КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 407–414

Ки&М

ПЛЕНАРНЫЕ ДОКЛАДЫ

УДК: 004.75

ARC-CE: updates and plans

O. Smirnova[1,a], B. Kónya[1], D. Cameron[2], J. K. Nilsen[2] and A. Filipčič[3]

[1] Dept. of Physics, Lund University, 221 00, Lund, Sweden, Professorsgatan 1
[2] Dept. of Physics, University of Oslo, N-0316 Oslo, Norway, 1048 Blindern
[3] Institute Jožef Stefan, 1000 Ljubljana, Slovenia, Jamova 39

E-mail: [a]oxana.smirnova@hep.lu.se

Received October 27, 2014

ARC Compute Element is becoming more popular in WLCG and EGI infrastructures, being used not only in the Grid context, but also as an interface to HPC and Cloud resources. It strongly relies on community contributions, which helps keeping up with the changes in the distributed computing landscape. Future ARC plans are closely linked to the needs of the LHC computing, whichever shape it may take. There are also numerous examples of ARC usage for smaller research communities through national computing infrastructure projects in different countries. As such, ARC is a viable solution for building uniform distributed computing infrastructures using a variety of resources.

Keywords: distributed computing, grid computing, volunteer computing

ARC-CE: новости и перспективы

О. Смирнова[1], Б. Коня[1], Д. Кэмерон[2], Й. К. Нильсен[2] и А. Филипчич[3]

[1] Отд. физики, Университет Лунда, 221 00 Лунд, Швеция
[2] Отд. физики, Университет Осло, N-0316 Осло, Норвегия
[3] Институт Йожефа Стефана, 1000 Любляна, Словения

Вычислительный элемент ARC приобретает всё большую популярность в инфраструктурах WLCG и EGI, и используется не только в контексте систем Грид, но и как интерфейс к суперкомпьютерам и облачным ресурсам. Развитие и поддержка ARC опирается на вклады членов пользовательского сообщества, что помогает идти в ногу со всеми изменениями в сфере распределённых вычислений. Перспективы развития ARC тесно связаны с требованиями обработки данных БАК, в любых их проявлениях. ARC также используется и для нужд небольших научных сообществ, благодаря государственным вычислительным инфраструктурам в различных странах. Таким образом, ARC представляет собой эффективное решение для создания распределённых вычислительных инфраструктур, использующих разнообразные ресурсы.

Ключевые слова: распределённые вычисления, грид-вычисления, добровольные вычисления

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 404–414.

1. Introduction

ARC Compute Element (ARC-CE) is a Grid Compute Element at the core of the ARC middleware developed by NorduGrid [Ellert M et al., 2007]. As of today, it is a key component of ARC middleware, other components being clients for computing and data handling tasks, and information services that advertise ARC-CE status and capacity. ARC-CE is a key enabler of the Nordic Tier-1 operated by NeIC [NeIC Web site], which main characteristic is the distributed nature of both computing and data services. While storage pools across different Nordic countries are federated into a single instance by dCache [dCache Web site], computing services rely on ARC-CE thanks to its capability of caching input data.

Inclusion of ARC into EMI [European Middleware Initiative] middleware stack made it readily available to all sites that support Worldwide LHC Computing Grid (WLCG) [Worldwide LHC Computing Grid] and participate in the European Grid Infrastructure (EGI) [EGI Web site]. It is now used by all large LHC experiments, with ATLAS being the largest user of ARC-CE, followed by an increasing usage by CMS, ALICE and LHCb. In some countries, like the Nordic and Baltic states, as well as Slovenia, ARC-CE is the only Compute Element in use. As can be seen in Figure 1, ARC-CE deployment steadily increases over the past two years.

