ARC–CE: updates and plans

Oxana Smirnova, NeIC/Lund University
1 July 2014
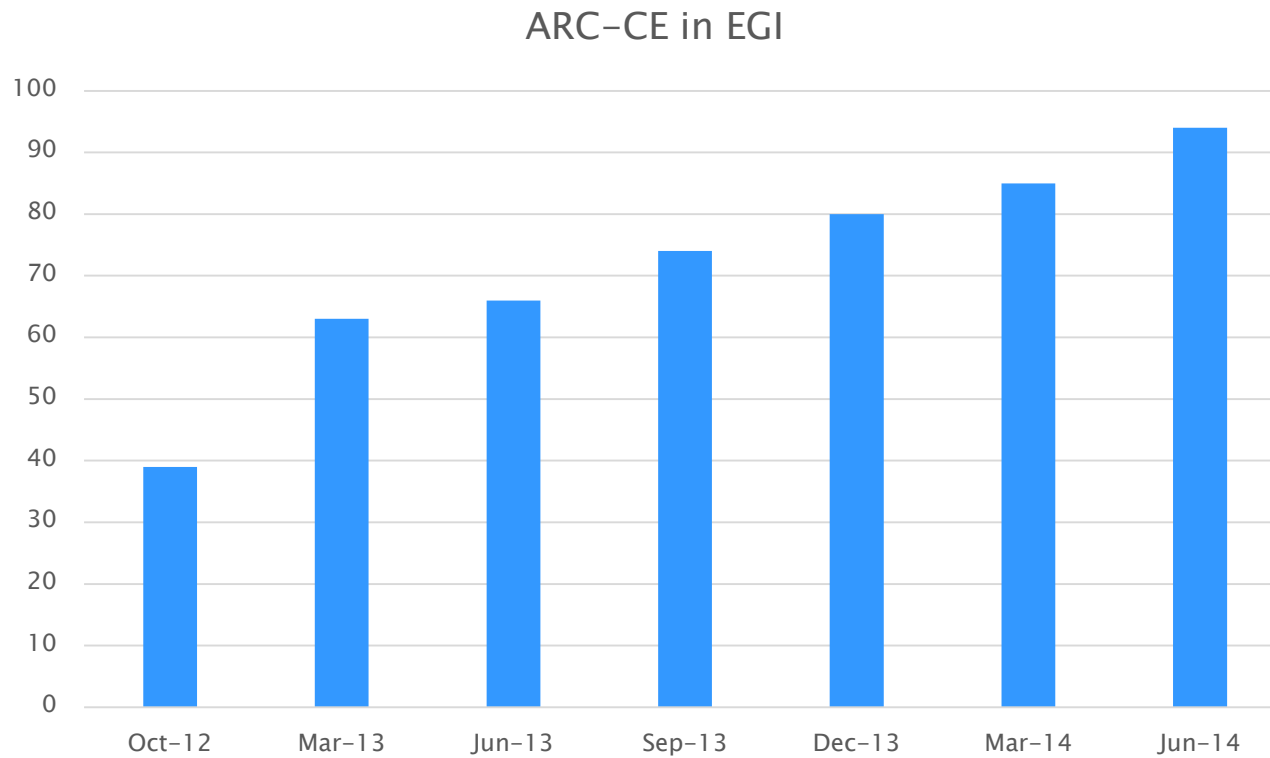*Grid 2014, Dubna*

