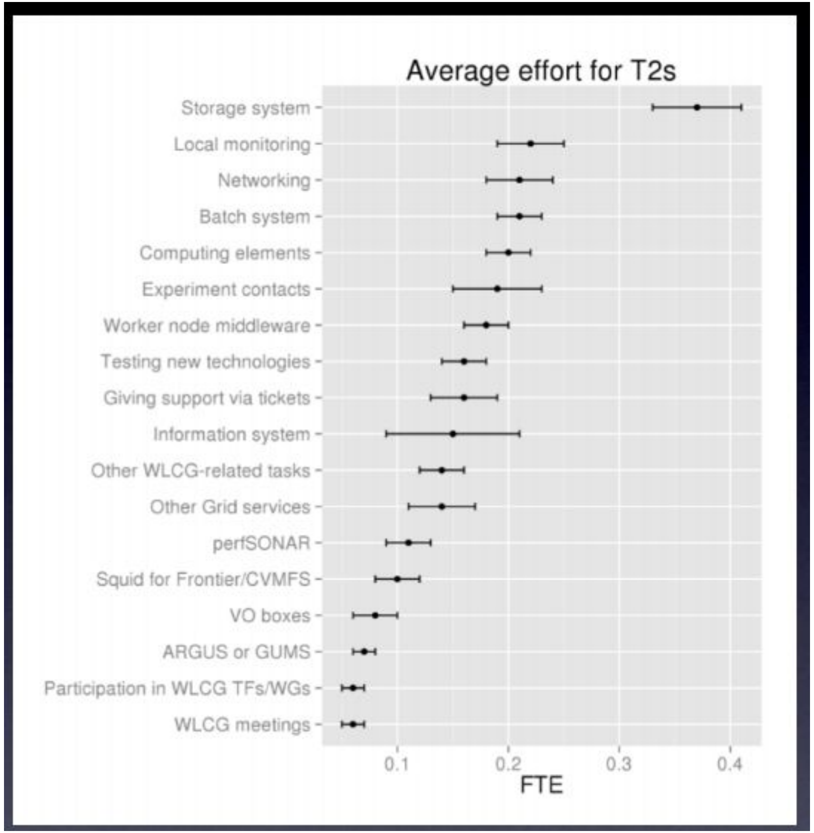
# Lightweight Sites

David Cameron
University of Oslo

Many slides taken from GDB May 2016 dedicated to the subject

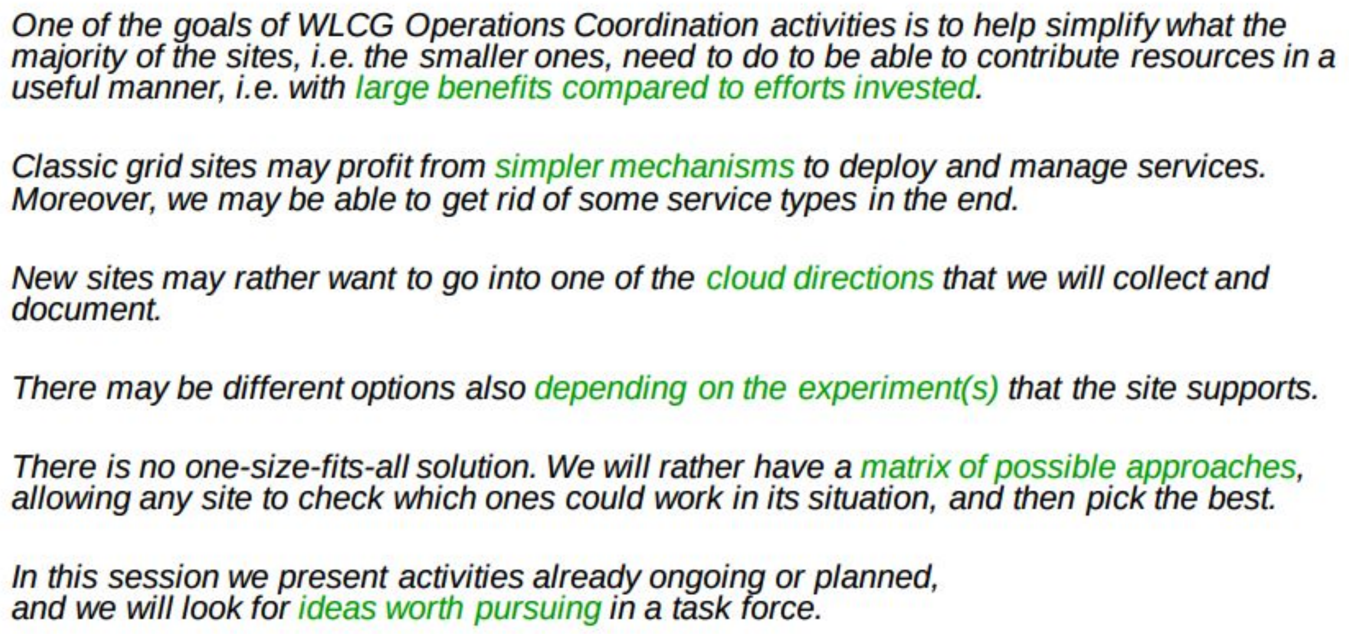
# Why?

