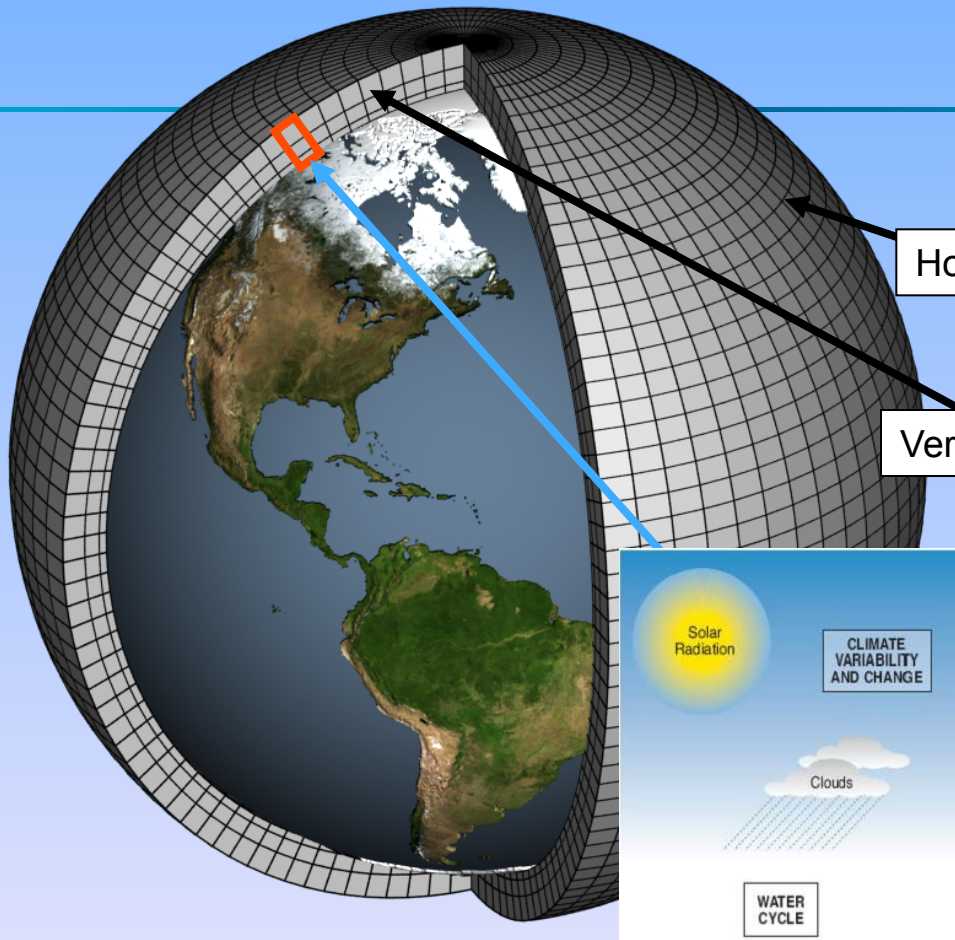
NCRC Grid Scheduling Environment

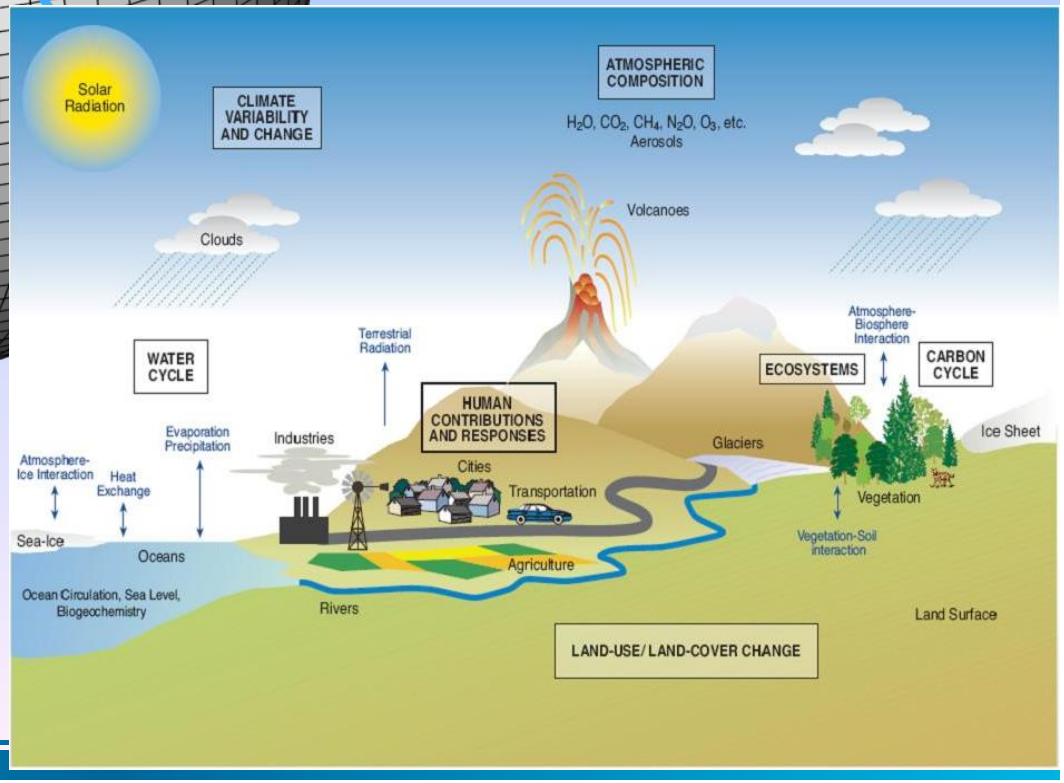
Frank Indiviglio & Don Maxwell

Schematic Global Climate Model



Horizontal Grid (Latitude-Longitude)

Vertical Grid (Height or Pressure)



A GCM is a mathematical representation of the major climate system components and their interactions. The GCM equations operate on a global grid and are solved on a computer.

Moving to a Remote Computing Model

- Users are used to local resources.
  - Workflows require very tightly coupled events.
  - Requirements for the workflow:
    - Needs to be fully automated.
    - Needs a scheduling platform capable of supported multi-site events.

Moving to a Remote Computing Model

- New model has archival and post-processing at user's local centers.
- Shared computing resources located remotely.
- Computing resources had to be allocated between centers and groups.

Climate Modeling and Research System: Initial Capability (CMRS.1)

Cray XT6 LC

- 2,576 Socket G34 AMD 2.1 GHz 12-core Magny-Cours processors
- 30,912 compute cores, 1,288 24-core nodes
- 82.4 TB DDR3 memory, 64 GB/node, 2.67 GB/core
- Peak performance: 260 TF
- 14 cabinets in a 2x7 cabinet configuration
- Liquid cooled using Cray ECOphlex cooling technology
- Peak Electrical Consumption: 792 kVA
