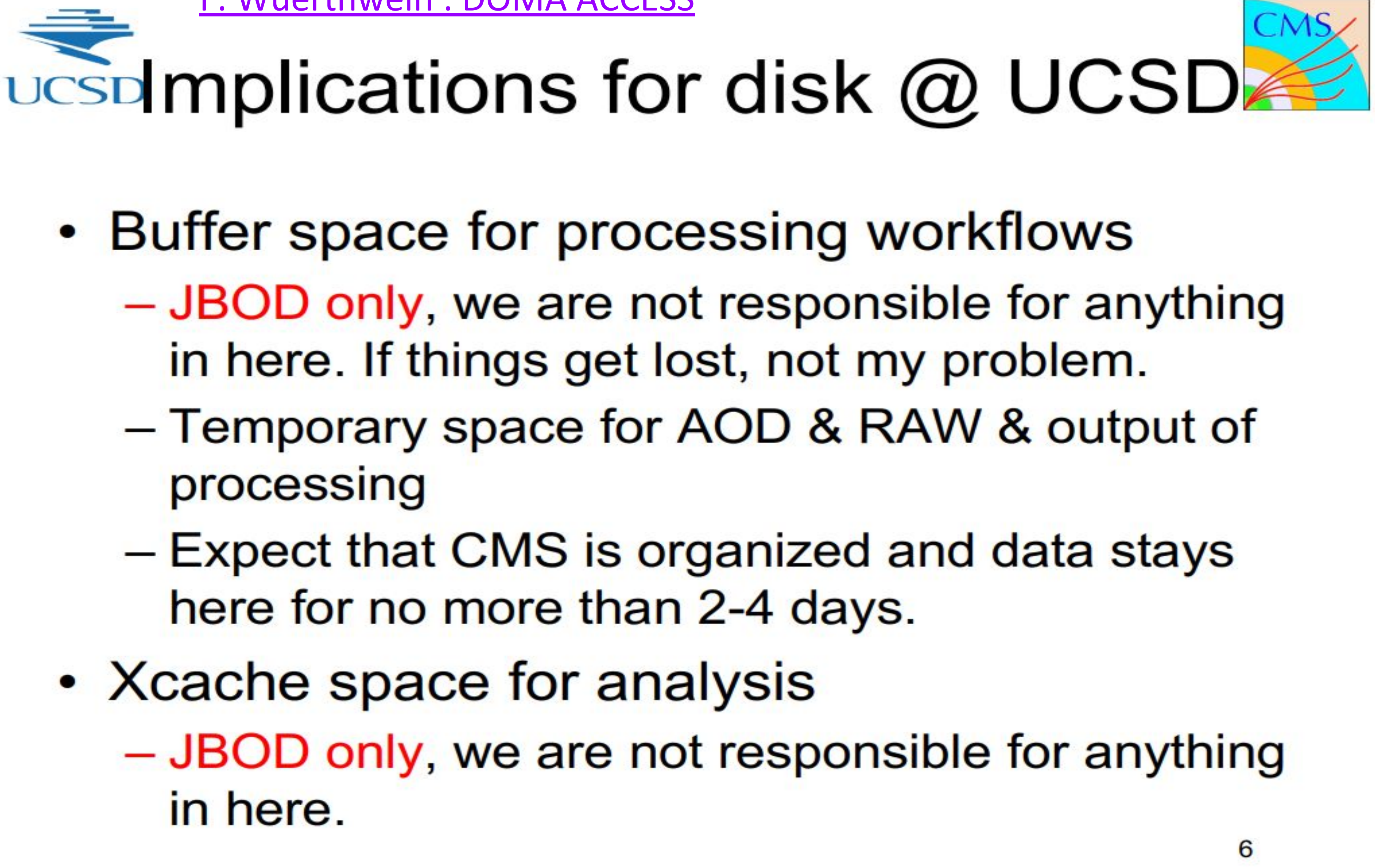
Laboratoire d'Annecy de Physique des Particules

