

# Advanced HPC User Group Meeting

2019-12-06

Jan Moren

Scientific Computing and  
Data analysis section

Say hello to



**DE·i·GO**

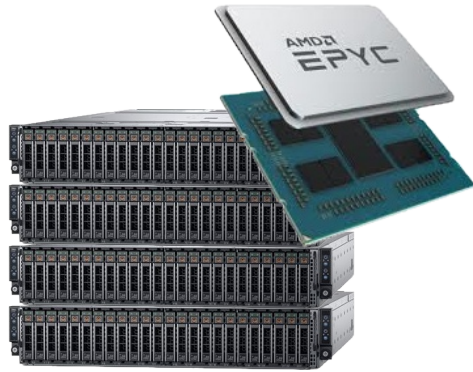
High-core AMD  
and Intel nodes

next-generation  
networking

ultra-high  
speed storage



# DE·i·GO

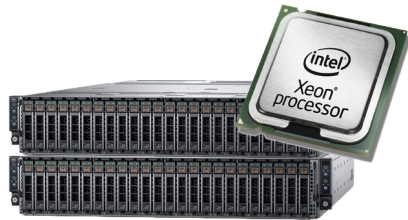


## 456 AMD nodes

2 x EPYC 7702 2.0GHz  
128 cores  
512GB memory

---

**58368 cores**



## 192 Intel nodes

2 x Xeon 6230 2.1GHz  
40 cores  
512GB memory

---

**7680 cores**

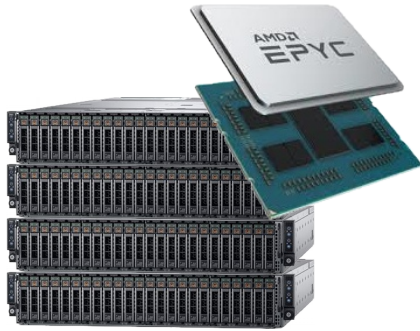
Sango: 9600 cores

**Deigo: 66048 cores**





# DE·i·GO

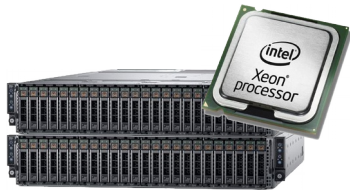


## 456 AMD nodes

2 x EPYC 7702 2.0GHz  
128 cores  
512GB memory

---

**58368 cores - 88%**



## 192 Intel nodes

2 x Xeon 6230 2.1GHz  
40 cores  
512GB memory

---

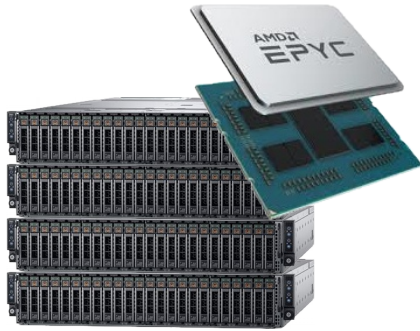
**7680 cores - 12%**

## Why both Intel and AMD?

- Per core:
  - AMD is a bit faster for integer, I/O
  - Intel is a bit faster FPU (esp. AVX512)
  - Depends *a lot* on your code
- Per node:
  - AMD **trounces** Intel



# DE·i·GO

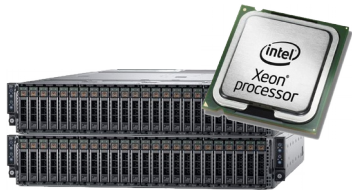


## 456 AMD nodes

2 x EPYC 7702 2.0GHz  
128 cores  
512GB memory

---

**58368 cores - 88%**



## 192 Intel nodes

2 x Xeon 6230 2.1GHz  
40 cores  
512GB memory

---

**7680 cores - 12%**

## Why both Intel and AMD?

- **But:** Intel MKL library can perform badly on non-Intel CPUs.
  - Intel checks CPU maker (not capability), selects operations based on that.
  - some BLAS operations - matrix multiplication - especially bad
  - some physics codes, Matlab affected
  - Can override maker check with:  
`export MKL_DEBUG_CPU_TYPE=5`  
MKL up to 600% faster on AMD...



