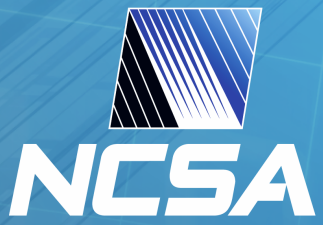
Storage for ICCP

Leveraging Spectrum Scale to isolate I/O and unify resources for campus

JD Maloney
Storage Engineer

Spectrum Scale User Group Meeting
June 10, 2016



National Center for Supercomputing Applications
University of Illinois at Urbana–Champaign

Illinois Campus Cluster Program

- Operated by NCSA, Campus makes investment & use
- Investor based, buy in cluster
- Investors get options of Infiniband or Ethernet
- Program is striving to bring prior non-HPC fields of study into the program
  - Fields such as linguistics, liberal arts, statistics
  - Users new to Linux environment

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

NCSA

Program Challenges

- Multiple fabrics in single cluster
- Users who still use Windows based applications for parts of their workflow
  - These usually run on machines across campus
  - Require batch file processing
- Users who still want “Windows like” view of file system, through Explorer
- Desire of users to have file system mounted in many places outside the cluster

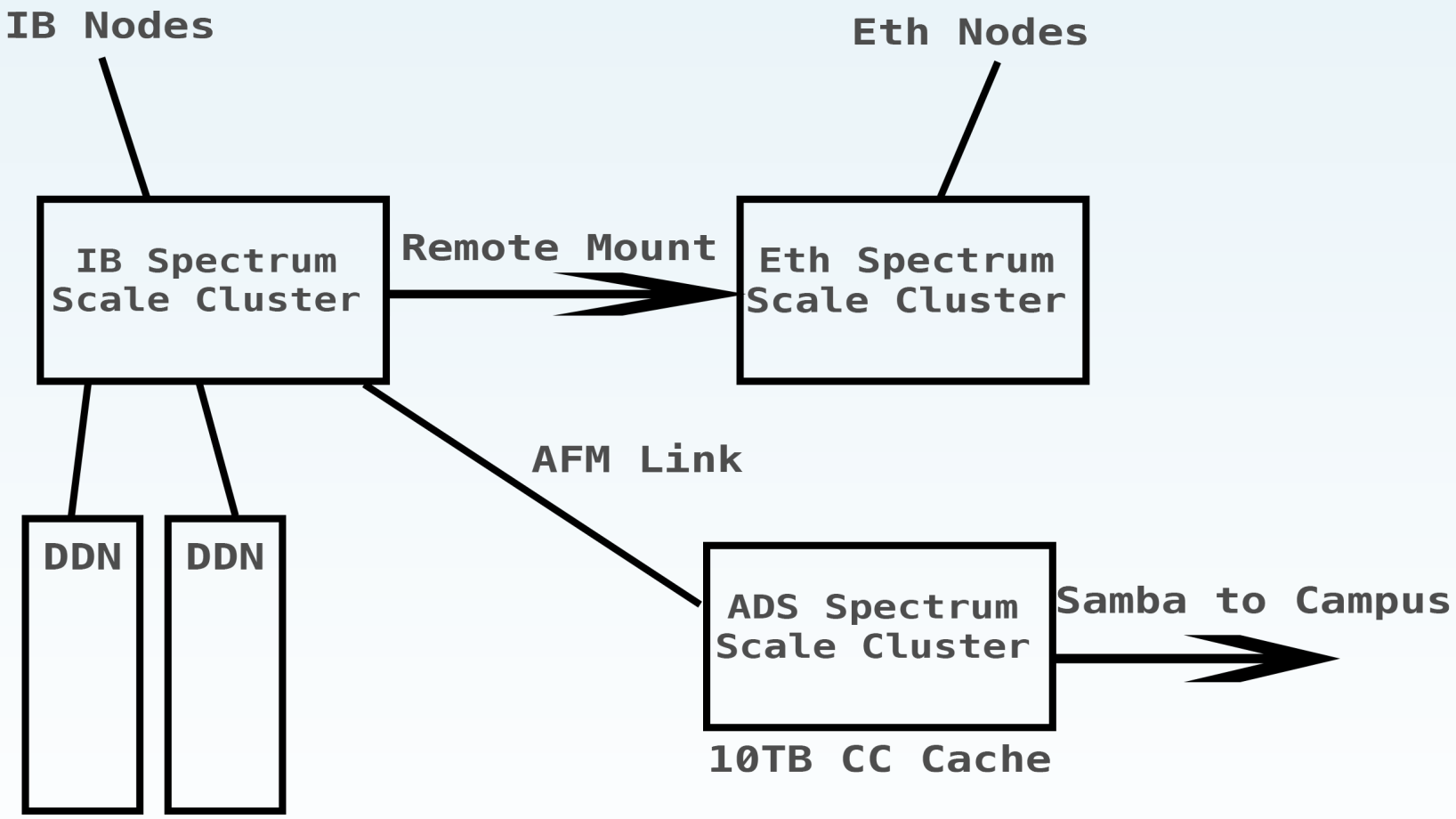
Campus Active Data Storage (ADS)