Total cost of ownership : T2 feedback

S. Jezequel summarising UCSD,Manchester,GROF

Wednesday 28 October 2020

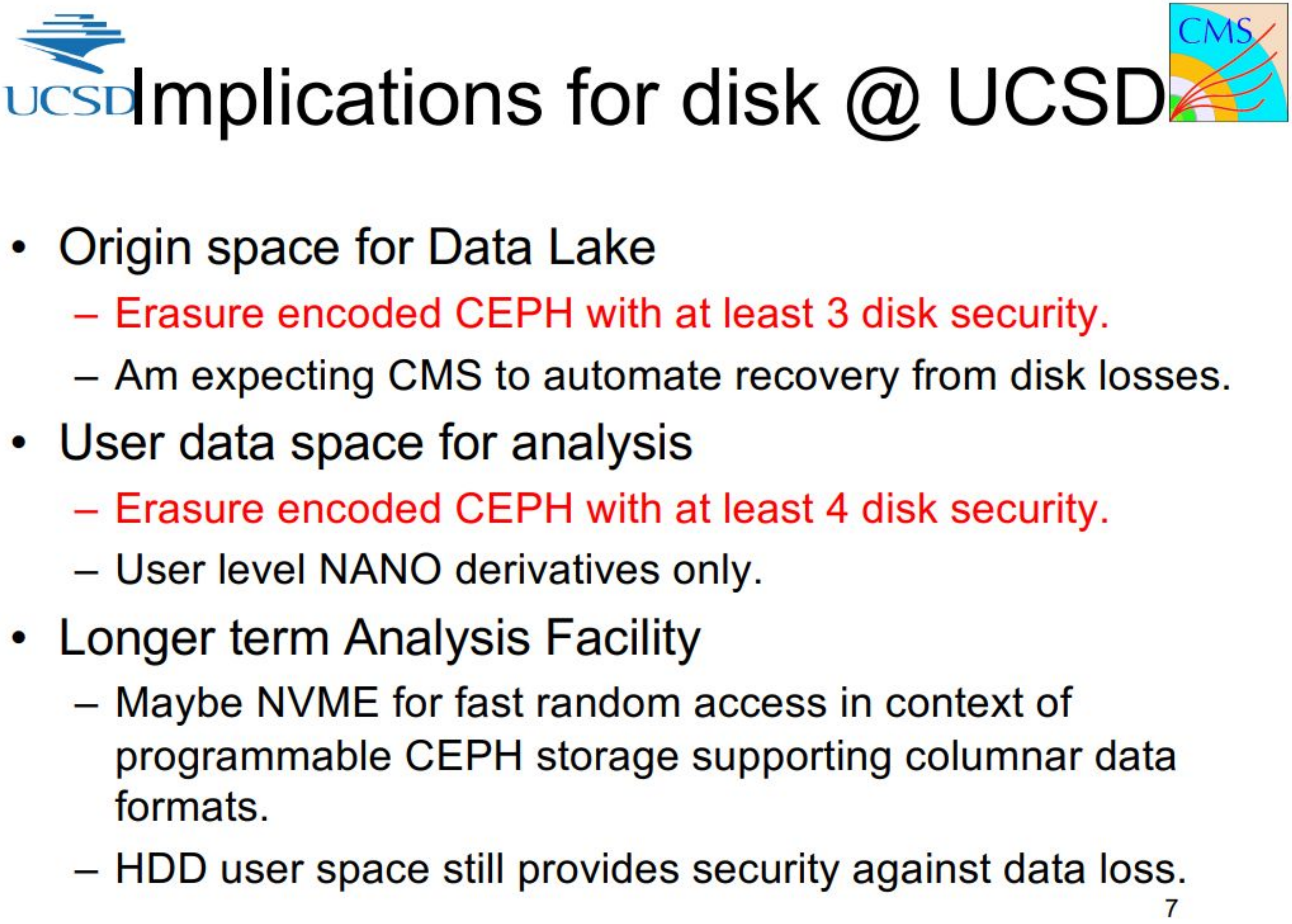
F. Wuerthwein : DOMA ACCESS



Implications for disk @ UCSD

- Buffer space for processing workflows
  - JBOD only, we are not responsible for anything in here. If things get lost, not my problem.
  - Temporary space for AOD & RAW & output of processing
  - Expect that CMS is organized and data stays here for no more than 2-4 days.
- Xcache space for analysis
  - JBOD only, we are not responsible for anything in here.