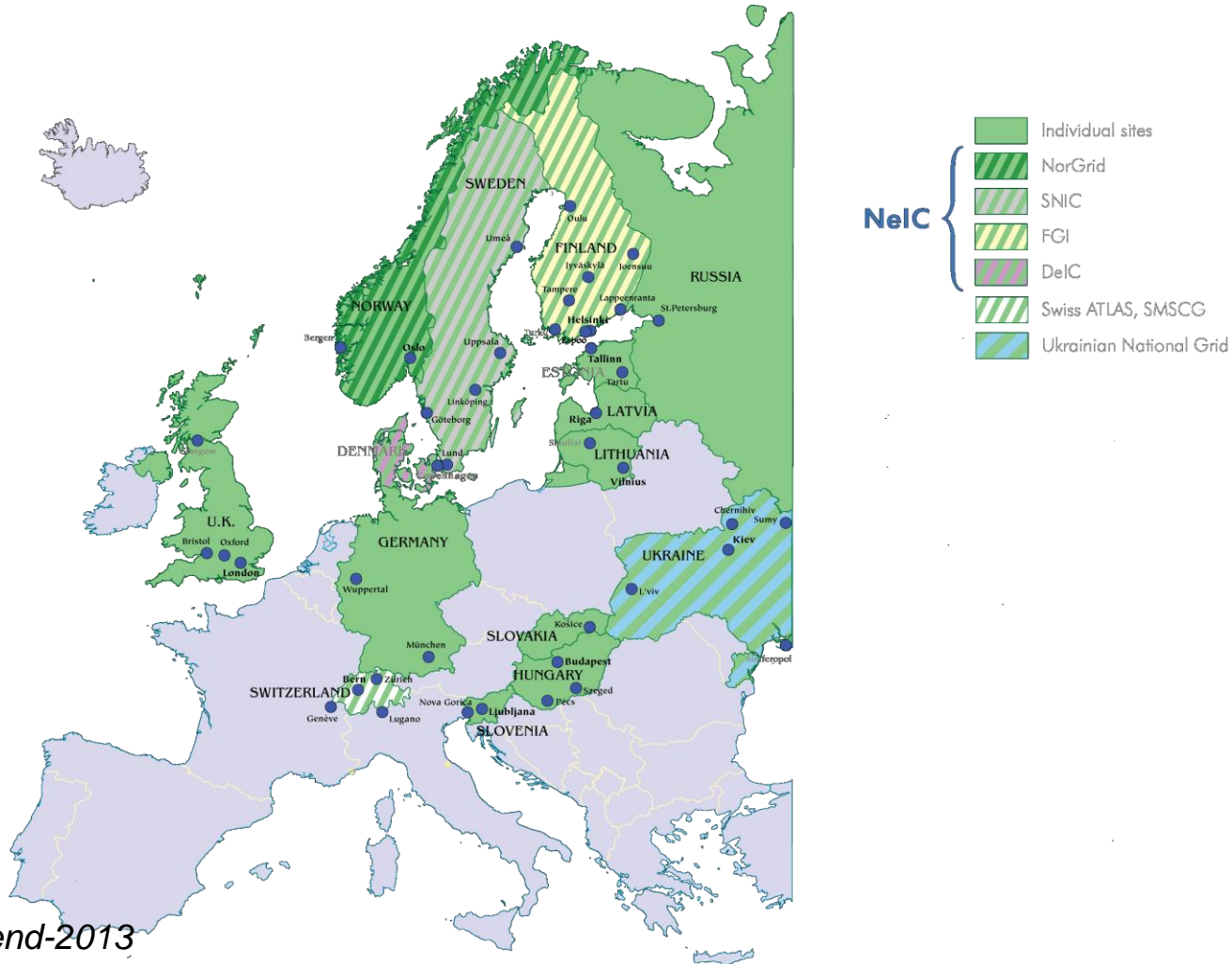*using input from: D. Cameron, A. Filipčič,*
J. Kerr Nilsen, B. Kónya