A grid site costs a lot to run



# How?

As usual, many (competing?) solutions



*One of the goals of WLCG Operations Coordination activities is to help simplify what the majority of the sites, i.e. the smaller ones, need to do to be able to contribute resources in a useful manner, i.e. with large benefits compared to efforts invested.*

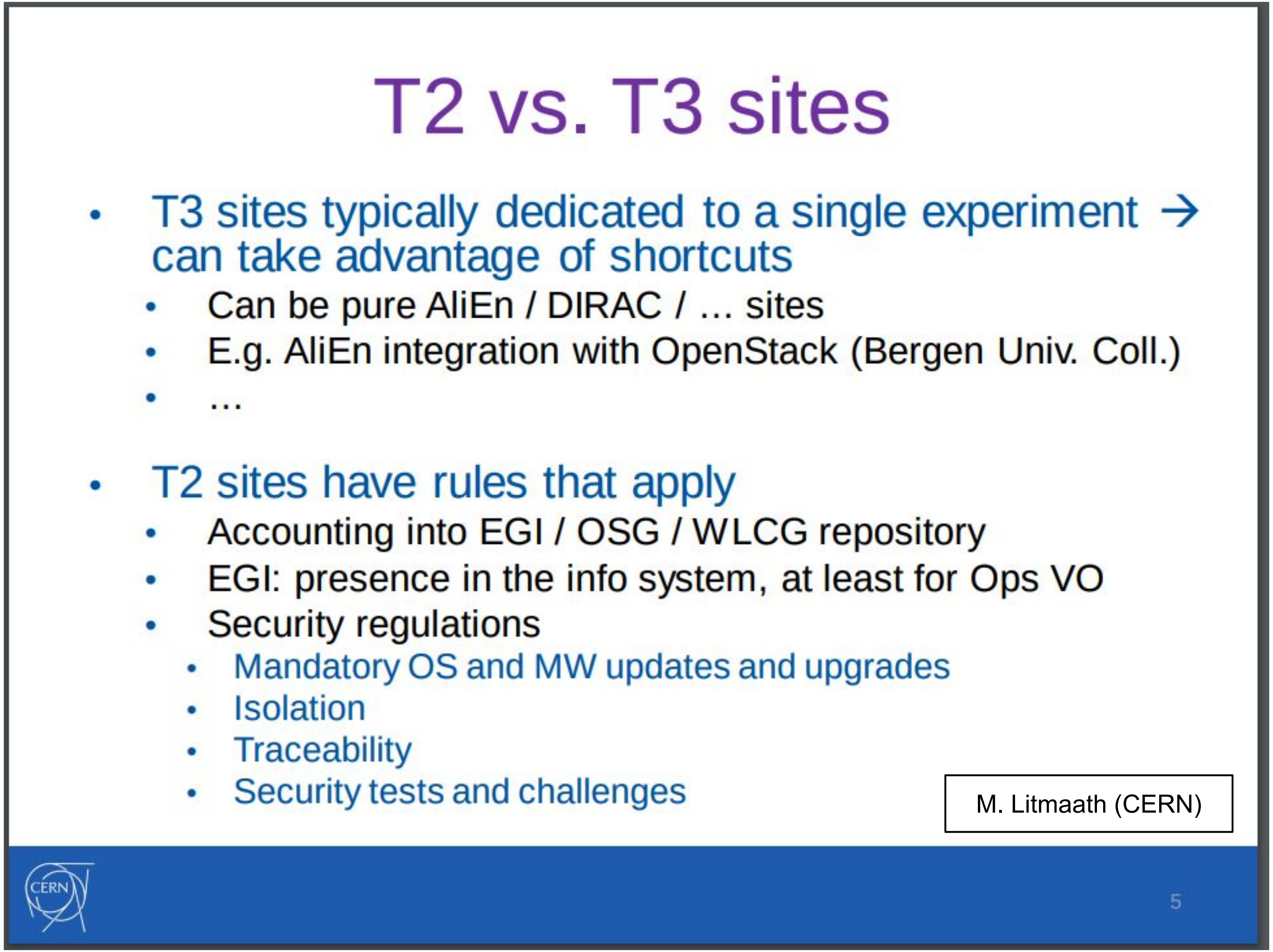
*Classic grid sites may profit from simpler mechanisms to deploy and manage services. Moreover, we may be able to get rid of some service types in the end.*

*New sites may rather want to go into one of the cloud directions that we will collect and document.*

*There may be different options also depending on the experiment(s) that the site supports.*

*There is no one-size-fits-all solution. We will rather have a matrix of possible approaches, allowing any site to check which ones could work in its situation, and then pick the best.*

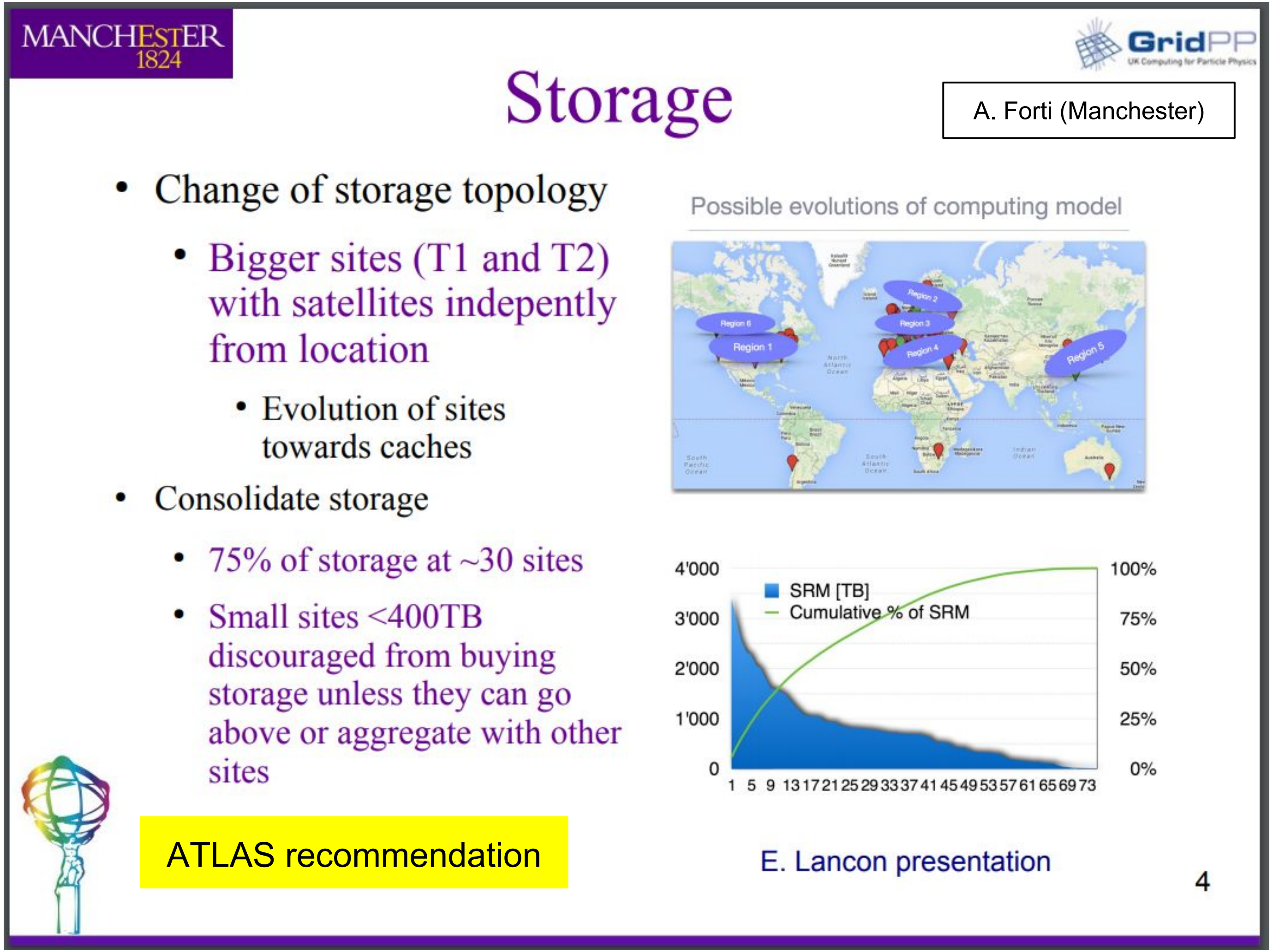
*In this session we present activities already ongoing or planned, and we will look for ideas worth pursuing in a task force.*