USCMS T2 : UCSD

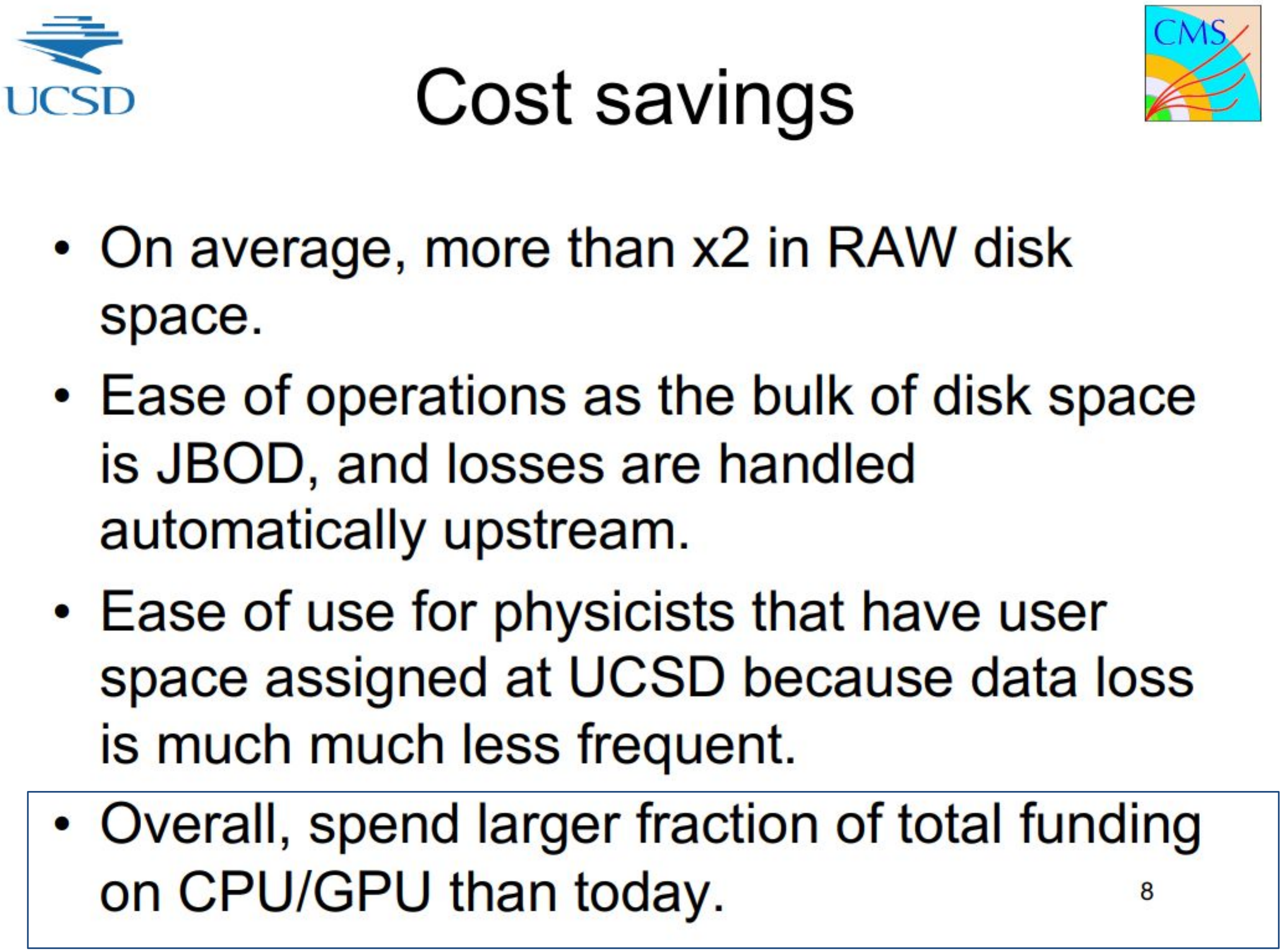


Implications for disk @ UCSD

- Origin space for Data Lake
  - Erasure encoded CEPH with at least 3 disk security.
  - Am expecting CMS to automate recovery from disk losses.
- User data space for analysis
  - Erasure encoded CEPH with at least 4 disk security.
  - User level NANO derivatives only.
- Longer term Analysis Facility
  - Maybe NVME for fast random access in context of programmable CEPH storage supporting columnar data formats.
  - HDD user space still provides security against data loss.

