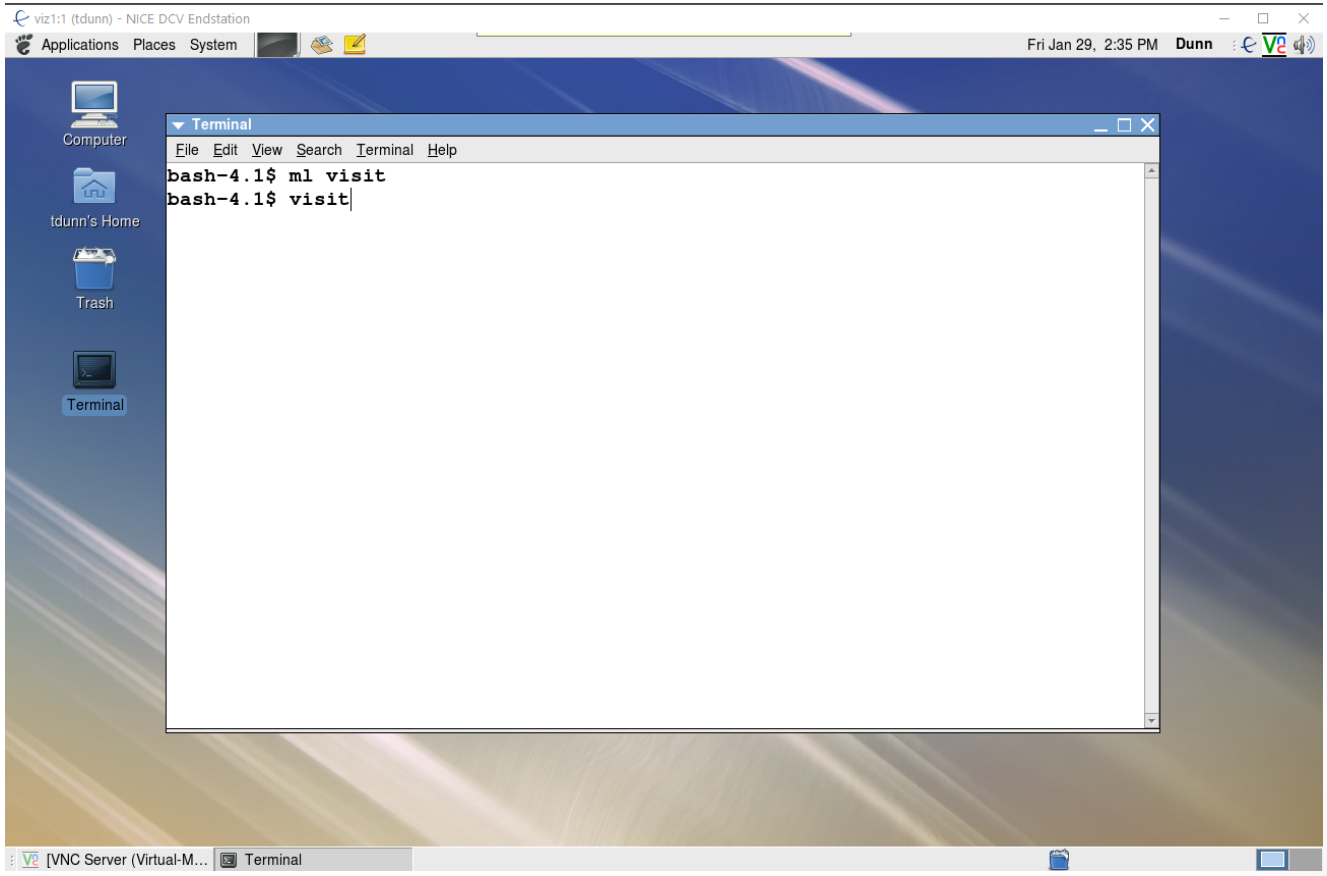
  touch ~/.lmodrc.lua

  Next restart your terminal session and you will now be setup to access LMOD modules.

- If you have your own visualization application installed on Janus (or Summit) you can just start it as defined by how you setup your installation of the application.

Running a Visualization Application

- To start a visualization application (e.g. Visit), open a terminal window and type the following commands.
- ml <*name of module*> where ml is LMOD's shorthand for module load.
- <*name of application to run*>



What NICE is and What it isn't

- It is:
  - A place to visualize data easily without knowing how to run jobs on a supercomputer.
  - Faster and easier than x-tunneling.
  - Easy to work with data because your data is next to you.
  - Provides a collaborative working environment.
  - It's one component of several parts of the Analytics Hub.

- What it isn't:
  - A place where plots are magically created
    - You still need to know your software!

Getting Help

- Come to us with questions on:
  - I need help with my script!
  - My script is running too slow…can it go faster?
  - I need to code a task, but don't know how
  - Can you make my visualization better?
  - How do I access this resource?
  - …
- When in doubt, ask!

Credit: Max Joseph

Getting Help

- Two methods of asking:
  - Ticket system
    - Send an email to el-help@colorado.edu
    - Will enter a ticket into our system
    - Best for us since once we talk to you we will submit a ticket anyway
  - Office hours!
    - Tim and Max are holding office hours in their Earth Lab office
    - Wednesdays, 1-2 pm
    - Stop by!

Questions

- Office hours, trainings, etc are on Earth Lab's calendar
- We keep it up to date!

- Questions? Shelley.Knuth@colorado.edu

- Slides: https://github.com/earthlab/trainings/blob/master/20160607_earth_lab_workflow.pptx

- Survey: http://tinyurl.com/curc-survey16