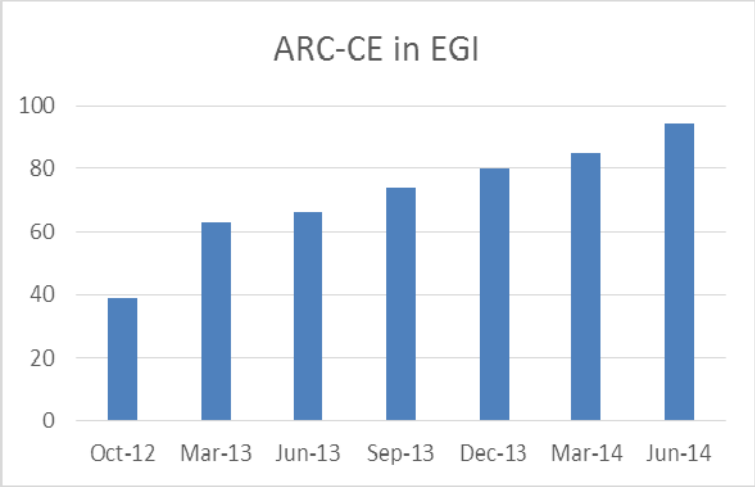


Fig. 1. Number of ARC-CE instances in the EGI database

Figure 2 shows geographical distribution of ARC services, including both ARC-CE and information indices. It serves as a basic service for several national Grid infrastructures, as indicated in the figure.

2. ARC-CE principles

ARC-CE is optimised for data-intensive jobs, taking particular care of staging input data to the local cache for eventual re-use, and staging output data to wherever the job description requests them to be staged.

Data movement is performed by dedicated processes hosted by ARC-CE, and a job will not be submitted to a local batch system until all input data are downloaded by the Compute Element. This may occur as a delay in the job start from the point of view of the user, but in fact this saves time and bandwidth by removing the necessity to move data by the jobs themselves, and moreover, availability of the input data in cache will make the next job that needs same data starting instantly. This approach maximises CPU utilization and minimizes bandwidth. In addition, worker nodes managed by ARC-CE do not require network connectivity if input and output data locations are known in advance, as shown in Figure 3.