7

USCMS T2 : UCSD



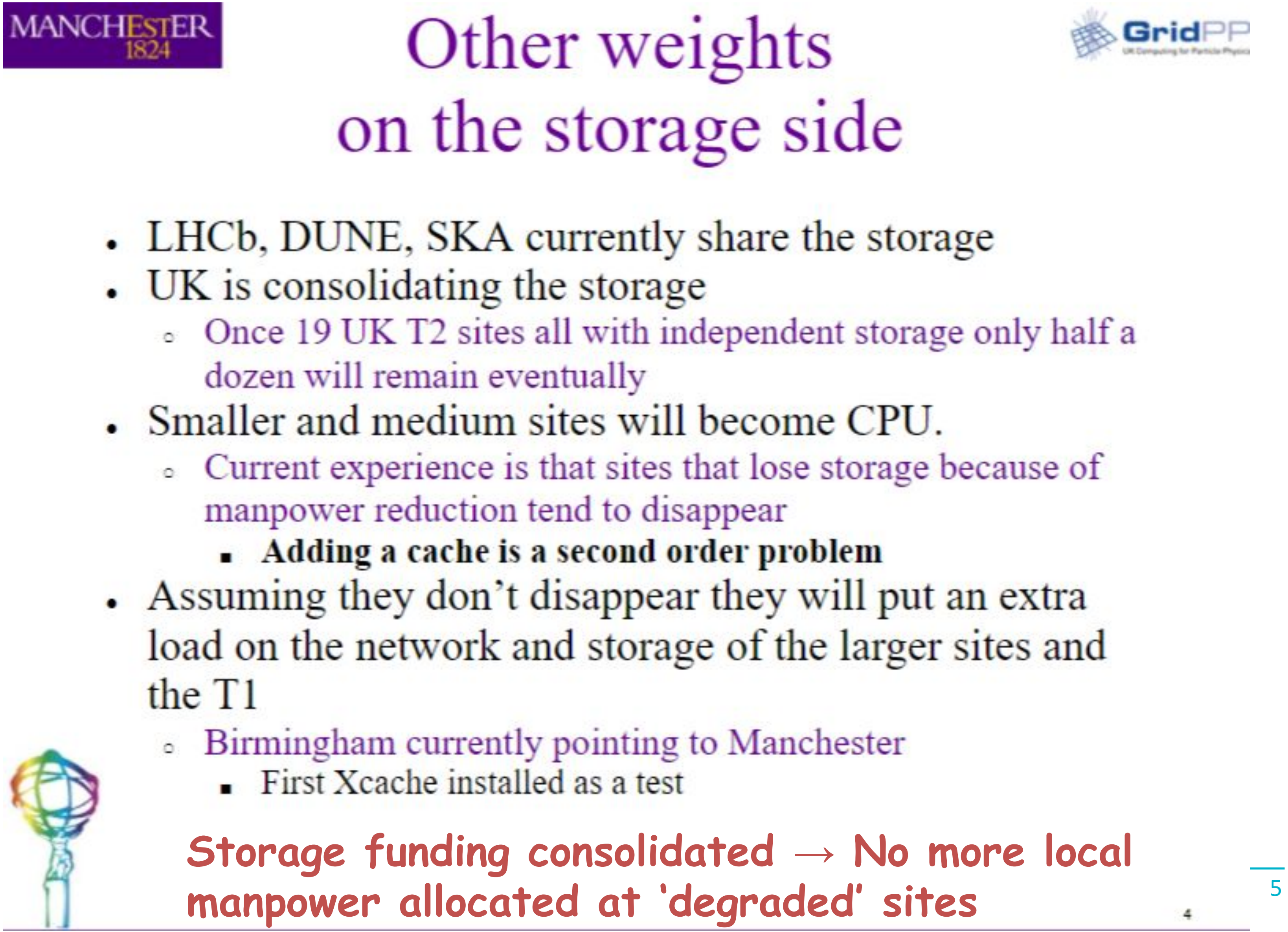
Cost savings

- On average, more than x2 in RAW disk space.
- Ease of operations as the bulk of disk space is JBOD, and losses are handled automatically upstream.
- Ease of use for physicists that have user space assigned at UCSD because data loss is much much less frequent.
- Overall, spend larger fraction of total funding on CPU/GPU than today.