# T2 vs. T3 sites

- T3 sites typically dedicated to a single experiment → can take advantage of shortcuts
  - Can be pure AliEn / DIRAC / … sites
  - E.g. AliEn integration with OpenStack (Bergen Univ. Coll.)
  - …
- T2 sites have rules that apply
  - Accounting into EGI / OSG / WLCG repository
  - EGI: presence in the info system, at least for Ops VO
  - Security regulations
    - Mandatory OS and MW updates and upgrades
    - Isolation
    - Traceability
    - Security tests and challenges

M. Litmaath (CERN)



A. Forti (Manchester)

# Storage

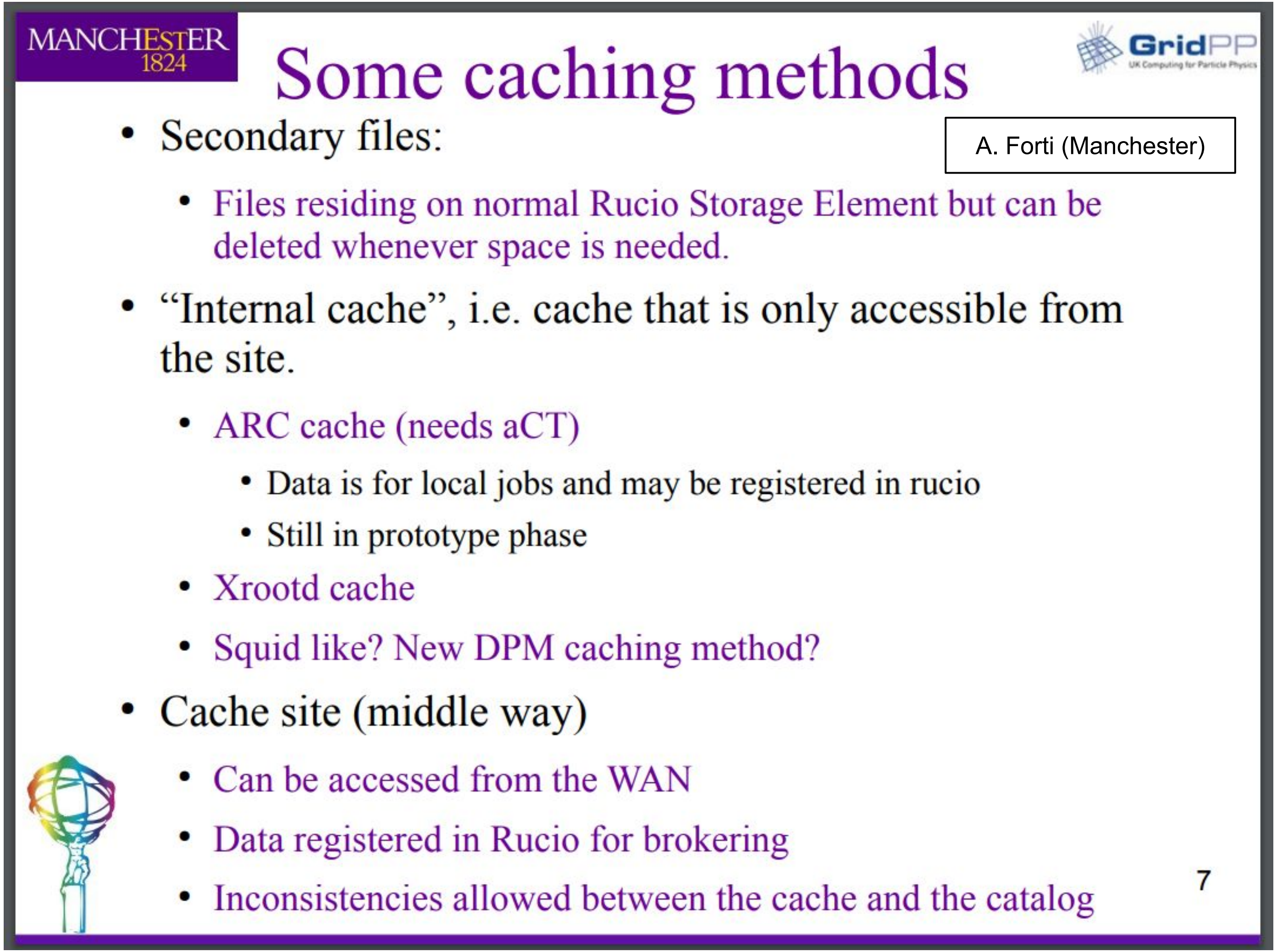
- Change of storage topology
  - Bigger sites (T1 and T2) with satellites indepently from location
    - Evolution of sites towards caches
- Consolidate storage
  - 75% of storage at ~30 sites
  - Small sites <400TB discouraged from buying storage unless they can go above or aggregate with other sites