- Compute element by NorduGrid
- A key component of ARC middleware
  - Other components: clients, information services
- A key enabler of the Nordic Tier-1
  - Along with the distributed dCache storage
- Used by all LHC experiments
  - Mostly by ATLAS (so far)
    - Increasing use by CMS
  - The only CE used in Nordic and Baltic countries
  - Main CE in Slovenia and Ukraine
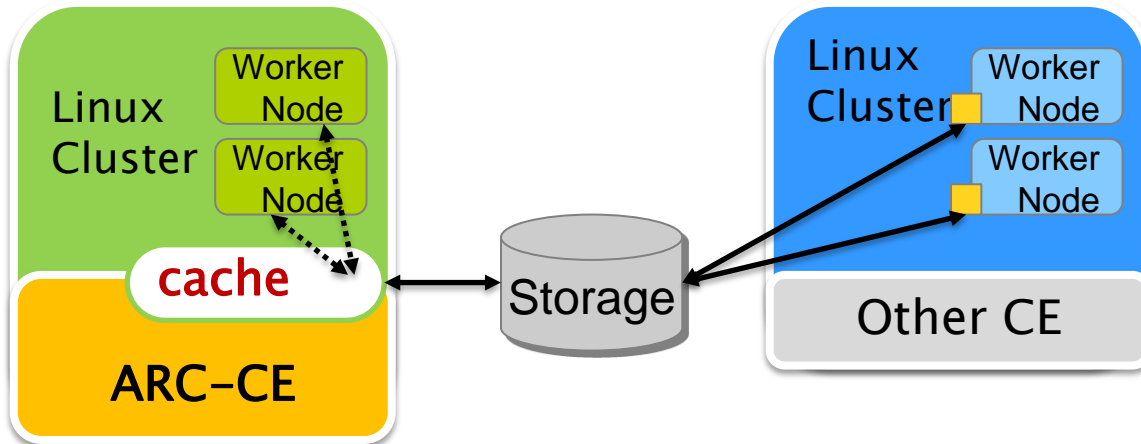 - Increasing use in the UK and Germany

ARC–CE instances in GOCDB

ARC–CE in EGI

cf. CREAM-CE: 566 instances as of today

ARC–CE geography

Data as of end-2013

Optimisation for data-intensive jobs

Linux Cluster
Worker Node
Worker Node
cache
ARC-CE

Storage

Linux Cluster
Worker Node
Worker Node
Other CE

ARC-CE can do all the data transfers

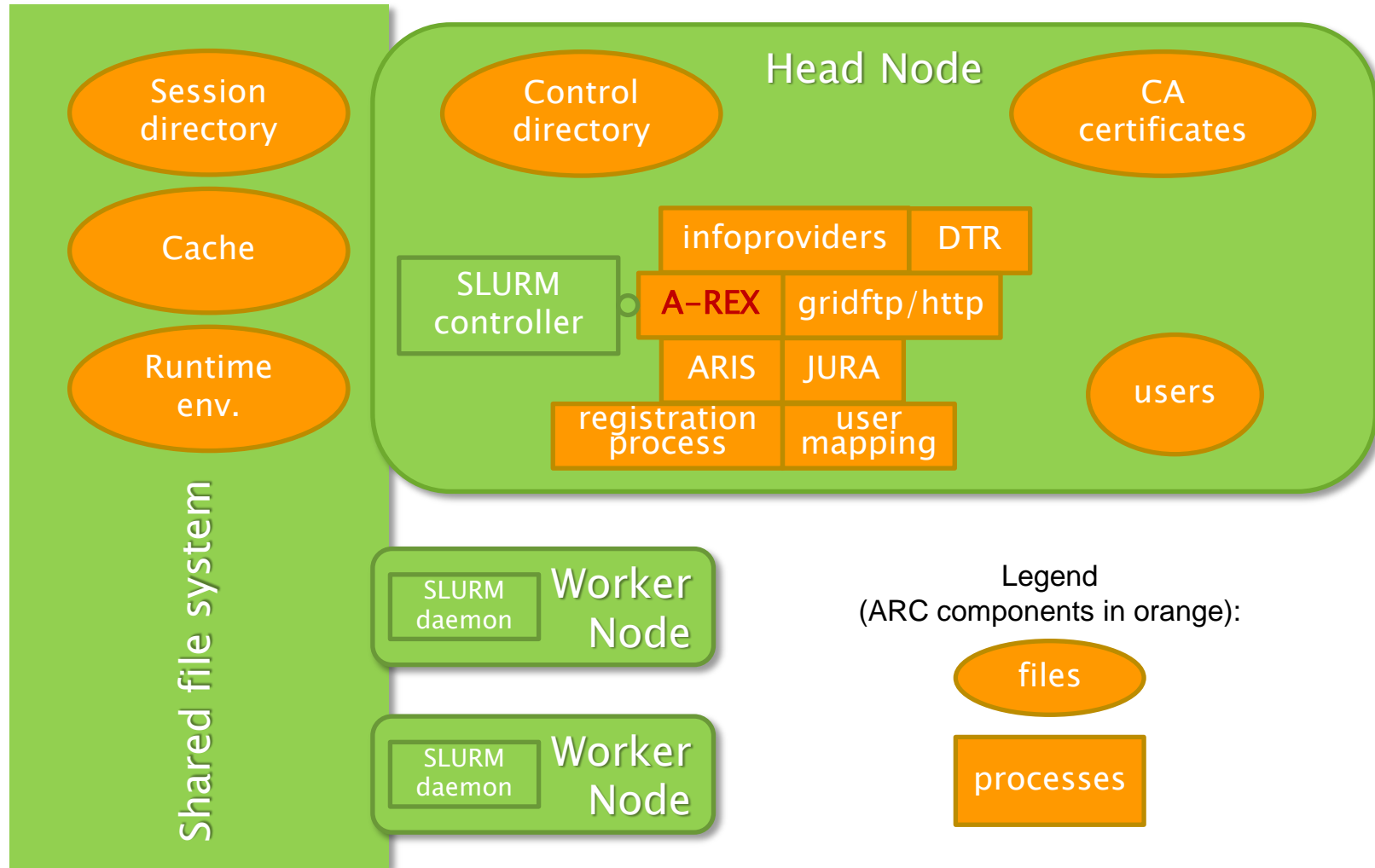Allows to cache **frequently used** files