- Spectrum Scale File System to hold research data that is still in use by researchers across campus in central location
- Bulk storage, no processing against file system
- Resides in same network space as the ICCP Cluster
- Departments buy into the storage on the system based on their needs $/TB

Goals

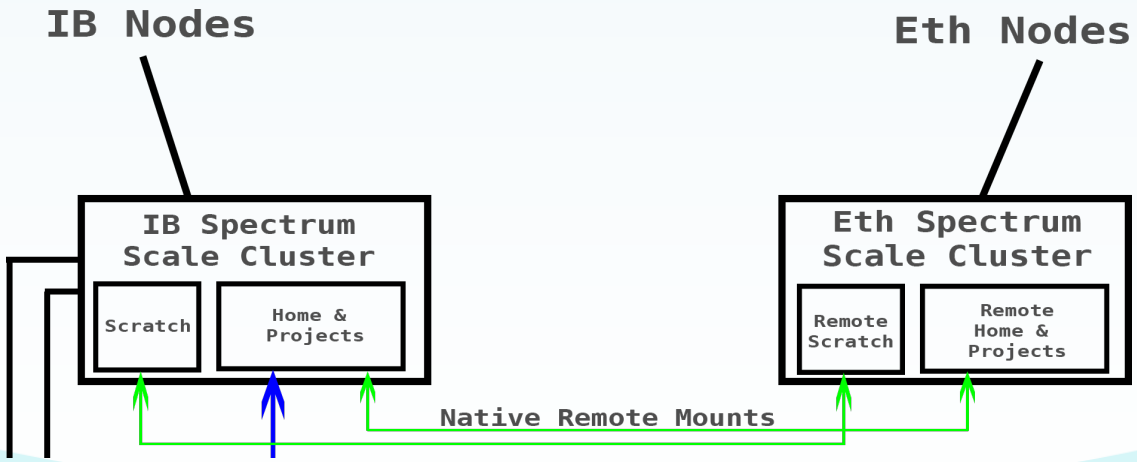
- Offer services users want
  - Samba/NFS/CIFS mounts to some machines on campus of the cluster file system (project & home space)
  - The ability to have different fabric options for purchase
  - Single file system across them all
- Efficiently use campus resources
  - Leverage existing systems to solve problems
  - Consolidate where possible
- Isolate disruptive I/O patterns to maintain performance
  - Prevent samba traffic and locks from hindering cluster performance
  - Leverage caching of common files for users off the main cluster system

Conceptual Overview



Remote Mounts to Bridge Fabrics

- Allows us to no longer use management interface on hosts to address nodes
- Helps reduce the amount of expulsion issues we see on the cluster
- Provides us with flexibility for changing the cluster management network topology and transitions from 1GbE to 10GbE and 40GbE on the data network