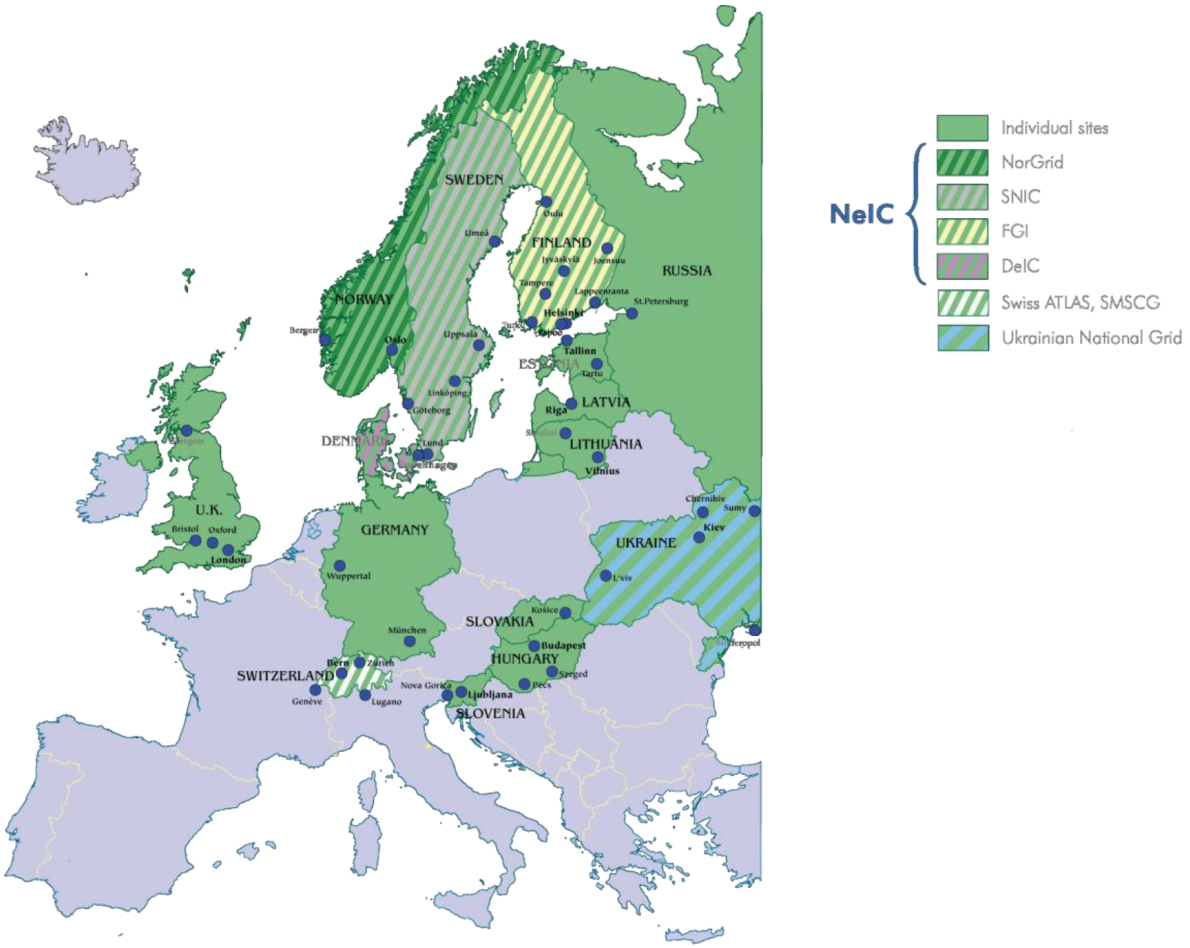


Fig. 2. Geographical distribution of ARC services deployment

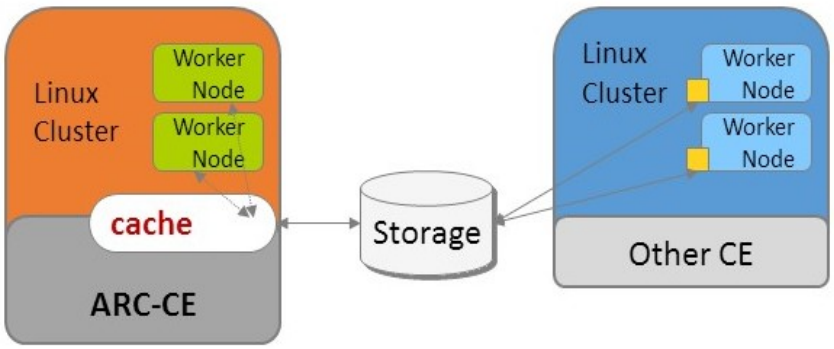


Fig. 3. Comparison of basic ARC-CE principles to those of other CEs: worker nodes managed by ARC-CE need no network connectivity

ARC-CE is a complex service, consisting of many services and utilities. It is quite demanding resource- wise, and needs a fast shared file system as well as a high-end storage server for the cache.

Figure 4 presents an overview of ARC-CE components when installed on a SLURM [SLURM Workload Manager] cluster. In general, ARC-CE supports a large variety of batch systems, including, in addition to SLURM: HTCondor [HTCondor Web site], PBS flavours, Grid Engine flavours, LoadLeveler and LSF. Support levels however differ, relying on community contributions.