Minimizes **bandwidth**

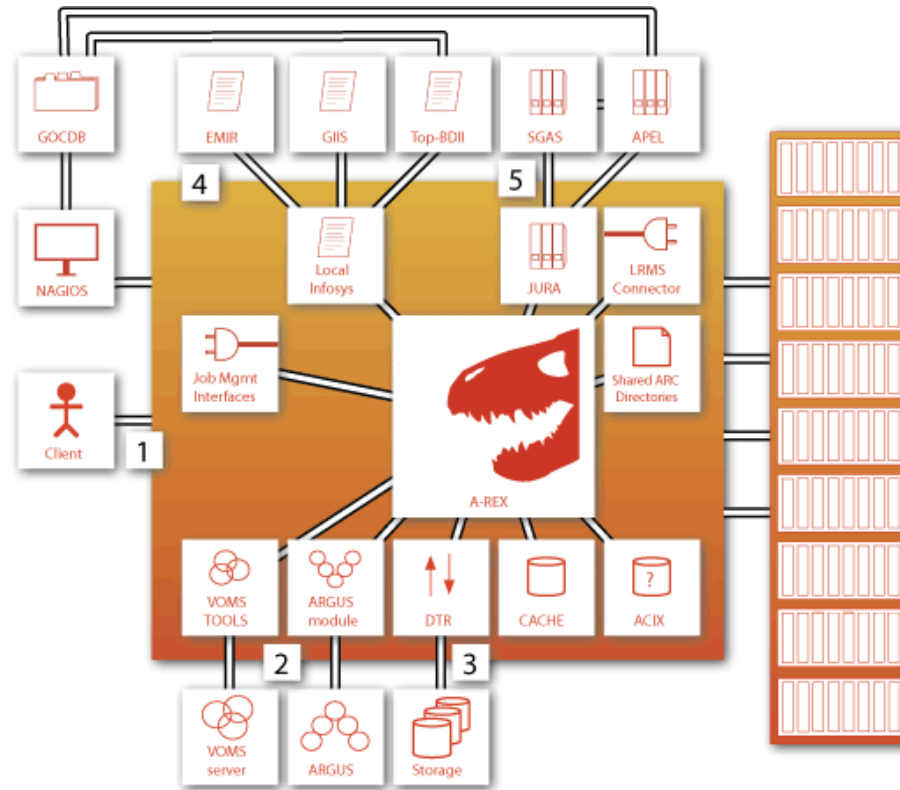Maximizes worker nodes usage **efficiency**