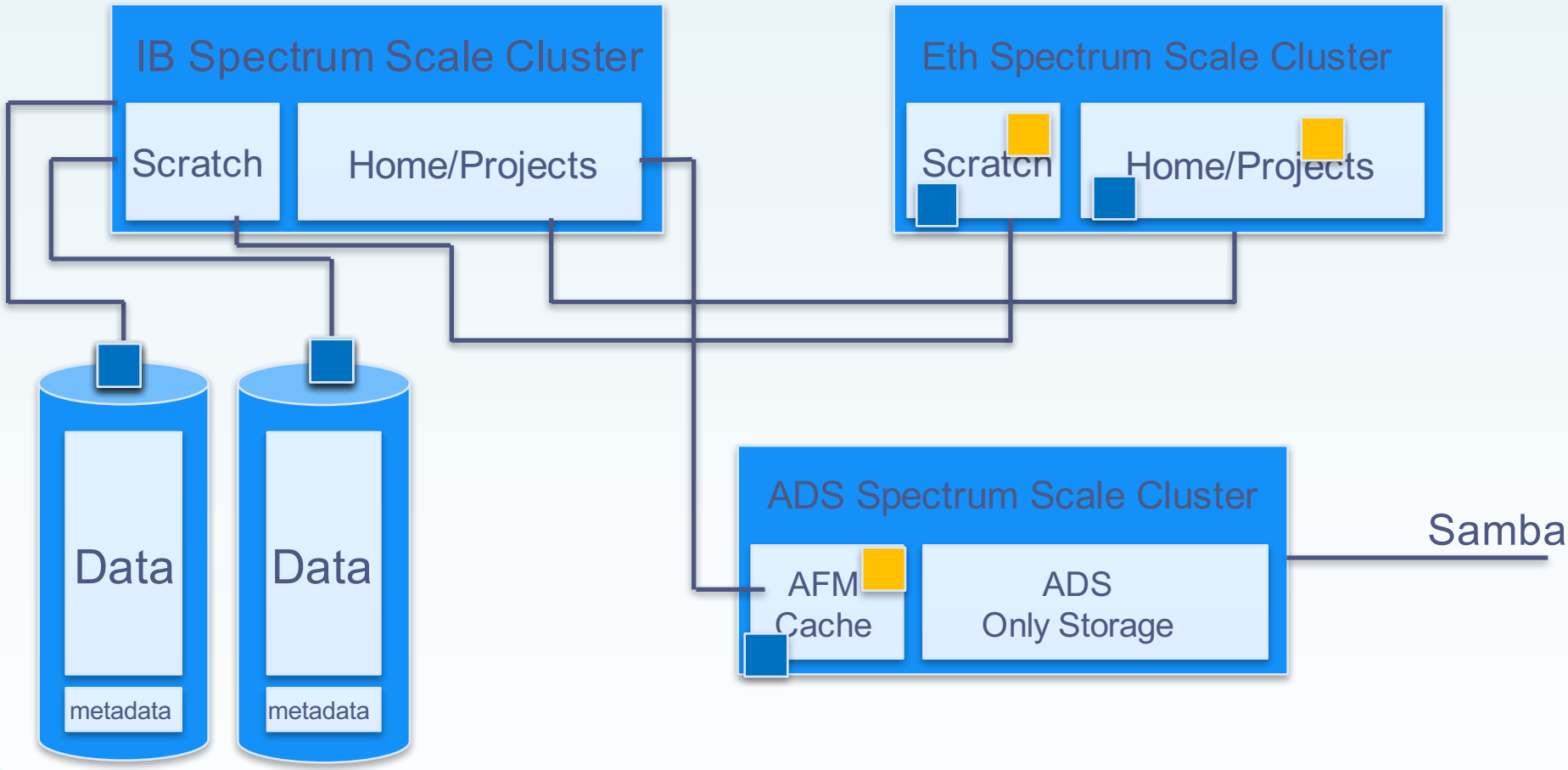
AFM to Offload Samba Exports

- AFM Cache sits on the ADS system which is separate disk from the cluster file system
- Already an export focused GPFS cluster to share with campus
- Connection between ADS and the Campus Cluster can act as a throttle for how much traffic, Samba services can consume
- Samba CPU load now sits off the production cluster NSD servers
- Great for users as performance needs are not high in terms of throughput