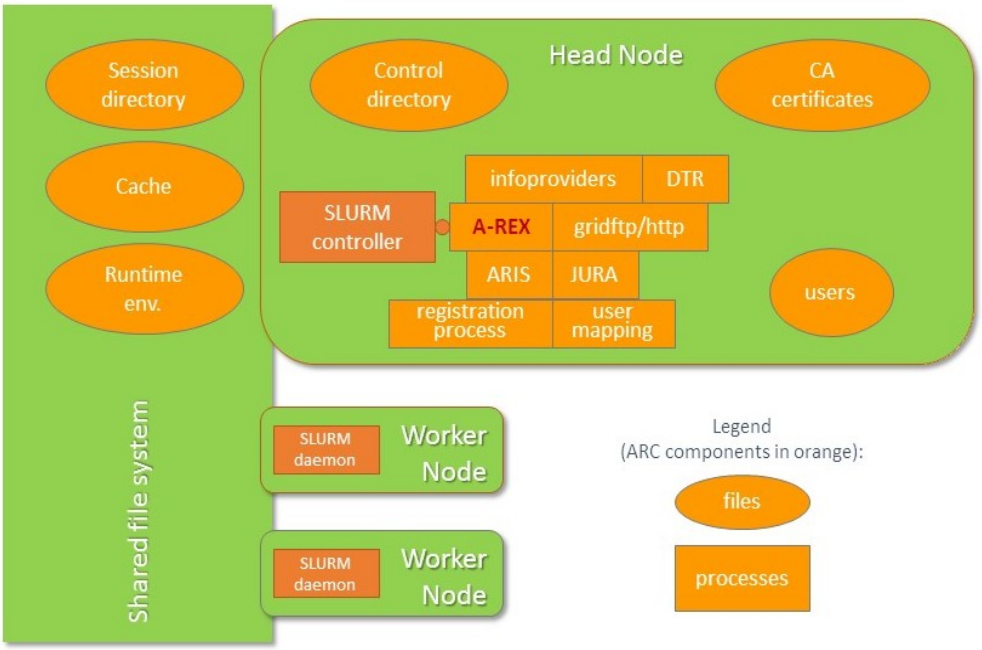


Fig. 4. Overview of ARC-CE components when installed on a SLURM-managed cluster

3. ARC Control Tower

While ARC-CE is designed to move data, it can be used as a generic CE for pilot jobs. This approach, though widely used, does not make use of all the benefits of ARC-CE. In order to optimize resource usage, a special service, ARC Control Tower (aCT), has been developed by NorduGrid and ATLAS [Filipčič A et al, 2011]. aCT is an external service functioning as a workload management system for ARC-CE, specializing on pilot jobs. Its function is to extract job descriptions from pilots, convert them to deterministic ARC-CE jobs, and schedule these "classic" jobs to best suited ARC-CEs. This service is extensively used by ATLAS, but its latest version, aCT2, would allow using it for other workflows as well. Figure 5 shows components of aCT in the current case of ATLAS production deployment. The ATLAS-specific modules are separate from the ARC-specific ones, and only share a common database instance. aCT is developed using ARC API and libraries (ARC SDK), and since it is modular, it is quite straightforward to replace ATLAS-specific components with those supporting other workflows.

In the ATLAS production scenario, aCT presents itself as a computational resource, picking job descriptions from PanDA [Maeno T., 2008], converting them into ARC-specific XRSL job descriptions, and then submitting and managing jobs on ARC-CEs. Upon job completion, aCT fetches output files, handles common failures in case such occurred, and updates PanDA with job status and other required information.

The modules of aCT are known as actors, each responsible for specific actions. ARC actors are: *submitter, status checker*, *fetcher* and *cleaner*, and ATLAS actors are: *autopilot*, *panda2arc*, *atlas status checker* and *validator*. The names of actors speak for themselves; detailed aCT documentation is expected to be released in near future.

4. Gateways to supercomputers

WLCG computing has so far relied mostly on dedicated resources, configured to meet the needs of LHC experiments. However, the current trend of streamlining research computing by investing into

large-scale HPC or Cloud centres motivates LHC experiments to investigate possibilities of using such non-traditional resources. In particular, national research HPC systems are attractive because of guaranteed long-term funding and massive CPU capacities. However, such systems are not designed for high-throughput data processing of the kind used by the LHC experiments. Still, using them for Monte Carlo generation is quite a feasible task.