ARC-CE is a rather complex service

Is built of many individual services and tools

Requires high-end storage for cache

NORDUGRID
Grid Solution for Wide Area
Computing and Data Handling

Head Node

Session directory

Control directory

CA certificates

Cache

infoproviders | DTR

SLURM controller

A-REX | gridftp/http

Runtime env.

ARIS | JURA

registration process | user mapping

users

Shared file system

SLURM daemon | Worker Node

SLURM daemon | Worker Node

Legend
(ARC components in orange):

files

processes

1. Job submission (brokering based on info from GIIS, EMIR, Local Infosys and ACIX)
2. Check credentials (VOMS, ARGUS, etc.)
3. Data staging from/to external storage
4. Registration to information indices (EGIIS, EMIR); serving information requests of global aggregators (Top-BDII)
5. JURA parses job logs, prepares and sends job usage records to either SGAS or APEL accounting databases

Control tower for pilots: aCT

- **ARC Control Tower is the layer between ATLAS and ARC**
  - Picks up job descriptions from Panda
  - Converts them to XRSL job description
  - Submits and manages jobs on ARC-CEs
  - Fetches output, handles common failures and updates Panda
- **aCT2: based on ARC API**
- **Modular**
  - ARC actors: submitter, status checker, fetcher, cleaner
  - ATLAS actors: autopilot, panda2arc, atlas status checker, validator

aCT2
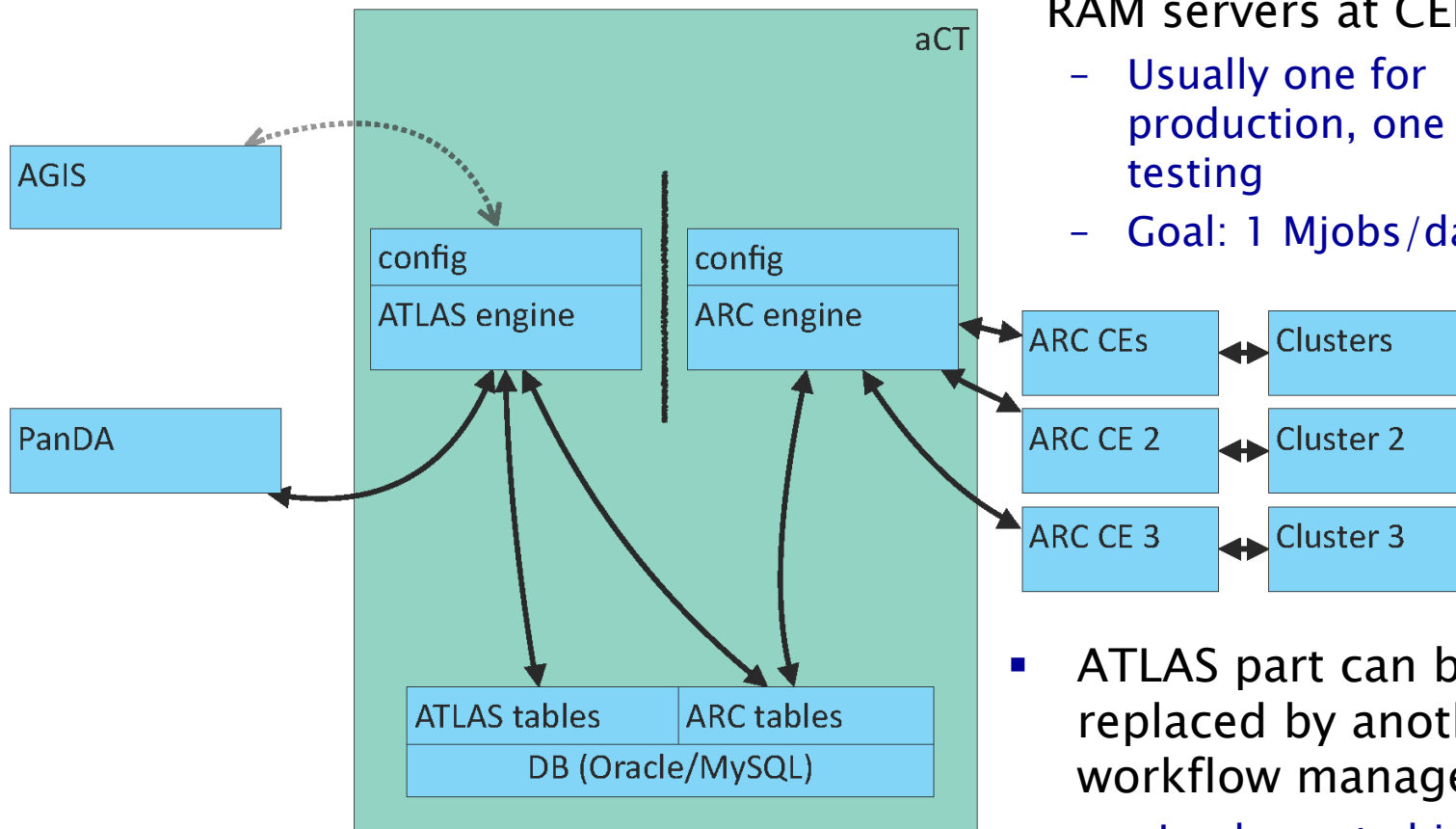
- Runs on two 24 CPU, 50GB RAM servers at CERN
  - Usually one for production, one for testing
  - Goal: 1 Mjobs/day