8

ATLAS T2 : Manchester

A. Forti : DOMA ACCESS



MANCHESTER 1824

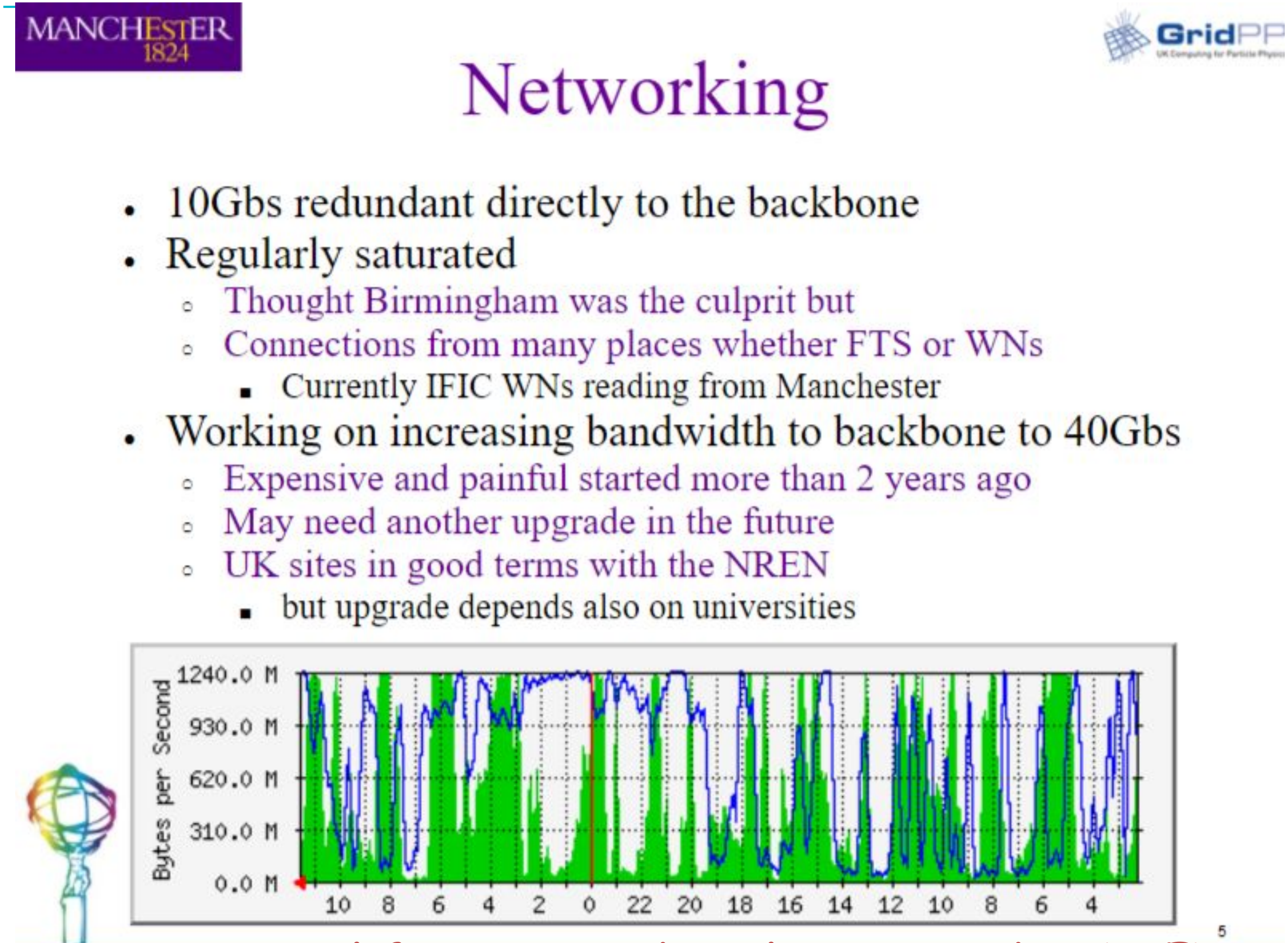
GridPP

Other weights on the storage side

- LHCb, DUNE, SKA currently share the storage
- UK is consolidating the storage
  - Once 19 UK T2 sites all with independent storage only half a dozen will remain eventually
- Smaller and medium sites will become CPU.
  - Current experience is that sites that lose storage because of manpower reduction tend to disappear
    - **Adding a cache is a second order problem**
- Assuming they don't disappear they will put an extra load on the network and storage of the larger sites and the T1
  - Birmingham currently pointing to Manchester
    - First Xcache installed as a test

Storage funding consolidated → No more local manpower allocated at 'degraded' sites