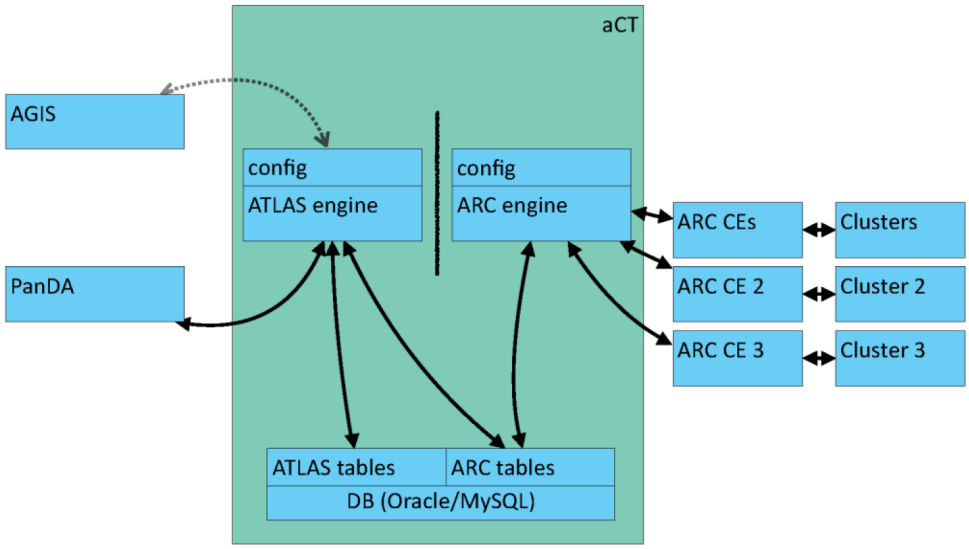


Fig. 5. aCT scheme for deployment with ATLAS' PanDA system; ATLAS-specific modules are clearly separated from ARC-specific ones

ARC does not require installation of any additional software on the worker nodes, neither it requires network connectivity for them. This makes it possible to use ARC-CE as a gateway for HPC systems, establishing interaction to the batch system via *ssh* or via aCT deployed on-site. These approaches are being tested on several HPC systems in Europe, such as *C2PAP/SuperMUC* and *Hydra* in Germany, and *Piz Daint* in Switzerland. Though initial tests are encouraging, and allow for certain types of ATLAS production jobs, still, a lot remains to be done on both HEP and HPC sides to make it useable.

In order to make a reasonable use of HPC resources, the following challenges have to be addressed:

- WAN access on worker nodes of HPC systems is limited (or absent), while it is still needed by many jobs in order to communicate to other services, such as databases
- Job scheduling has to be flexible enough to allow for cases like whole-node, whole-socket or whole-partition scheduling
- Shared file systems used by HPC sites are not necessarily optimal for heavy Input/Output, especially when thousands of processes do simultaneous read and write
- Traditionally, HPC sites offer dedicated login/edge nodes to the users; access to these nodes is strictly controlled and rather limited, in a manner not consistent with multi-user VOs
- In general, HPC policies and procedures are not suitable for WLCG use cases, where multi-user VOs use robot credentials and require installation of hundreds of different versions of proprietary software

Typically, HPC sites are tuned for few classes of massively parallel applications with relatively low I/O, allocate limited (not permanent) time slots to well-identified users, and only allow remote access via a *ssh*-login front nodes. Moreover, HPC systems rarely use Scientific Linux for OS, while it still remains the only supported OS for WLCG. This usage model is clearly not suitable for HEP computing.

There are still ways to meet requirements of both worlds, at least to some extent. Some site policies allow deployment of ARC-CE service machines: they can either be ported to the host operating

system, or binary-compatible packages can be used where possible. If site policies require deploying ARC-CE in a user mode, it can be adapted to run from a non-privileged user account; this however limits the usability of the system, as typically a UID can be mapped to only one batch account. In cases when site nodes have no WAN access, a more complex configuration using aCT as a gateway is possible; this requires manual or semi-automatic synchronization of necessary software and possibly of databases, for off-line usage. Tests showed that this may cause heavy load on shared file systems, thus not every site might be able to deploy such a setup. In cases when even the edge nodes have limited connectivity, ARC-CE needs to be deployed outside of the site, communicating via the *ssh*-back-end, which is currently being developed.