- ATLAS part can be replaced by another workflow manager
  - Implemented in Python

Gateways to supercomputers

- ATLAS is pursuing usage of HPC resources
  - Long-term funding guaranteed
  - Not suitable for all kinds of jobs still
- In Europe, ARC is often used as a "gateway"
  - C2PAP/SuperMUC, Hydra: via aCT
  - Piz Daint: via ARC-CE ssh back-end (testing)
- Still, a lot remains to be done on both HEP and HPC sides to make it useable
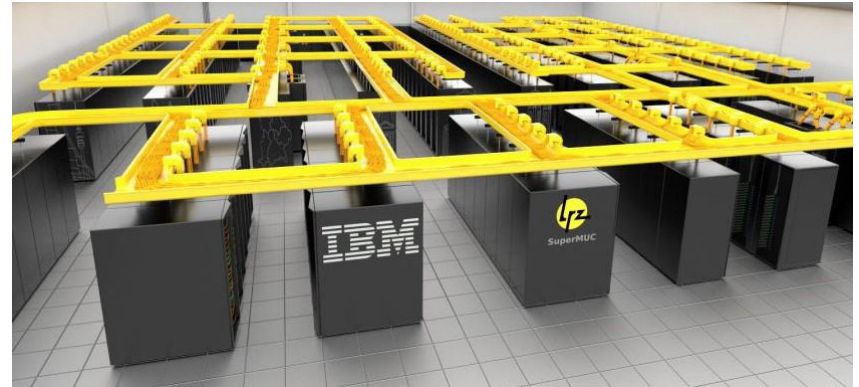
HPC challenges to overcome