Possible evolutions of computing model

E. Lancon presentation

ATLAS recommendation

4



MANCHESTER 1824

# Some caching methods

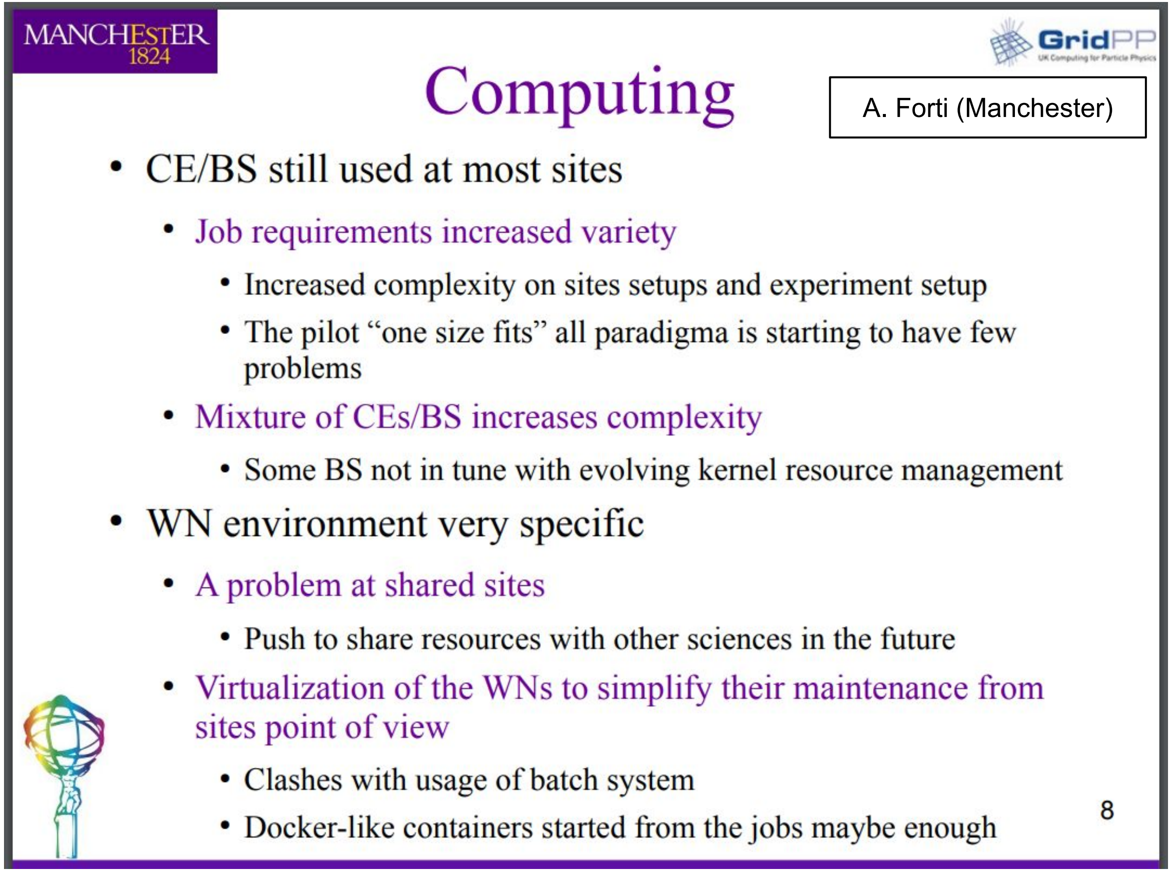
A. Forti (Manchester)

- Secondary files:
  - Files residing on normal Rucio Storage Element but can be deleted whenever space is needed.
- “Internal cache”, i.e. cache that is only accessible from the site.
  - ARC cache (needs aCT)
    - Data is for local jobs and may be registered in rucio
    - Still in prototype phase
  - Xrootd cache
  - Squid like? New DPM caching method?
- Cache site (middle way)
  - Can be accessed from the WAN
  - Data registered in Rucio for brokering
  - Inconsistencies allowed between the cache and the catalog

# Storage - the options

- No local storage - use a close big storage
  - Ok for places like UK with tightly-connected sites
  - What about a site in a remote area?
- Local cache
  - ARC cache
  - Xcache
  - Still requires running a service
    - ARC CE required for ARC cache
- Run jobs which are not data-intensive
  - Not so many (at least in ATLAS)