# DE·i·GO

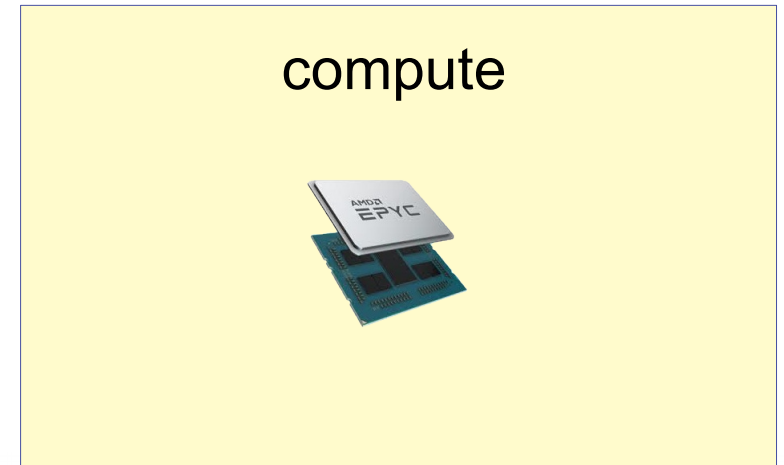
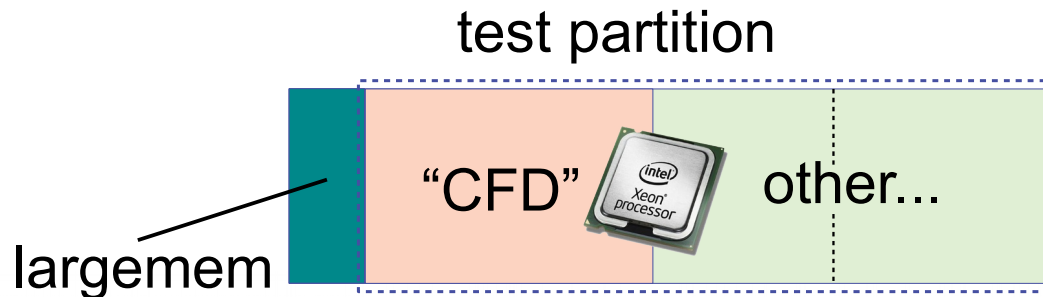
Work in Progress

Intel: 192 nodes, 12% cores

- Task-specific partitions
  - not per-user or per-unit
  - physics, MD
  - Intel-dependent code
  - largemem
- Overlap with low-priority test partition

AMD: 456 nodes, 88% cores

- General purpose compute partition
  - lots of cores, lots of users
  - **user memory and core limits**
- benefits from rebuilding your code



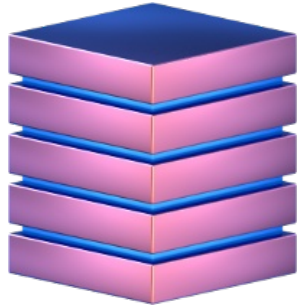


# DE·i·GO

## New /work

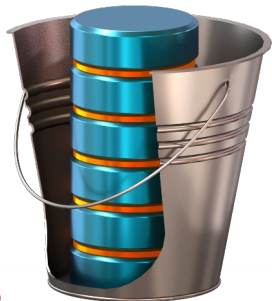
500TB SSD

10TB per unit



## Bucket

+6 PB



## Compute nodes



read and write



read-only

## Old /work

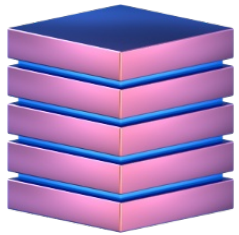


## Login nodes



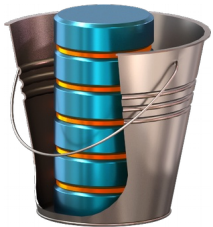
## New /work

500TB SSD  
10TB per unit



## Bucket

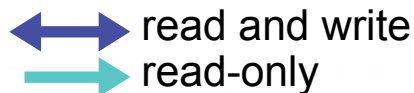
+6 PB



Compute nodes



Login nodes



## New workflow

- Similar to Saion:
  - /work is only scratch, **not storage**
  - 10TB per unit is a *hard limit*
  - read access to Bucket from compute nodes