ATLAS T2 : Manchester

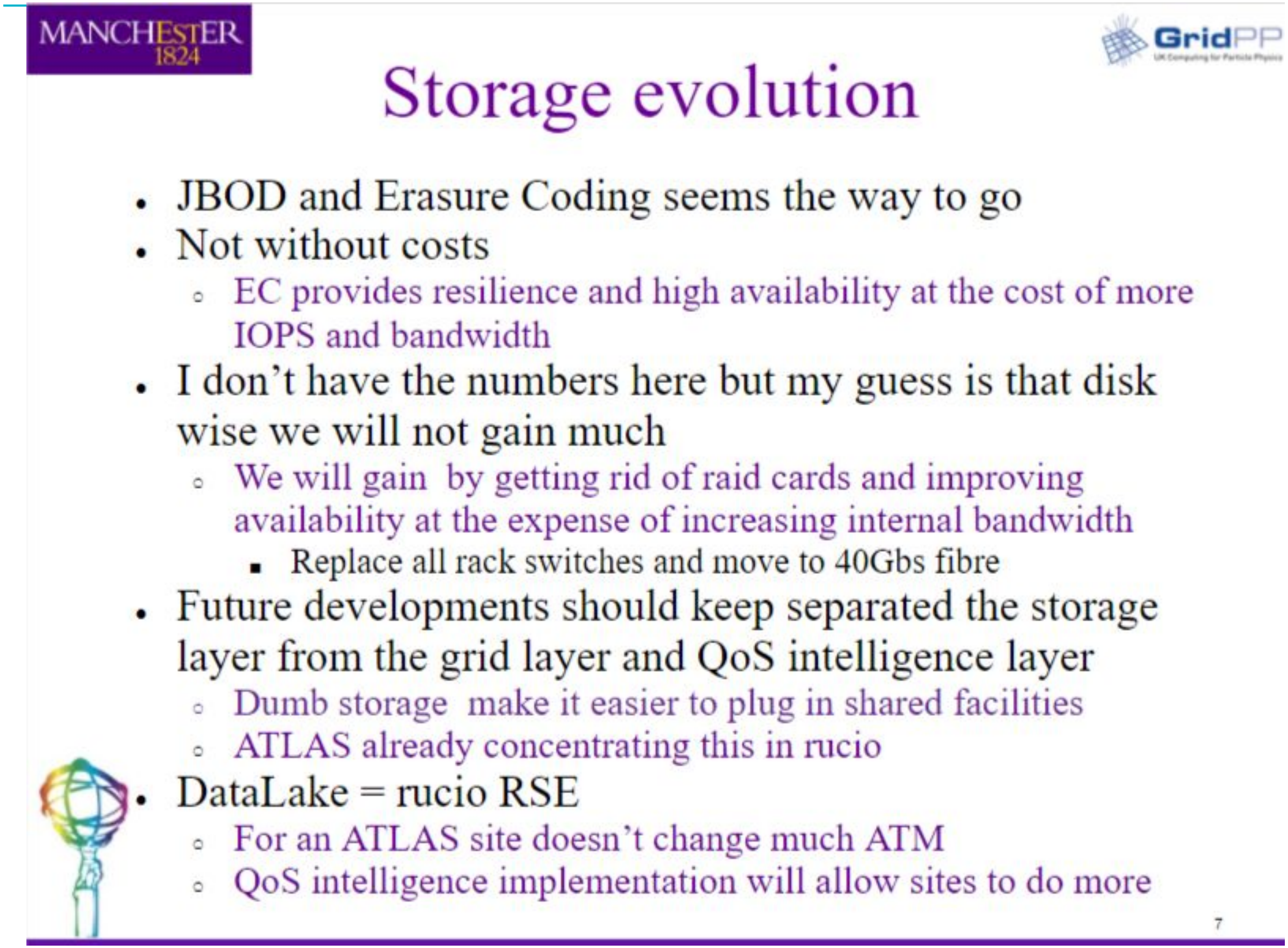


Networking

- 10Gbs redundant directly to the backbone
- Regularly saturated
  - Thought Birmingham was the culprit but
  - Connections from many places whether FTS or WNs
    - Currently IFIC WNs reading from Manchester
- Working on increasing bandwidth to backbone to 40Gbs
  - Expensive and painful started more than 2 years ago
  - May need another upgrade in the future
  - UK sites in good terms with the NREN
    - but upgrade depends also on universities