# CE/batch



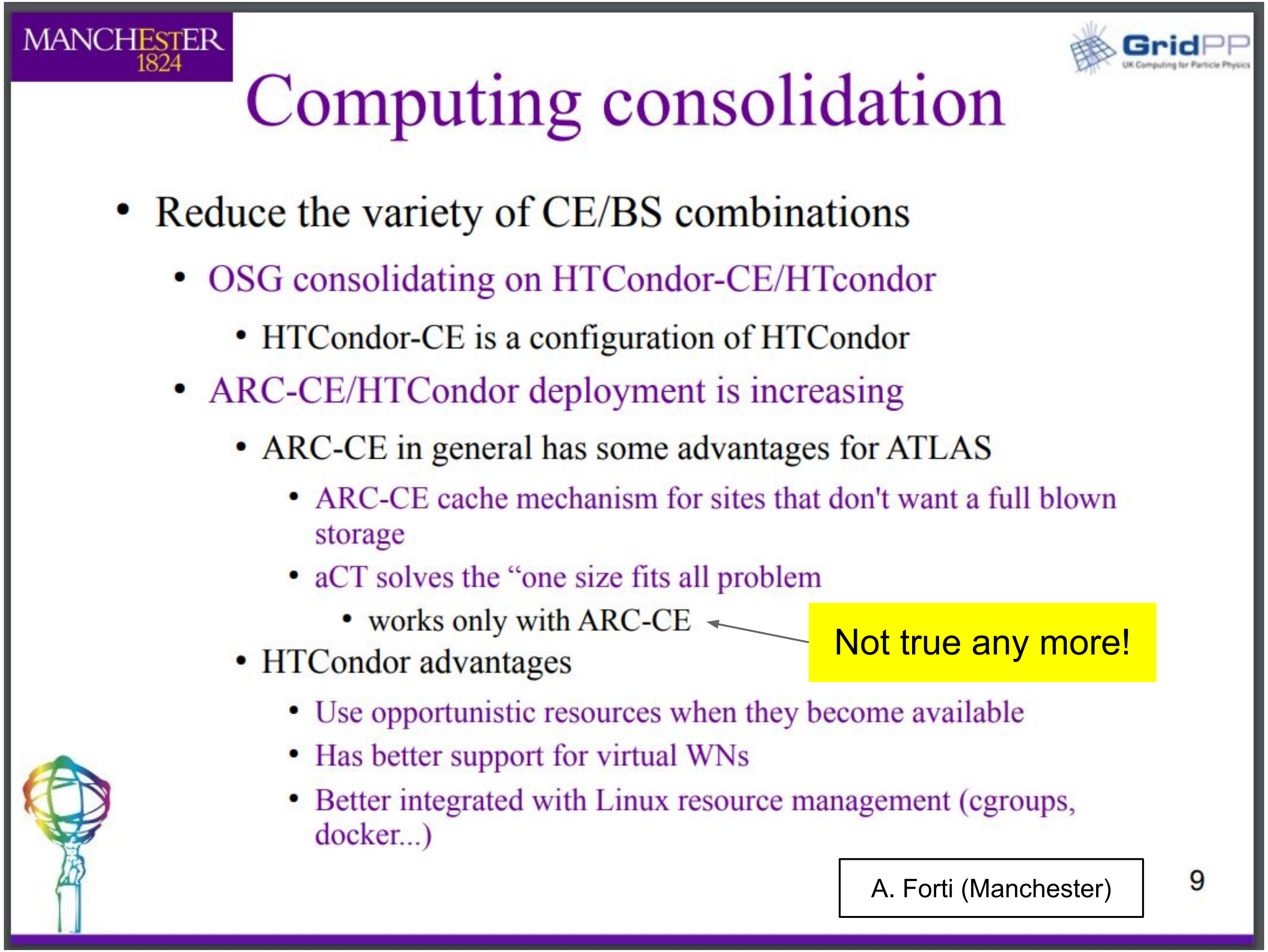
MANCHESTER 1824

GridPP UK Computing for Particle Physics

## Computing

A. Forti (Manchester)

- CE/BS still used at most sites
  - Job requirements increased variety
    - Increased complexity on sites setups and experiment setup
    - The pilot "one size fits" all paradigma is starting to have few problems
  - Mixture of CEs/BS increases complexity
    - Some BS not in tune with evolving kernel resource management
- WN environment very specific
  - A problem at shared sites
    - Push to share resources with other sciences in the future
  - Virtualization of the WNs to simplify their maintenance from sites point of view
    - Clashes with usage of batch system
    - Docker-like containers started from the jobs maybe enough

