HTCondor Workshop Autumn 2021

Report of Contributions

Contribution ID: **1**   Type: **not specified**

Dealing with dynamic and mixed workloads

Monday 20 September 2021 17:20 (30 minutes)

At INFN-T1 several competing groups submit their payloads to the HTCondor pool with a high level of heterogeneity. In particular, the same group can submit both multi core and single core jobs, and the ratio between these two can change quite rapidly; this and other unpredictable user side behaviours can make difficult for HTCondor administrators to provide user groups with a satisfactory fair share of the available computing resources.
As an attempt to reduce usage imbalances between different user groups, a system to self adjust disparities has been developed and it is being used with good results so far.

Desired slot length

Speaker release

Yes

Author: DAL PRA, Stefano (Universita e INFN, Bologna (IT))

Presenter: DAL PRA, Stefano (Universita e INFN, Bologna (IT))

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **2**   Type: **not specified**

Auto-scaling in the cloud: Intelligent HTCondor resource management

Tuesday 21 September 2021 15:00 (30 minutes)

HTCondor is an effective tool to rank and match execute resources against a set of jobs with explicit resource requirements. In the cloud, a subtly different challenge is presented: how to rank execute resource configurations that will be automatically created to run idle jobs (auto-scaled on-demand).

We describe recent work by the HTCondor team and Google Cloud to provide built-in support for commonly desired patterns in cloud auto-scaling. For example, a job can require co-location of execute resources with data stored as Google Cloud Storage objects. Alternatively, a group of jobs might seek to expand into as many cloud regions as possible in search of cost savings or to minimize the wall-clock time of a particular workflow.

Desired slot length

20

Speaker release

Yes

Authors: Dr THOMSON, Ross (Google); DOWNES, Tom (Google)

Presenters: Dr THOMSON, Ross (Google); DOWNES, Tom (Google)

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **3**   Type: **not specified**

A new HTCondor monitoring for CNAF Tier-1

Monday 20 September 2021 17:50 (30 minutes)

The CNAF Tier-1, composed of almost 1000 worker nodes and nearly 40000 cores, completed its migration to HTCondor more than one year ago. After having adapted existing monitoring tools (built with Sensu, Influx and Grafana) to work with the new batch system, an effort has started to collect a more rich and "condor oriented"set of metrics that are used to provide better insights on the pool status.
The data are collected into a PostgreSQL database, which makes them also available for further analysis or different applications, and presented by a specifically designed dashboard built using the dash and plotly python libraries.

Desired slot length

Speaker release

Yes

Authors: VERSARI, Federico (University of Bologna); DAL PRA, Stefano (INFN)

Presenter: VERSARI, Federico (University of Bologna)

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: 4

Type: **not specified**

HTCondor Integration with Hashicorp Vault for Oauth Credentials

Wednesday 22 September 2021 16:35 (30 minutes)

HTCondor now has an optional integration with open source Hashicorp Vault for managing Java Web Tokens (JWTs) such as Scitokens. In the integration, the condor_submit command calls out to htgettoken (developed at Fermilab) to communicate with a Vault service. Vault takes care of the Open ID Connect protocol (which is based on Oauth 2.0) to communicate with a token issuer and securely storing powerful refresh tokens while returning less powerful Vault tokens that can be used to obtain even less powerful access JWTs. In the initial authentication, htgettoken redirects the user to their web browser for approval, but subsequent requests for access JWTs use either the Vault token or renew the Vault token using Kerberos authentication. A Vault credmon component holds Vault tokens that it exchanges for access JWTs to renew in batch jobs. The submit file can specify just the name of a token issuer configured in Vault, and it can optionally specify specific scopes or audiences to further restrict the power of access JWTs. This talk will describe the HTCondor Vault integration in detail.