Critical for T2s not directly connected to NREN

ATLAS T2 : Manchester



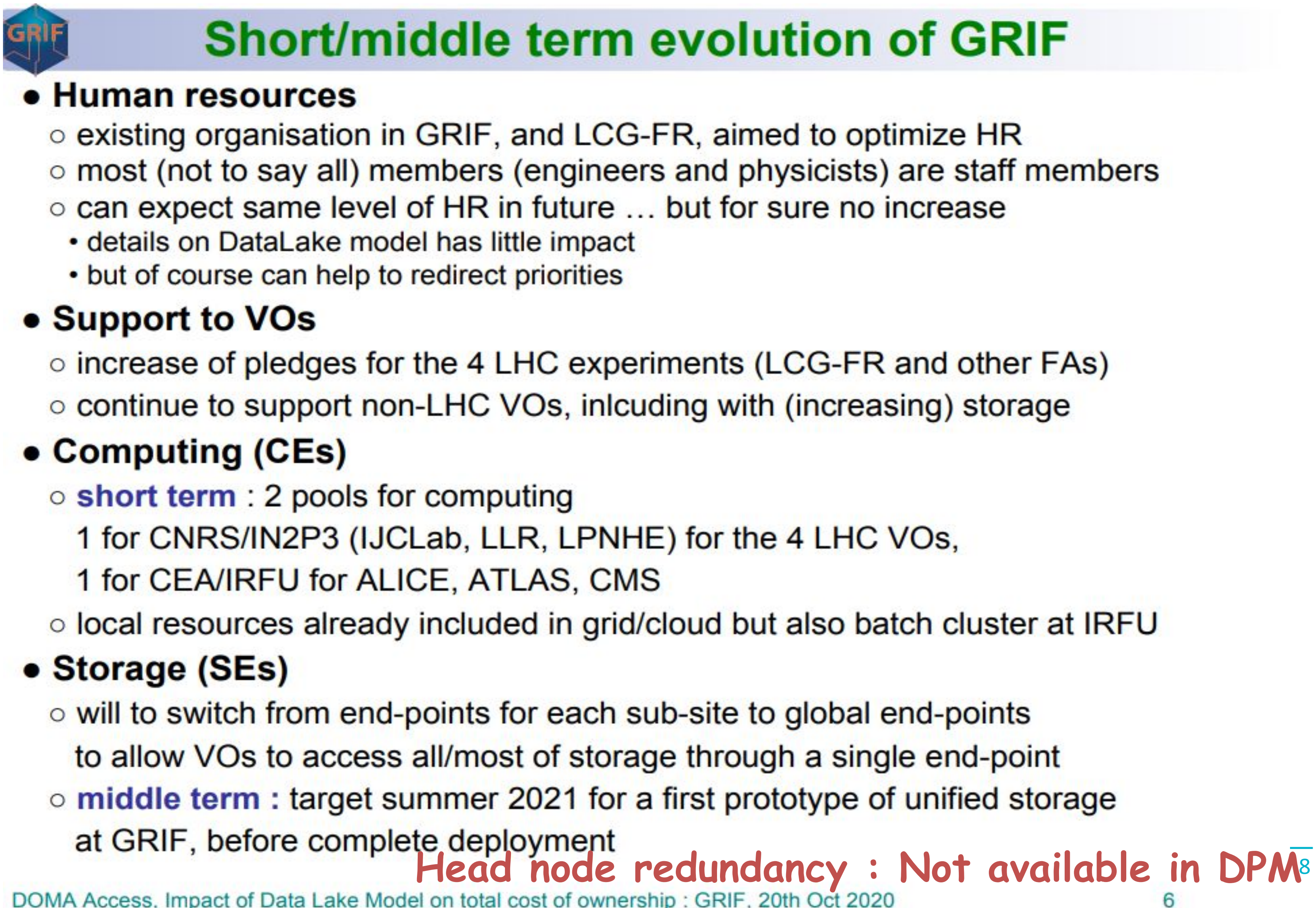
Storage evolution

- JBOD and Erasure Coding seems the way to go
- Not without costs
  - EC provides resilience and high availability at the cost of more IOPS and bandwidth
- I don't have the numbers here but my guess is that disk wise we will not gain much
  - We will gain by getting rid of raid cards and improving availability at the expense of increasing internal bandwidth
    - Replace all rack switches and move to 40Gbs fibre
- Future developments should keep separated the storage layer from the grid layer and QoS intelligence layer
  - Dumb storage make it easier to plug in shared facilities
  - ATLAS already concentrating this in rucio
- DataLake = rucio RSE
  - For an ATLAS site doesn't change much ATM
  - QoS intelligence implementation will allow sites to do more

7

ATLAS T2 : GRIF (Paris region)