  - Peak demand to date: 512 kVA (64.6%)
- Cooling Requirement: 225 tons
  - Peak demand to date: 145 tons (64.4%)
- Connectivity to the external Lustre-based Fast Scratch and Long Term File Systems

02/16/2011

Remote Computing

Challenges

- Managing Job Streams that span multiple sites.
- Data Transfer for every job
- Multiple batch resources
- Several types of workflows
- Lots of moving parts

Grid Scheduling

- Grid scheduler:
  - Responsible for:
    - accounting
    - prioritization
    - and scheduling of jobs – even across centers!

- Jobs get handed off between instances
  - For example, when a batch run completes on the Research system, data is staged to GFDL, a job is submitted to the post-processing nodes, and the results are put into the Archive.
  - Some of the Data Movement is *currently* scheduled through the meta-scheduler.
  - Authentication is done through x.509 certificates.

Data Movement

OAK RIDGE National Laboratory

- *cp*, *cxfscp*, *gridftp*, *mpscp*, *mcp*, *rsync*, *scp*, and *spdcp* are just some of the initial data copy tools.
  - Currently evaluating other tools.
  - Data integrity is difficult to achieve.
- NOAA is now using a general copy tool, *gcp*, to wrapper the underlying utilities.
  - Users cannot be expected to know every copy utility.
  - Failure modes on the underlying utilities need to be handled.