Desired slot length

Speaker release

Yes

Author: DYKSTRA, Dave (Fermi National Accelerator Lab. (US))

Presenter: DYKSTRA, Dave (Fermi National Accelerator Lab. (US))

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **5**   Type: **not specified**

In silico detection of (CRISPR) spacers matching Betacoronaviridae genomes in gut metagenomics sequencing data

Friday 24 September 2021 17:20 (30 minutes)

In silico detection of (CRISPR) spacers matching Betacoronaviridae genomes in gut metagenomics sequencing data

Leoni G.1,2, Petrillo M.2, Puertas-Gallardo A.2, Sanges R.1, Patak A.2

1. Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste (Italy);
2. Joint Research Center (JRC), Ispra (Italy).

The CRISPR-Cas system is the major component of the prokaryotic adaptive immune system (Horvath
& Barrangou, 2010). CRISPR, which stand for "Clustered Regularly Interspaced Short Palindromic
Repeats", are genomics arrays found in the DNA of many bacteria. They consist in short repeated
sequences (size 23-47 base pairs), separated by unique sequences of similar length (spacers), that
often
derives from phages and viral infections, plasmids or mobile genetic elements (Shmakov et al.,
2017).
CRISPRs are coupled to specific "CRISPR-associated genes"(Cas) to form the so called CRISPR-Cas
system. This system has the primary role to protect prokaryotes from virus and other mobile genetic
elements activity by conferring immunological memory from past infections (Garneau et al., 2010;
Nussenzweig & Marraffini, 2020).
Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a single-stranded RNA virus
that
rapidly emerged in 2019. In humans, it causes coronavirus disease 2019 (COVID-19), an influenza-
like
disease that is primarily thought to infect the lungs with transmission through the respiratory
route.
However, clinical evidence suggests that the intestine may present another viral target organ, a
potential
hiding place for the virus, which may explain the persistence of COVID-19 symptoms after months
from patients recovery (Lamers et al., 2020). Furthermore, extra-pulmonary clinical manifestations
of
COVID-19 are reported. Nonetheless, although a link between SARS-CoV-2 infection and the mis-
regulation
of the gut microbiome was suggested, its involvement remains largely unexplored (Brooks
& Bhatt, 2021).
To simultaneously verify both the potential existence of SARS-CoV-2 in gut and to test whether
the
human gut microbiome may be stressed by SARS-CoV-2 infection, we developed a bioinformatic
workflow based on the detection of Betacoronaviridae-specific CRISPR spacers from ˜28,000 public
available gut metagenomics data. To process such "Big Biological Data"in a reasonable CPU time,
we
relied on a HTCondor High Throughput Computing System, characterized by 10 Tflops of computing
capacity and more than 80 Tbytes of storage. Computing block was composed by 8 Nodes IBM
x3550
with two Intel Xeon processor E5-2600 v3 product family CPUs with10 cores 2.6 GHz, two QPI

links
up to 9.6 GT/s each and 256 GB of RAM. While our work is still ongoing, preliminary results revealed
the presence of some Betacoronavirus-specific spacers in the human gut metagenomics data, proving
that SARS-like viruses can target human gut and suggesting that the human microbiome can be
stressed by the systemic viral infection. By collecting further data, we aim to strengthen our results as
well as to investigate the effects of the SARS-COV-2-induced microbiome stress to the host.