1. read your input from bucket
  - Read directly or copy to /work beforehand
2. use /work for ongoing computation
3. copy results back to bucket
4. Clean up /work

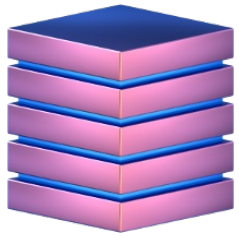




# DE·i·GO

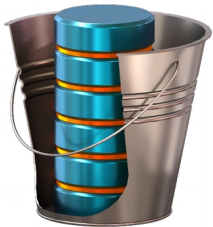
## New /work

500TB SSD  
10TB per unit



## Bucket

+6 PB



Compute nodes



## Old /work



- Will **disappear** during 2020
  - Soon out of warranty
  - Full, not expandable
- Will be available read-only
- You **must** copy data you need to Bucket
- The rest will be archived and effectively *unavailable* once shut down



Login nodes



Old /work

 read and write  
 read-only



## Provided software

- CentOS 8
- GCC 8 (or 9), AOCC
- BLIS, LibFLAME
- User Software (modules)
  - Popular open source modules will be rebuilt.
  - Other modules will run directly or through “sango” container, available as module (**best effort**)

## Your software

- For best results, rebuild with modern compilers, libraries
- Many will run OK unchanged
- Use “sango” container to run those that won’t:

```
module load bowtie  
myprog -o xyz
```



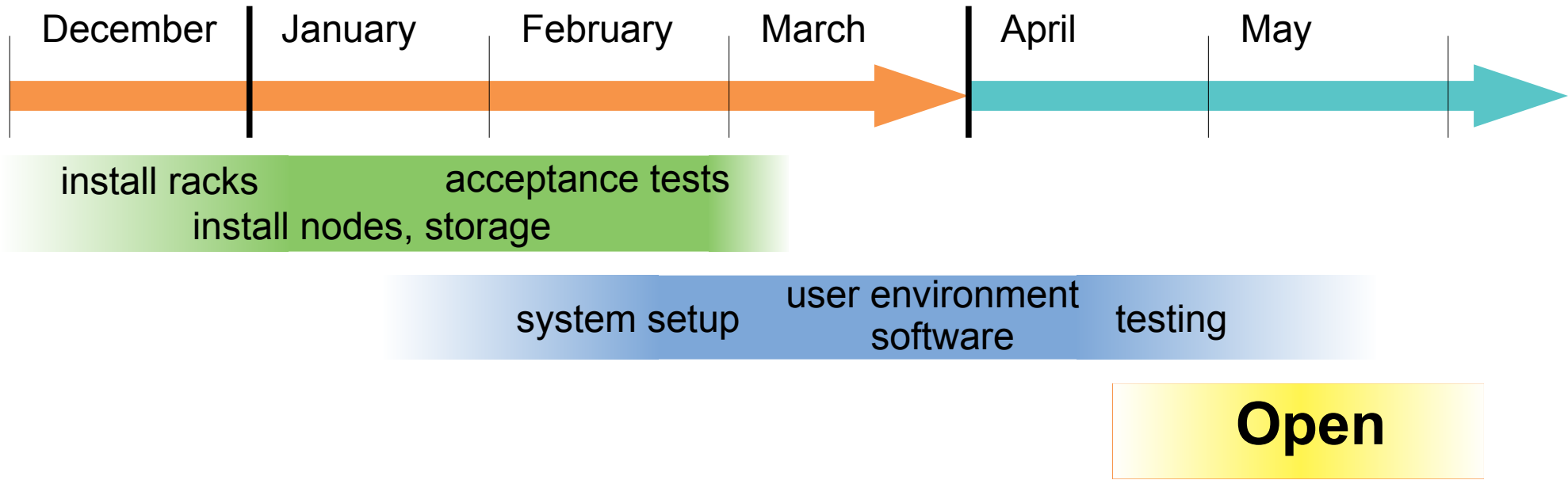
```
module load sango  
sango -m bowtie myprog -o xyz
```