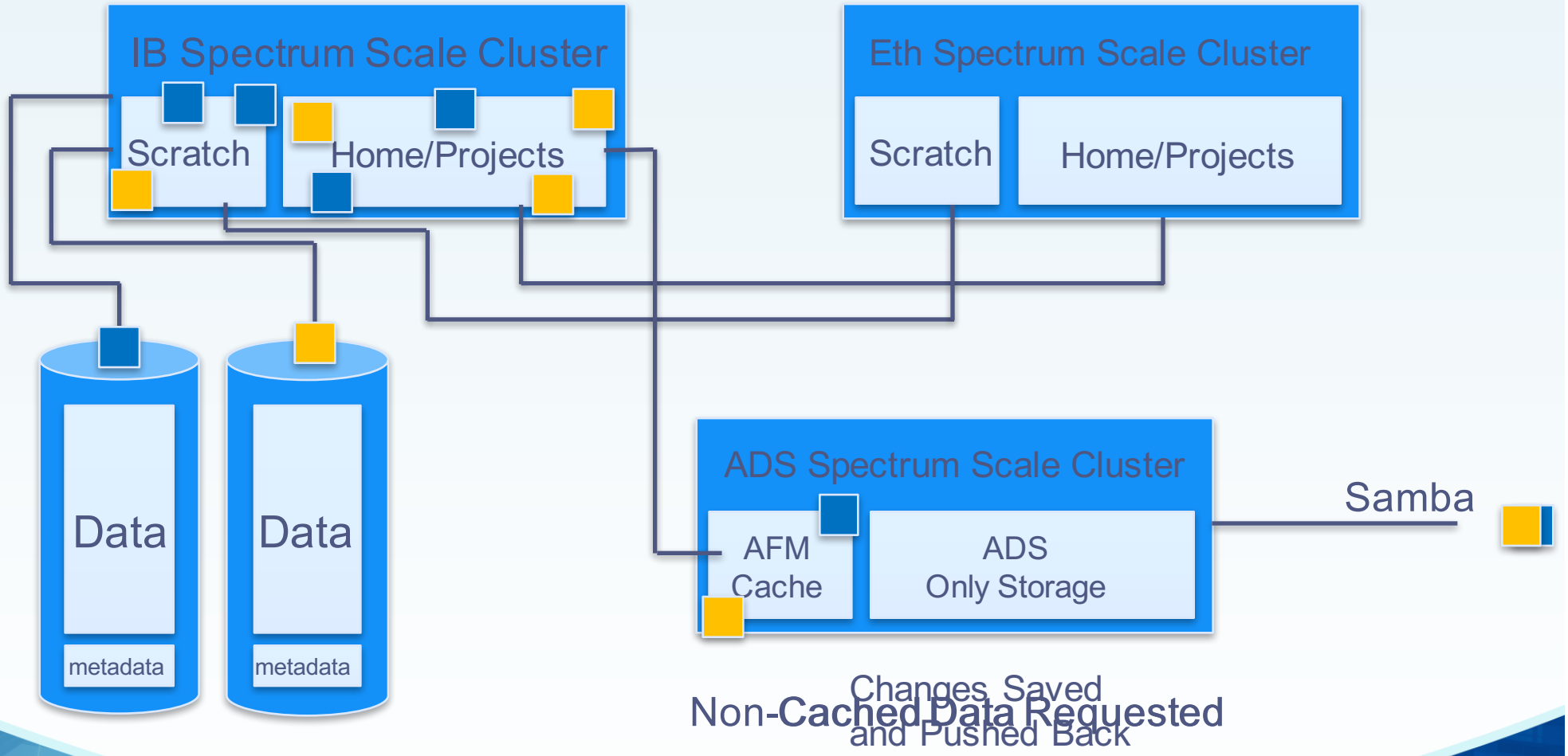
AFM Benefits over Remote Cluster or CES

- Allows cached files to be accessible during maintenance periods of the cluster file system to let users keep doing work
  - Clusters aren't tied together necessarily for maintenance
  - Certain investors have critical uptime needs
- CES from cluster file system wouldn't have the I/O cache of the ADS system and consume NSD servers that are already busy