- WAN access on nodes is limited or not available
- Whole-node, whole-socket, whole-partition scheduling
- Shared filesystem might suffer with heavy I/O jobs
- Limited access to login/edge nodes
- HPC policies and procedures
  - Sites are tuned for few classes of massively parallel applications with relatively low I/O
  - Limited time slots allocated
  - Access via a SSH-login front node
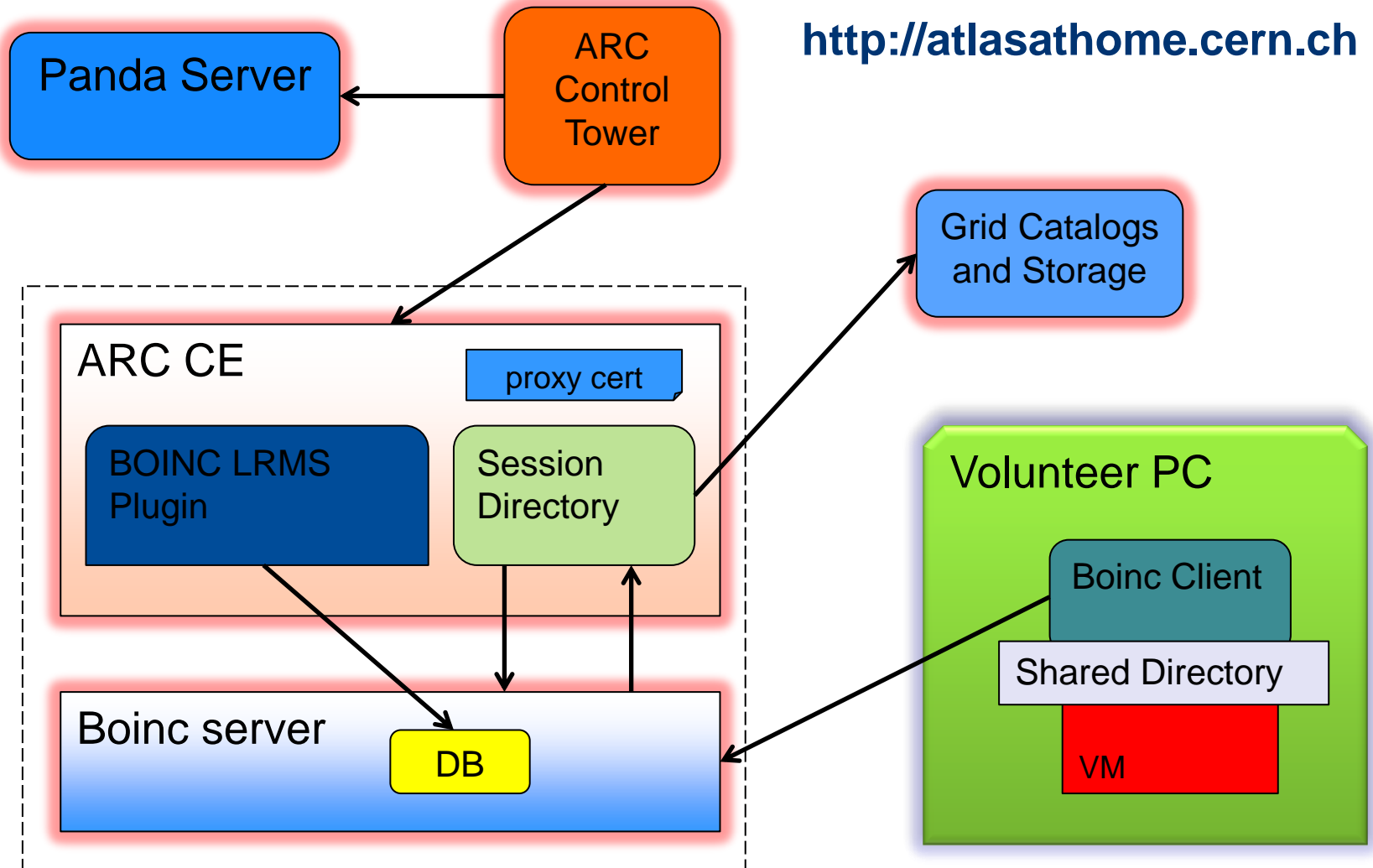
Overcoming the difficulties