**Preliminary**



# DE·i·GO

## Timeline

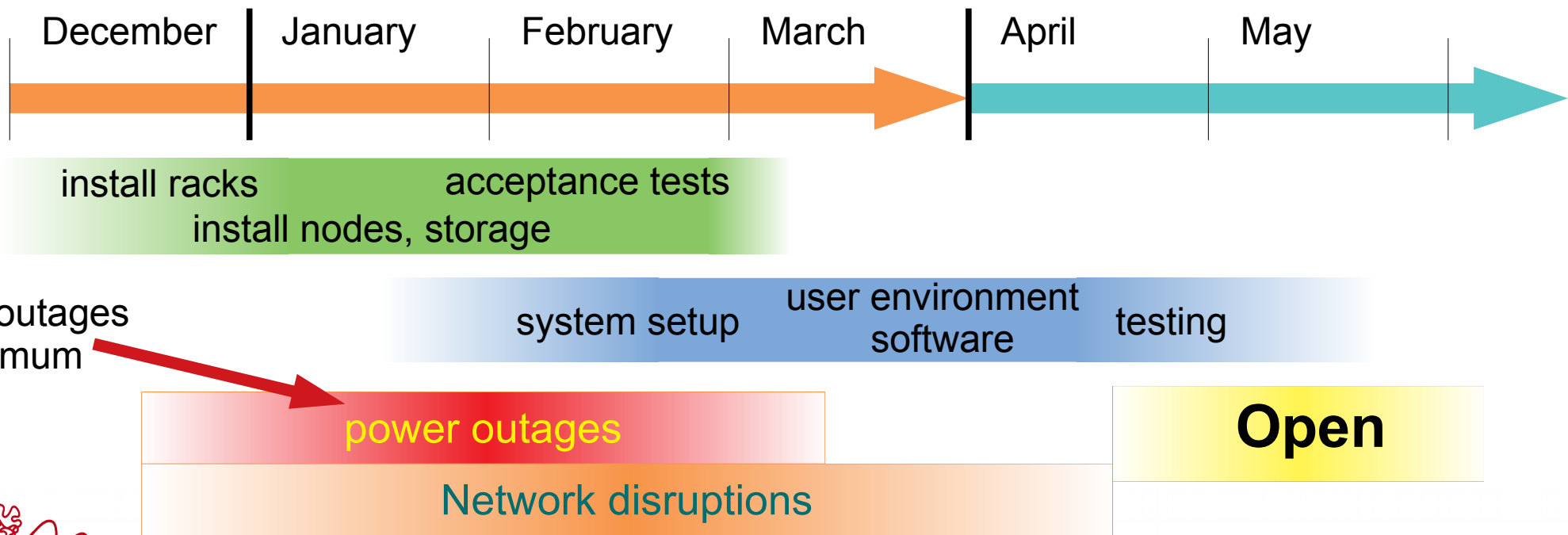


**Preliminary**



# DE·i·GO

## Timeline



# Students and RUAs storage problems

## Symptoms

- They get “out of space” errors when trying to save data on work or bucket
- The unit has plenty of space left
- Their files/directories have group “allstudents” or “allruas”

```
$ ls -l  
drwxr-xr-x 2 jan-moren allstudents 4096 14 nov 15.41 jan-moren/  
$
```

# Students and RUAs storage problems

## Groups

- All users have 1 *primary* group, multiple secondary groups.
  - files get primary group by default
- Most OIST members have their unit or section as primary group:

```
$ groups  
scicomsec oist scicom-data ...
```

- But students do rotations, and may work with multiple units.