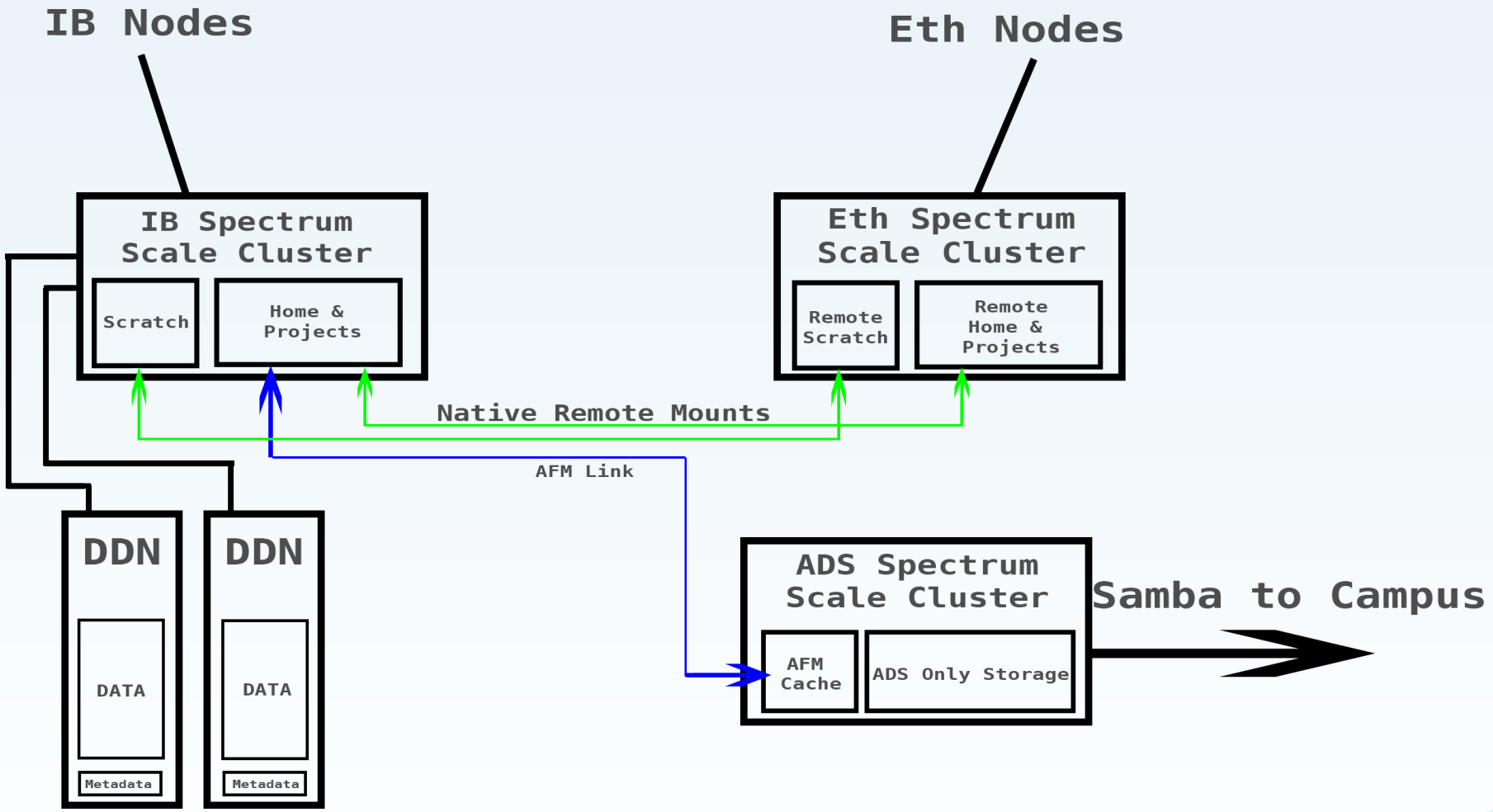
Data Flow Cluster Traffic



Data Flow AFM/Samba Traffic



Data Flow (no animation)



Rollout Ongoing

- Clusters recently updated to Spectrum Scale 4.2 from GPFS 3.5
- AFM cache preparing to go into friendly user mode
- Awaiting new hardware for the logical split of IB & Ethernet Spectrum Scale Clusters
- Working with users to retrain for mounts from new system

What We're Watching for

- Reduction in node expulsion due to fabric issues
- Improved performance on batch system due to reduced Samba load
- Amount of cache space that is adequate for users
  - Is 10TB enough? Should we add more?
  - Performance vs Capacity, small fast disks vs big slow disks
- Mount scaling across the campus network as usage increases

Questions?

Thoughts?