5. ATLAS@Home

Volunteer computing is credited as being one of the Grid pre-cursors, but it is rarely considered as a serious resource for the LHC needs. Still, BOINC-based LHC@home project clocked 2 Tera-FLOPS, which is perhaps not as impressive as hundreds of TeraFLOPS of SETI@home or Einstein@home, but still a valuable contribution. Apart of providing Cloud-like resources for free, it is a very useful mean for public outreach and popularization of LHC physics.

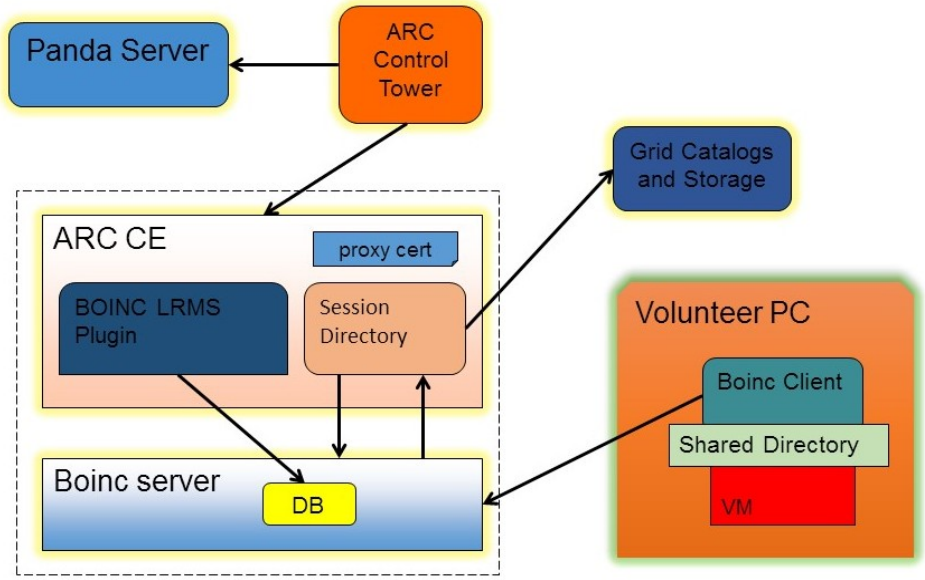


Fig. 6. Scheme of the ATLAS@home volunteer computing project: ATLAS jobs are retrieved from the PanDA server by aCT, and submitted to ARC-CE with a BOINC back-end

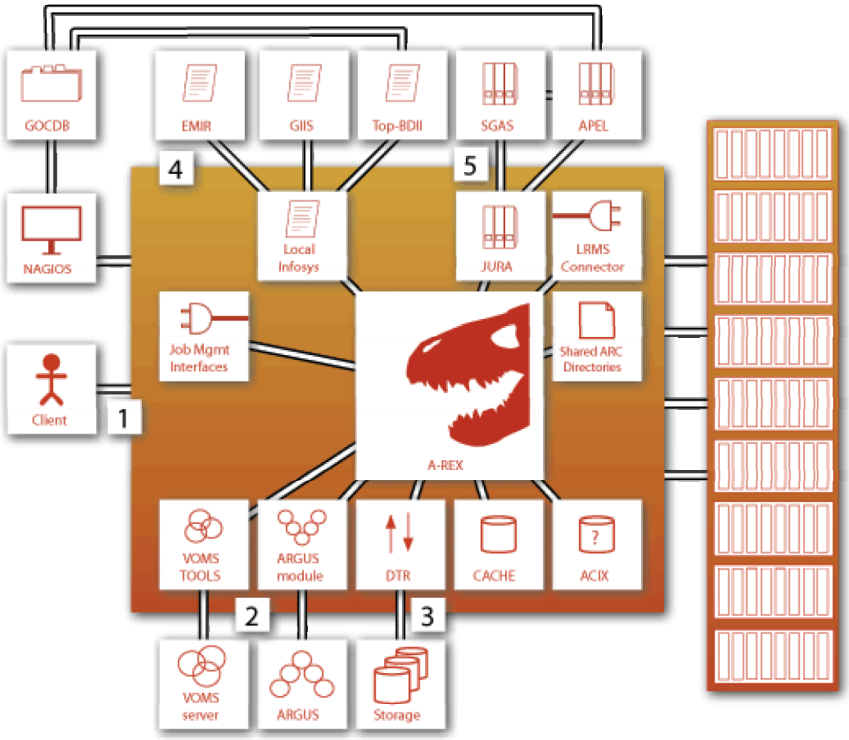
ATLAS recently has launched a BOINC-based volunteer computing project ATLAS@home [ATLAS@home Web site], which makes use of the ARC Control Tower and ARC-CE with a specially crafted BOINC back-end to populate BOINC server with real production jobs. The overall setup is shown in Figure 6.