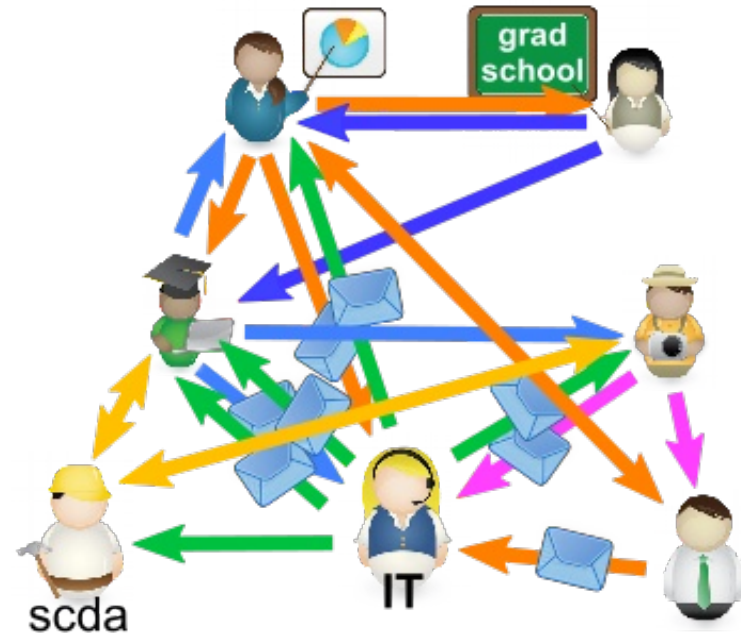
## Problem

- Only IT can change primary groups.
- PI's have the legal responsibility for who is a member, and need to give approval
  - PI have authority but not means
  - IT has means but not authority
- Can only have one primary group, so for collaborations you need ugly, brittle hacks.

# Students and RUAs storage problems

## Problem

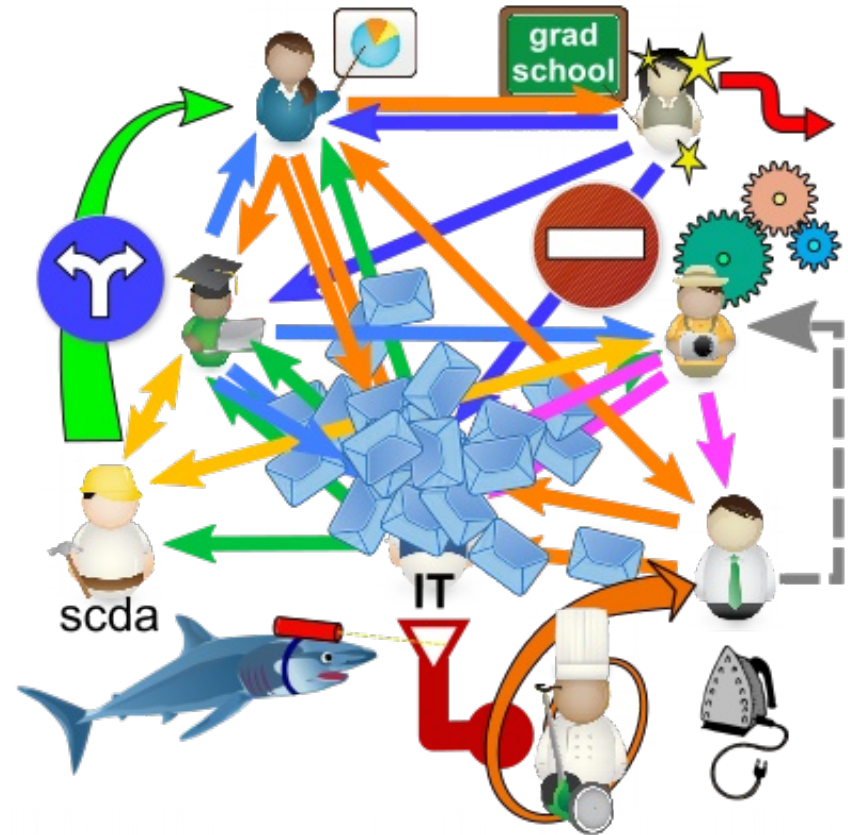
- Only IT can change primary groups.
- PI's have the legal responsibility for who is a member, and need to give approval
  - PI have authority but not means
  - IT has means but not authority
- Can only have one primary group, so for collaborations you need ugly, brittle hacks.