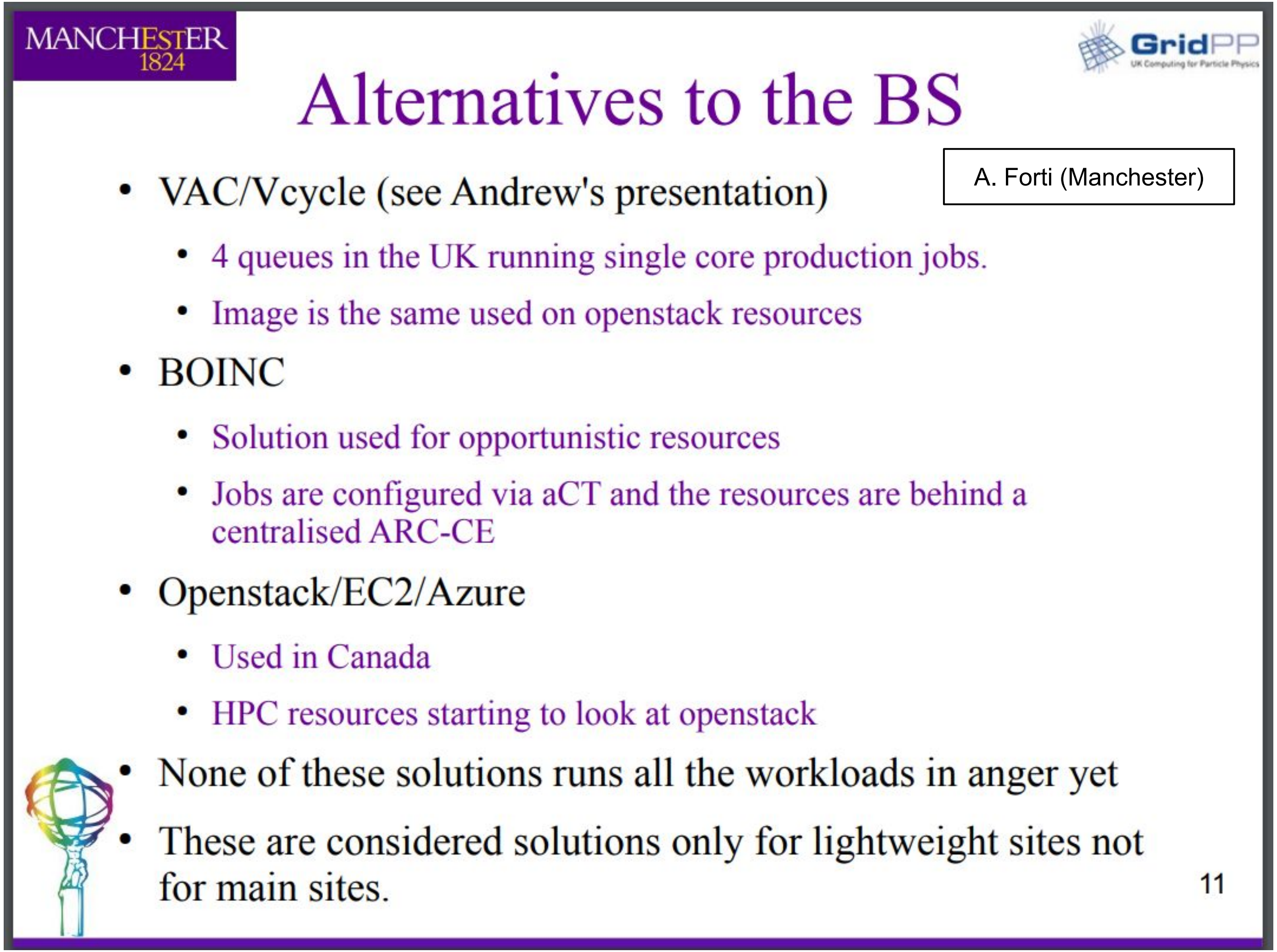


# Computing consolidation

- Reduce the variety of CE/BS combinations
  - OSG consolidating on HTCondor-CE/HTcondor
    - HTCondor-CE is a configuration of HTCondor
  - ARC-CE/HTCondor deployment is increasing
    - ARC-CE in general has some advantages for ATLAS
      - ARC-CE cache mechanism for sites that don't want a full blown storage
      - aCT solves the "one size fits all problem
        - works only with ARC-CE
    - HTCondor advantages
      - Use opportunistic resources when they become available
      - Has better support for virtual WNs
      - Better integrated with Linux resource management (cgroups, docker...)

Not true any more!

A. Forti (Manchester)



# Alternatives to the BS

A. Forti (Manchester)

- VAC/Vcycle (see Andrew's presentation)
  - 4 queues in the UK running single core production jobs.
  - Image is the same used on openstack resources
- BOINC
  - Solution used for opportunistic resources
  - Jobs are configured via aCT and the resources are behind a centralised ARC-CE
- Openstack/EC2/Azure
  - Used in Canada
  - HPC resources starting to look at openstack
- None of these solutions runs all the workloads in anger yet
- These are considered solutions only for lightweight sites not for main sites.

# Vacuum model

- A VM appears out of the vacuum with a pilot ready to run
- Vcycle can be used to instantiate VMs
- i.e. a site can set up an Openstack cloud and doesn't need grid services or batch system

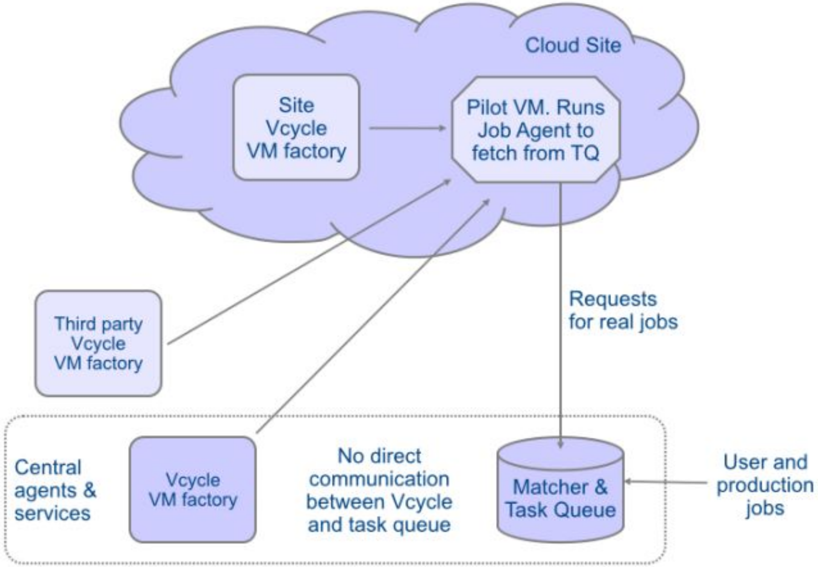
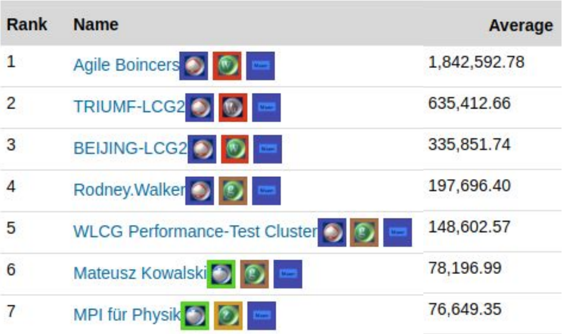

Cloud Site
Site Vcycle VM factory
Pilot VM. Runs Job Agent to fetch from TQ
Third party Vcycle VM factory
Requests for real jobs
Central agents & services
Vcycle VM factory
No direct communication between Vcycle and task queue
Matcher & Task Queue
User and production jobs

**Figure 2.** The Vacuum model as implemented by Vcycle

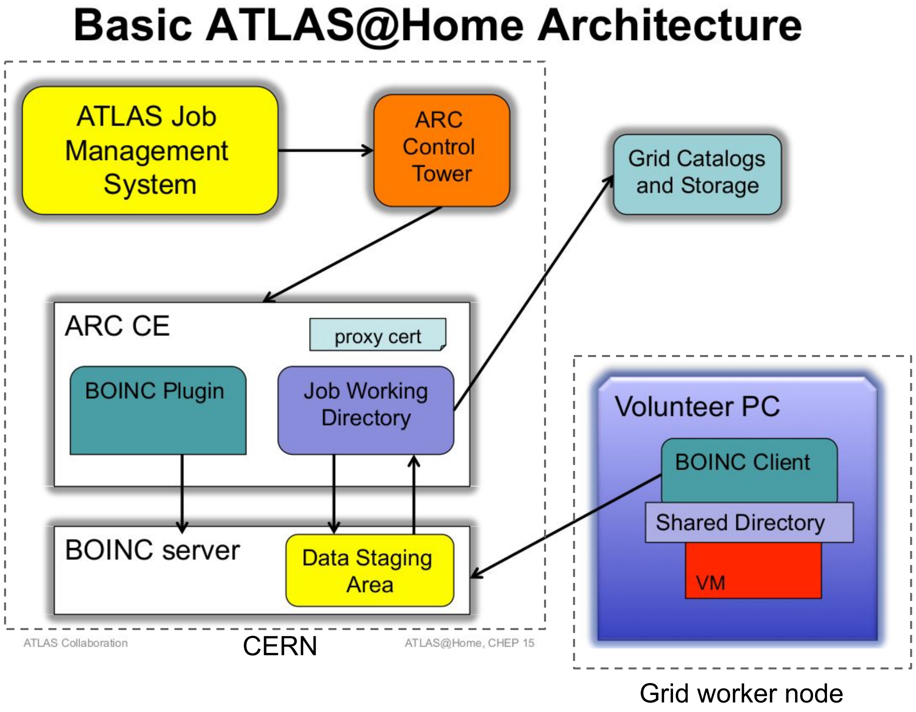
http://iopscience.iop.org/article/10.1088/1742-6596/664/2/022031/pdf