There are certain considerations to be taken into account, related to the nature of volunteer computing resources. Clearly, one cannot rely on them for top-priority or data-intensive tasks, thus jobs suitable for ATLAS@home are low-priority jobs with high ratio of CPU to I/O, such as non-urgent Monte Carlo simulations. Virtualization is needed for ATLAS software environment, which is achieved through usage of CERNVM images and CVMFS-based software deployment. Since volunteer hosts have no Grid credentials and have access to Grid storages, data need to be staged by ARC middleware components. Ultimately, aCT makes the volunteer resources looking much like a regular queue from the PanDA point of view.

ATLAS@home enjoys a rather unexpected popularity, and at times provides more processing power than some Tier2 centres. Despite the fact that the individual contributors are unknown to the physicists and are not subject to standard WLCG operational procedures, control or monitoring, the service they offer is very valuable and much appreciated.

6. Integration with EGI operations

Thanks to the EMI efforts, ARC-CE is now well-integrated with EGI. Figure 7 shows relation of ARC-CE and its internal services to such EGI components as accounting, VO management, authorisation, monitoring, indexing and cataloguing. This allows for quite a smooth migration from other WLCG CEs (usually, CREAM-CE) to ARC-CE, and an increasing number of WLCG sites are switching to ARC-CE these days. Even if occasional glitches are discovered during deployment of ARC-CEs in previously untested configurations, such issues are quickly solved thanks to the active community of code contributors and openness for new contributions.



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

Fig. 7. Relation of ARC-CE and its internal services to EGI components and other relevant Grid services

7. Summary and outlook

ARC-CE is a well-established Grid computing service, used by WLCG and other Grid sites well beyond its Nordic origin. Being a community-driven effort, ARC benefits from knowledge and expertise brought in by every new site, and the list of ARC code contributors keep growing. Particularly active contribution area is back-ends to various flavors of batch systems, and even to such non-traditional resources as volunteer computing or *ssh*-accessible HPC systems.

ARC Control Tower, serving as a gateway between production systems and ARC-based resources, opens up many new possibilities of adding computing power to WLCG. Most notable examples are aCT usage in conjunction with HPC resources, and an amazing success of the ATLAS@home project, which makes use of aCT, ARC-CE and BOINC. aCT is still work in progress, and more tuning is needed to optimize its performance. Documentation and proper packaging of aCT are other important tasks that welcome contributors.

Future of ARC is inevitably linked to the ever increasing LHC computing requirements; immediate focus is on enhancing support for inclusion of HPC systems into WLCG, and adding support for more batch system options, particularly those related to multi-core processing. aCT2 experience will also help to develop more user-friendly task schedulers, possibly even for other communities outside WLCG.

References

ATLAS@home Web site URL http://atlasathome.cern.ch

dCache Web site URL http://www.dcache.org

EGI Web site URL http://www.egi.eu

Ellert M et al. Future Gener. Comput. Syst. 2007. 23 219-240 ISSN 0167-739X

European Middleware Initiative Web site URL http://www.eu-emi.eu

Filipčič A et al 2011 J. Phys.: Conf. Ser. 331 072013

HTCondor Web site URL http://research.cs.wisc.edu/htcondor/

Maeno T. J. Phys.: Conf. Ser. 119 062036. 2008

NeIC Web site URL http://neic.nordforsk.org

SLURM Workload Manager Web site URL http://slurm.schedmd.com

Worldwide LHC Computing Grid Web site URL http://lcg.cern.ch