F. Derue : DOMA ACCESS



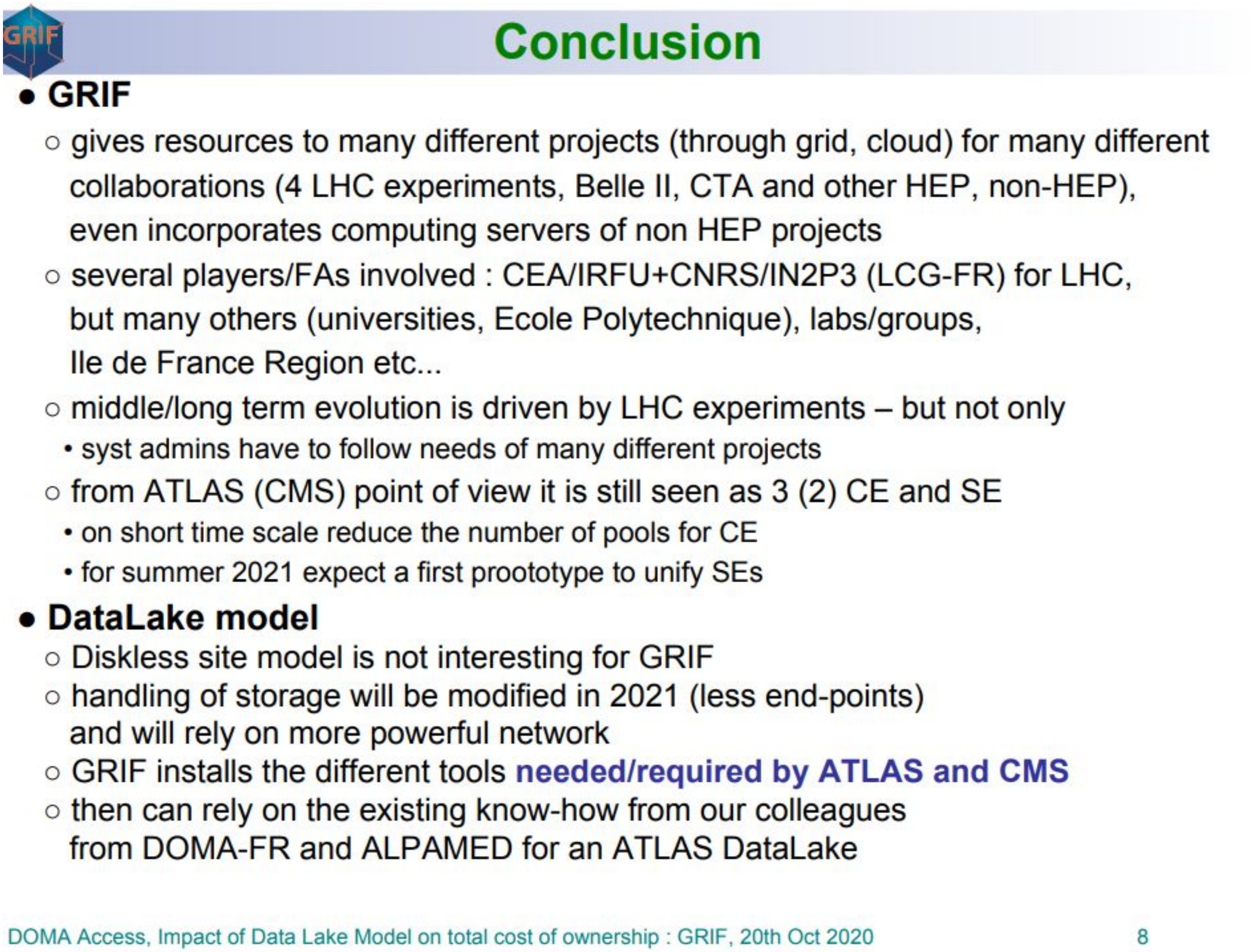
Short/middle term evolution of GRIF

- **Human resources**
  - existing organisation in GRIF, and LCG-FR, aimed to optimize HR
  - most (not to say all) members (engineers and physicists) are staff members
  - can expect same level of HR in future … but for sure no increase
    - details on DataLake model has little impact
    - but of course can help to redirect priorities
- **Support to VOs**
  - increase of pledges for the 4 LHC experiments (LCG-FR and other FAs)
  - continue to support non-LHC VOs, inlcuding with (increasing) storage
- **Computing (CEs)**
  - **short term** : 2 pools for computing
    1 for CNRS/IN2P3 (IJCLab, LLR, LPNHE) for the 4 LHC VOs,
    1 for CEA/IRFU for ALICE, ATLAS, CMS
  - local resources already included in grid/cloud but also batch cluster at IRFU
- **Storage (SEs)**
  - will to switch from end-points for each sub-site to global end-points to allow VOs to access all/most of storage through a single end-point
  - **middle term :** target summer 2021 for a first prototype of unified storage at GRIF, before complete deployment

DOMA Access. Impact of Data Lake Model on total cost of ownership : GRIF, 20th Oct 2020 6

Head node redundancy : Not available in DPM

ATLAS T2 : GRIF (Paris region)



Conclusion

- **GRIF**
  - ○ gives resources to many different projects (through grid, cloud) for many different collaborations (4 LHC experiments, Belle II, CTA and other HEP, non-HEP), even incorporates computing servers of non HEP projects
  - ○ several players/FAs involved : CEA/IRFU+CNRS/IN2P3 (LCG-FR) for LHC, but many others (universities, Ecole Polytechnique), labs/groups, Ile de France Region etc...
  - ○ middle/long term evolution is driven by LHC experiments – but not only
    - syst admins have to follow needs of many different projects
  - ○ from ATLAS (CMS) point of view it is still seen as 3 (2) CE and SE
    - on short time scale reduce the number of pools for CE
    - for summer 2021 expect a first proototype to unify SEs
- **DataLake model**
  - ○ Diskless site model is not interesting for GRIF
  - ○ handling of storage will be modified in 2021 (less end-points) and will rely on more powerful network
  - ○ GRIF installs the different tools **needed/required by ATLAS and CMS**
  - ○ then can rely on the existing know-how from our colleagues from DOMA-FR and ALPAMED for an ATLAS DataLake