- Custom ARC-CE service machines:
  - Using RHEL6 packages on unsupported OS
  - Porting to unsupported OS (SLES11)
- User mode services:
  - Adapting ARC-CE to run as a non-privileged user
  - Limited usability (uid mapped to one batch account)
- No WAN access:
  - aCT, manual software sync, Parrot, no db access
    - Experienced GPFS lock-up due to heavy software area usage (bug fixed by IBM)
- Limited connectivity to edge nodes:
  - ARC SSH backend in the works

NORDUGRID
Grid Solution for Wide Area
Computing and Data Handling

- **The scale of volunteer computing:**
  - At Amazon EC2 price: 1.16USD/hour for a CPU intensive Windows/Linux instance (~1.5GFLOPS) All BOINC activities would cost ~ (7.2PetaFLOPS/1.5GFLOPS)*1.16USD=
    $$=\textbf{5.56M USD/hour}$$
  - Successful Individual projects:
    - Einstein@home (280 TeraFLOPS),
    - SETI@home (581 TeraFLOPS)
    - LHC@home (2TeraFLOPS)

Slide by Wenjing Wu

NORDUGRID
Grid Solution for Wide Area
Computing and Data Handling

- **Why use volunteer computing?**
  - It's free!
  - Public outreach
- **Considerations**
  - Low priority jobs with high CPU-I/O ratio
    - Non-urgent Monte Carlo simulation
  - Need virtualisation for ATLAS sw environment
    - CERNVM image and CVMFS
  - No grid credentials or access on volunteer hosts
    - ARC middleware for data staging
  - The resources should look like a regular Panda queue
    - ARC Control Tower

NORDUGRID
Grid Solution for Wide Area
Computing and Data Handling

http://atlasathome.cern.ch

Panda Server

ARC Control Tower

Grid Catalogs and Storage

ARC CE

proxy cert

BOINC LRMS Plugin

Session Directory

Boinc server

DB

Volunteer PC

Boinc Client

Shared Directory

VM

gilda117.**ihep.ac.cn**/ATLAS/top_users.php?sort_by=expavg_credit

Google

Top participants

David Cameron · lo

Rank	Name	Recent average credit	Total credit	Country	Participant since
1	Andrej Filipcic	592	17,237	Slovenia	19 Feb 2014, 12:45:04 UTC
2	Bok	409	5,126	United States	29 Apr 2014, 15:24:12 UTC
3	David Cameron	264	3,843	Switzerland	15 Jan 2014, 13:18:02 UTC
4	lancon	144	2,175	Switzerland	20 Jan 2014, 16:10:41 UTC
5	Tomas Kouba	129	1,372	Czech Republic	12 May 2014, 8:22:51 UTC
6	Simone Campana	86	956	Switzerland	9 May 2014, 9:21:13 UTC
7	kxt	69	765	Czech Republic	12 May 2014, 13:09:16 UTC
8	zombie67 [MM]	43	450	United States	9 May 2014, 13:15:39 UTC
9	Andrej Gorisek	22	340	France	7 May 2014, 19:29:49 UTC
10	fialal	14	145	Czech Republic	12 May 2014, 13:07:16 UTC
11	kanwx				4 Apr 2014, 5:47:34 UTC
12	Laurence				29 Apr 2014, 12:03:12 UTC
13	wuwj	1	880	International	31 Dec 2013, 2:11:20 UTC
14	APOATLAS	0	0	France	29 Apr 2014, 10:08:41 UTC

No idea who they are!

Summary and outlook

- ARC–CE is well established beyond Nordics now
  - More contributors, too
- aCT2 opens up many more possibilities, such as usage of HPC and volunteer computing
  - Very new, much optimisation is still needed
  - … and also documentation and packaging
- Future directions:
  - Focus on LHC requirements
  - Enhance support of HPC systems, more batch system options
  - Develop more user–friendly task schedulers, using aCT2 experience