Moab features support NCRC mission

- Users
  - Showstart
  - Showbf
  - Checkjob
    - Why is my job not running?
- Systems
  - Advance Reservations
    - Maintenance
    - Troubleshooting and testing hardware (target individual node)
  - Good diagnostic tools
  - Dynamic Backfill for high resource utilization
  - Standing Reservation for debug and interactive work

Requirements

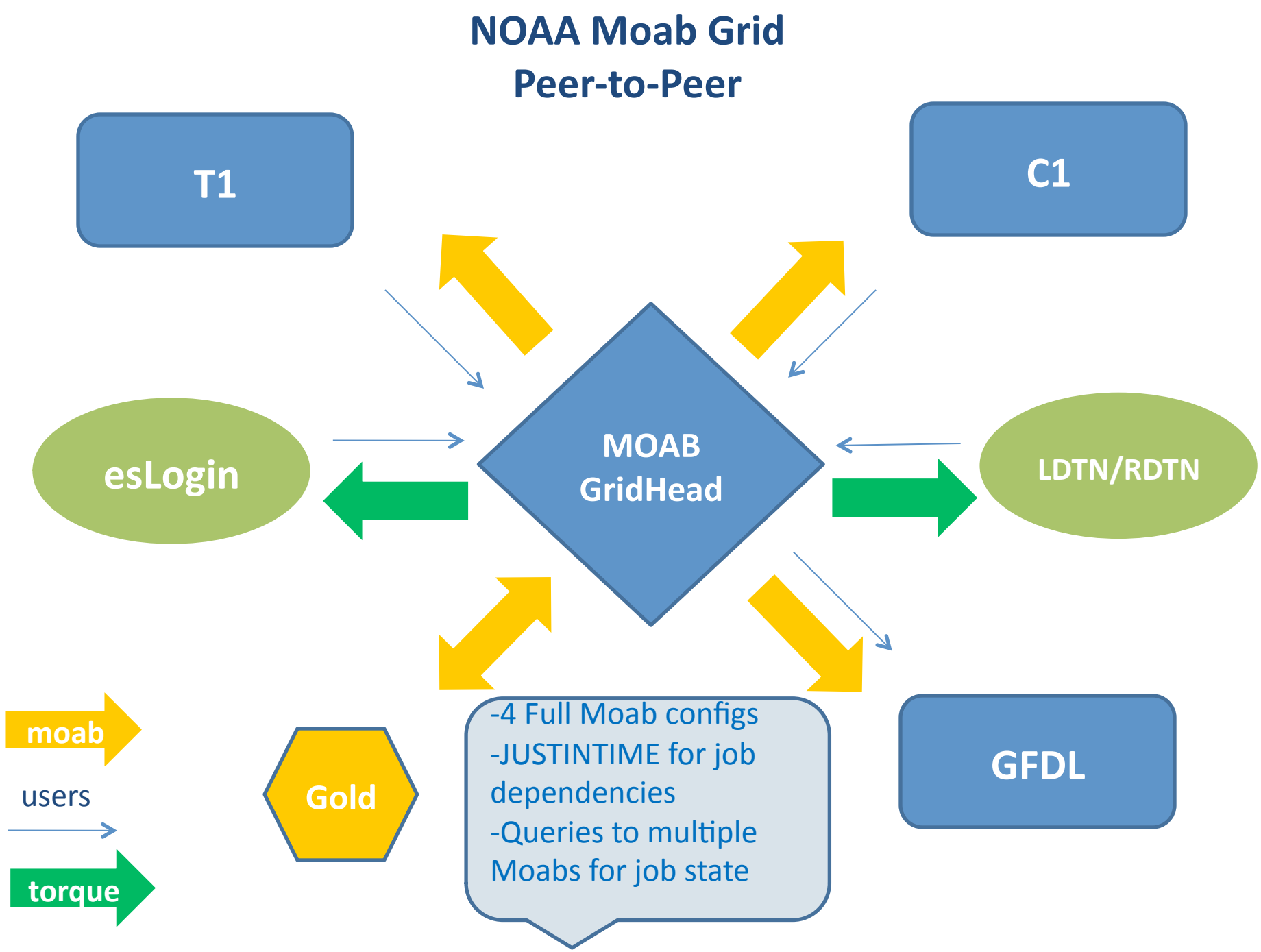
- Use fairshare to attempt to promote steady allocation usage throughout the month
- 50% of the available time for high-priority persistent and urgent work
- Novel queue for jobs that have unusual resource requirements typically needing more than 25% of the system
- Windfall queue for work that would not be charged against an allocation