# BOINC

- Used by ATLAS, CMS & LHCb @home projects
  - ATLAS uses it for backfilling some grid sites
- Lightest possible way to run jobs
- Only BOINC client (+ virtualbox/singularity) needs to be installed
- No middleware environment
- No storage
- No problem killing jobs
- ATLAS recommendation for compute-only T3

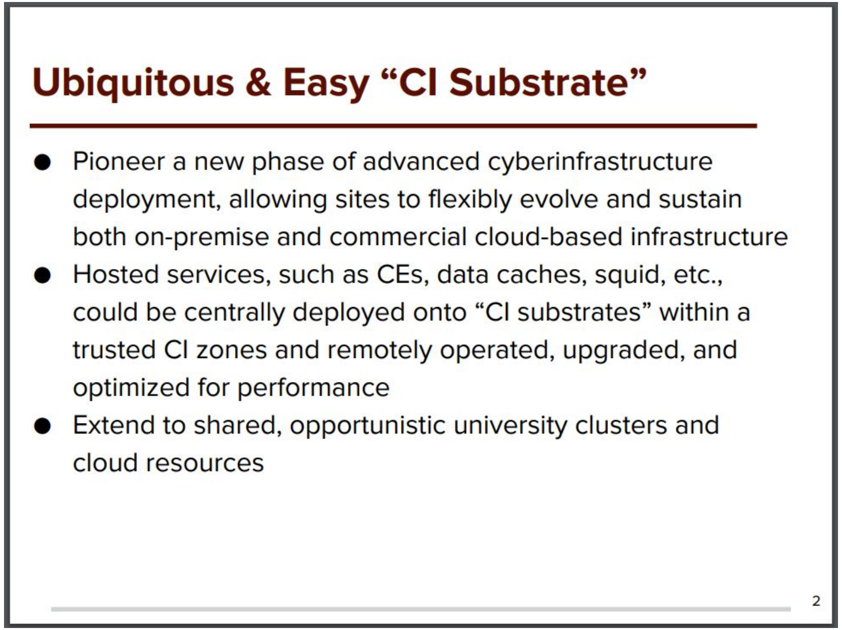


| Rank | Name | Average |
|---|---|---|
| 1 | Agile Boincers | 1,842,592.78 |
| 2 | TRIUMF-LCG2 | 635,412.66 |
| 3 | BEIJING-LCG2 | 335,851.74 |
| 4 | Rodney.Walker | 197,696.40 |
| 5 | WLCG Performance-Test Cluster | 148,602.57 |
| 6 | Mateusz Kowalski | 78,196.99 |
| 7 | MPI für Physik | 76,649.35 |



# Alternative approach - central management

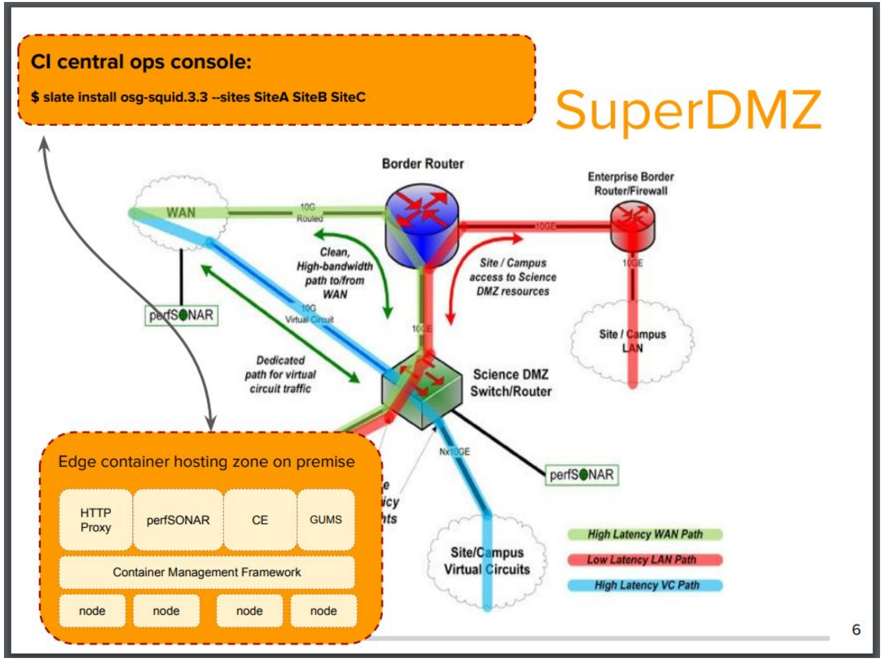
OSG (US) approach



## Ubiquitous & Easy “CI Substrate”

- Pioneer a new phase of advanced cyberinfrastructure deployment, allowing sites to flexibly evolve and sustain both on-premise and commercial cloud-based infrastructure
- Hosted services, such as CEs, data caches, squid, etc., could be centrally deployed onto “CI substrates” within a trusted CI zones and remotely operated, upgraded, and optimized for performance
- Extend to shared, opportunistic university clusters and cloud resources