# Students and RUAs storage problems

## Problem

- Only IT can change primary groups.
- PI's have the legal responsibility for who is a member, and need to give approval
  - PI have authority but not means
  - IT has means but not authority
- Can only have one primary group, so for collaborations you need ugly, brittle hacks.





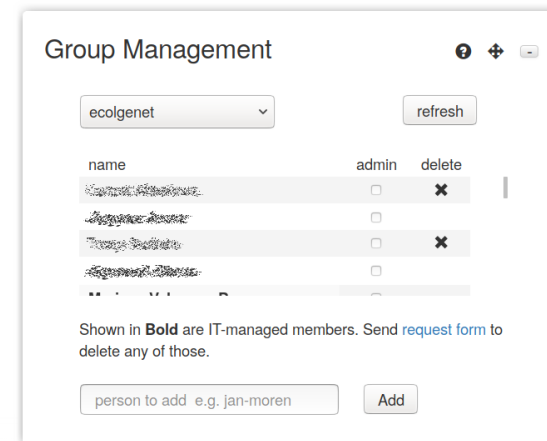
# Students and RUAs storage problems

## Problem

- Only IT can change primary groups.
- PI's have the legal responsibility for who is a member, and need to give approval
  - PI have authority but not means
  - IT has means but not authority
- Can only have one primary group, so for collaborations you need ugly, brittle hacks.

## Solution

1. Put students in “allstudents”
2. PIs can use “grouper” to add anybody as a *secondary* group.
  - can delegate to unit members
  - *can't* remove regular members



Group Management

ecolgenet refresh

name	admin	delete
<b>Jan-Moren</b>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<b>Jan-Moren</b>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<b>Jan-Moren</b>	<input type="checkbox"/>	<input type="checkbox"/>
...	...	...

Shown in **Bold** are IT-managed members. Send [request form](#) to delete any of those.

person to add e.g. jan-moren Add

# Students and RUAs storage problems

## Problem

- Only IT can change primary groups.
- PI's have the legal responsibility for who is a member, and need to give approval
  - PI have authority but not means
  - IT has means but not authority
- Can only have one primary group, so for collaborations you need ugly, brittle hacks.

## Solution

1. Put students in “allstudents”
  2. PIs can use “grouper” to add anybody as a *secondary* group.
  3. Set unit directories to override the users' primary group (setguid)
    - new files get unit group, not “allstudents”
- PIs now do all changes themselves
  - Students (or anybody) can get access to multiple units

# Students and RUAs storage problems

## Solution

1. Put students in “allstudents”
  2. PIs can use “grouper” to add anybody as a *secondary* group.
  3. Set unit directories to override the users’ primary group (setgid)
    - new files get unit group, not “allstudents”
- PIs now do all changes themselves
  - Students (or anybody) can get access to multiple units

## Problem

- group ownership is sometimes overridden, “setgid” removed:
  - ex. “rsync -a”
  - “cp -a”, “cp -p”
  - using SMB (unclear when)

```
$ ls -l  
...jan-moren allstudents ... myfolder/
```

- “allstudents” has no quota, so writing files fails with out of space error.

# Students and RUAs storage problems

## Solution

1. Put students in “allstudents”
  2. PIs can use “grouper” to add anybody as a *secondary* group.
  3. Set unit directories to override the users’ primary group (setgid)  
→ new files get unit group, not “allstudents”
- PIs now do all changes themselves
  - Students (or anybody) can get access to multiple units

## Solution

- “fixdirs” script fixes both group and setgid bit:

```
$ fixdirs myfolder myuni  
$ ls -l  
... jan-moren myuni ... myfolder/
```

- Avoid setting group, permissions when moving data:

```
$ rsync -a --no-group --no-perms ...  
$ cp -a --no-preserve=mode,ownership
```

# New List, New Members

## Advanced HPC User Group

- “advanced” - reasonable, willing to learn, HPC users
  - No PIs, no section leaders - your boss (or mine) won't see you asking questions.
- The mailing list is a simple, direct way to contact us and each other.
- We will move the mailing list to new system. This might interrupt things. Tell us if there's a problem.
- **We want new members.** If you know somebody you think would benefit, tell me.