NCRC
Moab Priority Implementation

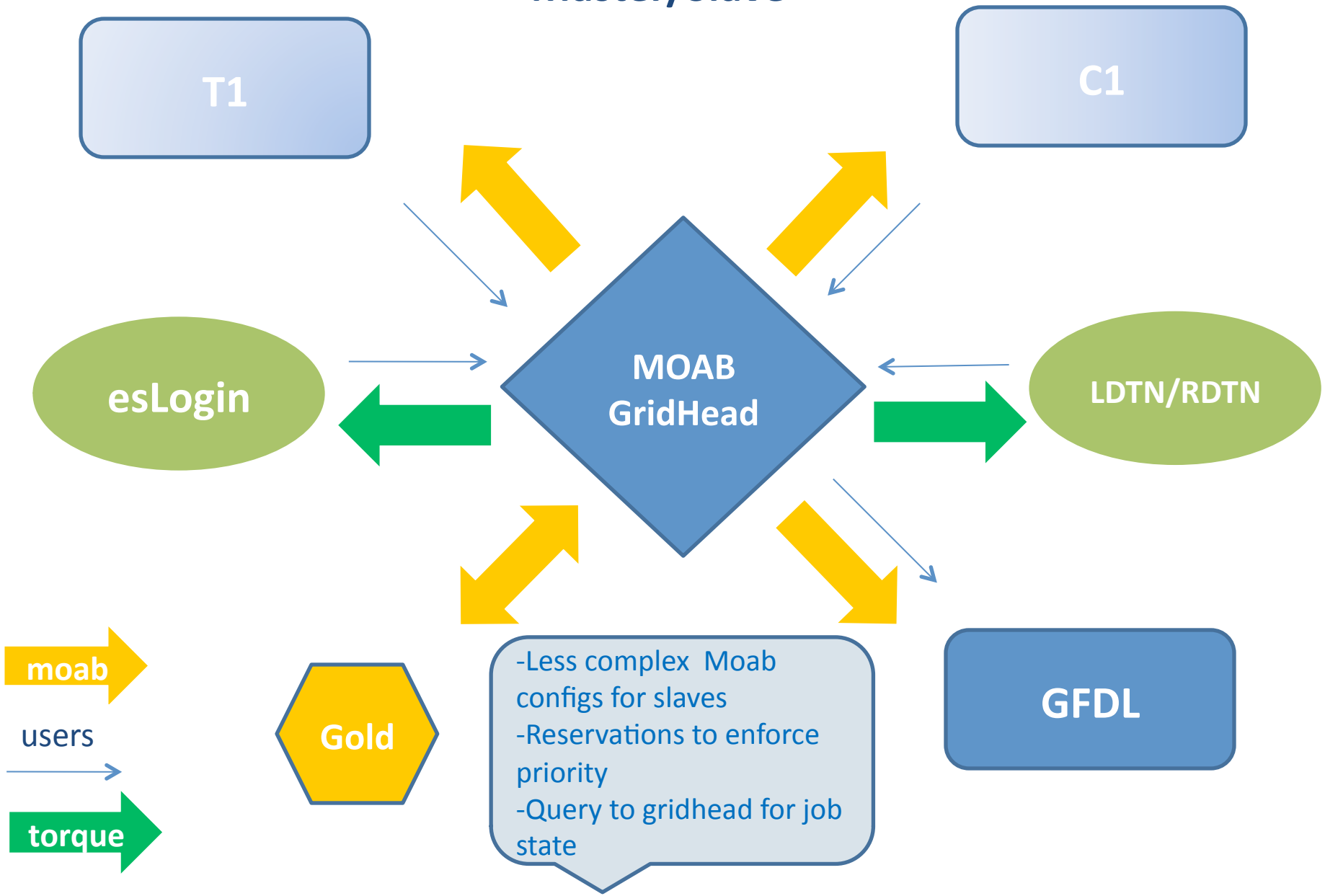
Factor	Unit of Weight	Actual Weight (Minutes)	Value
Class	# of days	1440	Urgent (10) Persistent (5) Debug (2) Batch (1) Windfall (-365)
Fairshare	# of minutes	1	(<>)5% user (+/-) 30 minutes (<>)5% class (+/-) 60 minutes
Queue Time	1 minute	1	Provided by Moab