2



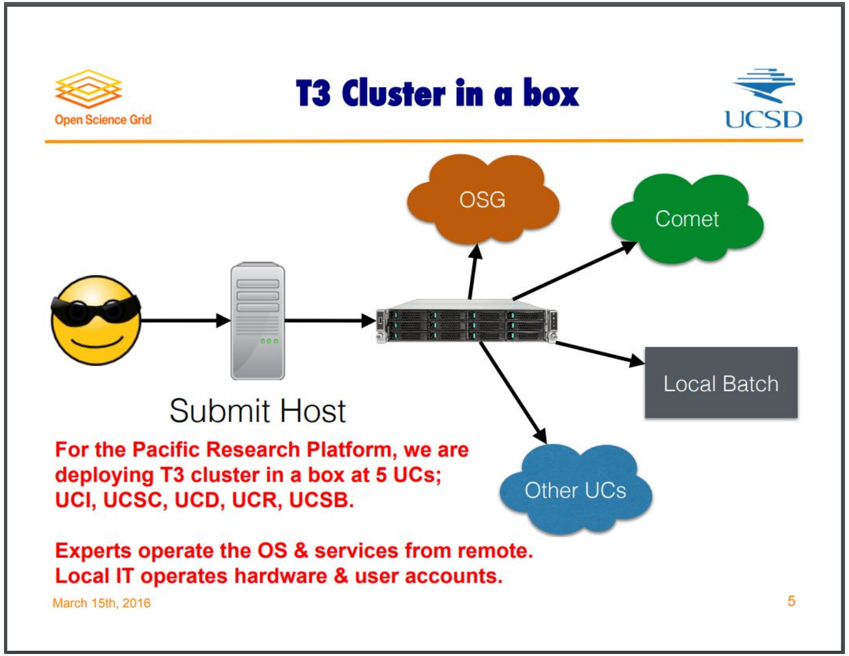
**CI central ops console:**

```
$ slate install osg-squid.3.3 --sites SiteA SiteB SiteC
```

SuperDMZ
Border Router
Enterprise Border Router/Firewall
WAN
10G Routed
10GE
Clean, High-bandwidth path to/from WAN
Site / Campus access to Science DMZ resources
10GE
perfSONAR
10G Virtual Circuit
10GE
Site / Campus LAN
Dedicated path for virtual circuit traffic
Science DMZ Switch/Router
Nx10GE
perfSONAR
Edge container hosting zone on premise
HTTP Proxy
perfSONAR
CE
GUMS
Container Management Framework
node
node
node
node
Site/Campus Virtual Circuits
High Latency WAN Path
Low Latency LAN Path
High Latency VC Path
6

L. Bryant (U Chicago)

# Tier-3 in a box



## T3 Cluster in a box

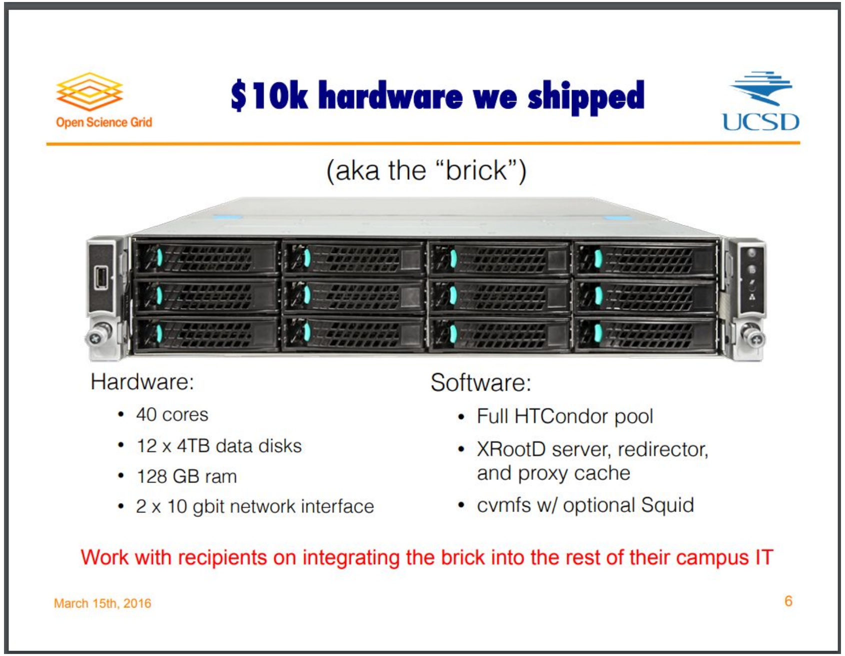
Submit Host

**For the Pacific Research Platform, we are deploying T3 cluster in a box at 5 UCs; UCI, UCSC, UCD, UCR, UCSB.**

**Experts operate the OS & services from remote. Local IT operates hardware & user accounts.**

March 15th, 2016

5



## $10k hardware we shipped

(aka the "brick")

Hardware:

- 40 cores
- 12 x 4TB data disks
- 128 GB ram
- 2 x 10 gbit network interface

Software:

- Full HTCondor pool
- XRootD server, redirector, and proxy cache
- cvmfs w/ optional Squid

Work with recipients on integrating the brick into the rest of their campus IT

March 15th, 2016

6

F. Wurthwein (UCSD/SDSC)

# Where does ARC fit here?

- ARC was always advertised as a lightweight grid middleware
    - Was true when the alternative was CREAM + full worker node
    - Still true in the era of out-of-the-box services in containers/CVMFS?
- Does the world still need CEs?
- Probably more promising for data management
    - ARC cache is truly lightweight, but coupled to ARC CE
    - Could decouple to just run cache + simple transfer service
    - Current R&D to test this in context of Rucio, as alternative to FTS + grid storage
- ARC cache vs xrootd cache
    - xrootd only works for xrootd, ARC is protocol-agnostic
    - ARC cache implies a paradigm shift: requires ARC CE + shared FS, push model, aCT, ...
    - Xrootd cache fits with the dominating pilot pull model
    - How do we fit ARC into the pull model?

# Discussion