NCRC Gaea System

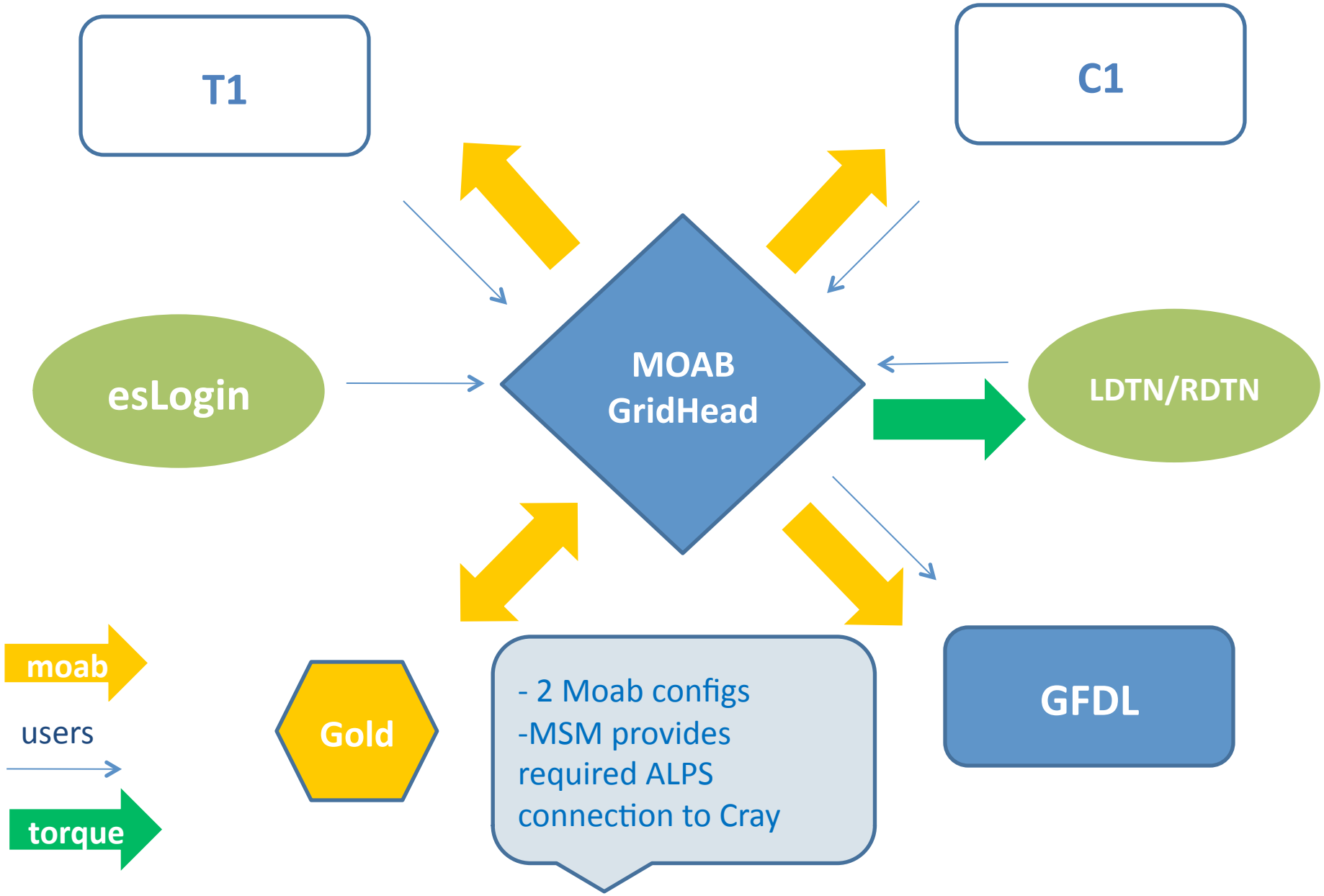
Compute Resource	Purpose
C1	Cray XT6 Compute Resource
T1	Cray XT6 Test Resource
esLogin	Login Nodes
LDTN	Local Data Transfer Nodes
RDTN	Remote Data Transfer Nodes
GFDL	Post Processing and Archival


NOAA Moab Grid
Peer-to-Peer
T1
C1
esLogin
MOAB
GridHead
LDTN/RDTN
GFDL
Gold
-4 Full Moab configs
-JUSTINTIME for job dependencies
-Queries to multiple Moabs for job state
moab
users
torque

NOAA Moab Grid
Master/Slave



NOAA Moab Grid
MSM



Along the way…

- **Lessons Learned**
  - **All Moab instances seeing jobs from all other Moab instances caused issues**
    - **Problematic with timeouts, hop counts being exceeded, job migration confusion, etc.**
    - **Unnecessary after pointing all clients at the gridhead**
- **New Features**
  - **Moab log management**
    - **LOGROLLACTION**
      - **First-failure data capture**
  - **One job number in the Moab grid**

Future

- Short-term
  - Master/Slave
  - Upgrade to Moab 6.x
- Long-term
  - Using MSM to externalize Moab from the XT
- Gold allocations to be incorporated into the fairshare configuration through the identity manager interface
- Add additional NOAA sites to Moab grid