Bibliography

Brooks, E. F., & Bhatt, A. S. (2021). The gut microbiome: A missing link in understanding the
gastrointestinal manifestations of COVID-19? Molecular Case Studies, 7(2), a006031.
https://doi.org/10.1101/mcs.a006031
Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C.,
Horvath, P., Magadán, A. H., & Moineau, S. (2010). The CRISPR/Cas bacterial immune system
cleaves bacteriophage and plasmid DNA. Nature, 468(7320), 67–71.
https://doi.org/10.1038/nature09523
Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea.
Science. https://www.science.org/doi/abs/10.1126/science.1179555
Lamers, M. M., Beumer, J., Vaart, J. van der, Knoops, K., Puschhof, J., Breugem, T. I., Ravelli, R. B.
G., Schayck, J. P. van, Mykytyn, A. Z., Duimel, H. Q., Donselaar, E. van, Riesebosch, S.,
Kuijpers, H. J. H., Schipper, D., Wetering, W. J. van de, Graaf, M. de, Koopmans, M., Cuppen,
E., Peters, P. J., ⋯Clevers, H. (2020). SARS-CoV-2 productively infects human gut
enterocytes. Science. https://www.science.org/doi/abs/10.1126/science.abc1669
Nussenzweig, P. M., & Marraffini, L. A. (2020). Molecular Mechanisms of CRISPR-Cas Immunity
in
Bacteria. Annual Review of Genetics, 54(1), 93–120. https://doi.org/10.1146/annurev-genet-
022120-112523
Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2017). The
CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. mBio,
8(5), e01397-17. https://doi.org/10.1128/mBio.01397-17

Desired slot length

20

Speaker release

Yes

Author: LEONI, Gabriele (SISSA)

Presenter: LEONI, Gabriele (SISSA)

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **6**

Type: **not specified**

Operations in the HTCondor pool at CERN

Tuesday 21 September 2021 17:20 (30 minutes)

During the last year the HTCondor pools at CERN have passed the milestone of 300K cores. In this presentation we will cover some of the operational challenges we have found and the various monitoring and automation solutions deployed to tackle them. We will review as well how we envision the evolution of the service in the coming years.

Desired slot length

Speaker release

Yes

Authors: JONES, Ben (CERN); FERNANDEZ ALVAREZ, Luis (CERN)

Presenter: FERNANDEZ ALVAREZ, Luis (CERN)

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **8**   Type: **not specified**

Upgrading to HTCondor 9.0

Wednesday 22 September 2021 15:40 (40 minutes)

The upgrade to HTCondor 9.0 isn't as smooth as previous upgrades. This talk discusses why and what to do.

Desired slot length

Speaker release

Yes

Author: MILLER, Todd Lancaster (University of Wisconsin Madison (US))

Presenter: MILLER, Todd Lancaster (University of Wisconsin Madison (US))

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **9**
Type: **not specified**

Running multiple experiment workflows on heterogeneous resources, the RAL experience

Tuesday 21 September 2021 17:50 (30 minutes)

The RAL Tier-1 runs an almost 50,000 core HTCondor batch farm which supports not only the four major LHC experiments but an increasing number of other experiments in the High Energy Physics, Astronomy and Space communities. Over the last few years there has been an increasing diversification both in the types of jobs the experiments expect to run and also in the hardware available to run jobs. It has proved very difficult to schedule jobs so they run efficiently on the correct hardware, while respecting the experiment fair shares and requiring minimum admin intervention. This talk describes our experiences over the last year, what we have tried and our future plans.

Desired slot length

Speaker release

Yes

Authors: DEWHURST, Alastair (Science and Technology Facilities Council STFC (GB)); ADAMS, James (STFC RAL); CABALLERO BEJAR, Jose (Science and Technology Facilities Council STFC (GB)); BIRKETT, Thomas (Science and Technology Facilities Council)

Presenter: DEWHURST, Alastair (Science and Technology Facilities Council STFC (GB))

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **10**   Type: **not specified**

The CMS Submission Infrastructure deployment

Thursday 23 September 2021 15:00 (30 minutes)

The CMS experiment at CERN requires vast amounts of computational power in order to process, simulate and analyze the high energy particle collisions data that enables the CMS collaboration to fulfill its research program in Fundamental Physics. A worldwide-distributed infrastructure, the LHC Computing Grid (WLCG), provides the majority of these resources, along with a growing participation from international High Performance Computing facilities. The combined processing power is harnessed for CMS use by means of a number of HTCondor pools operated by the CMS Submission Infrastructure team. This contribution will present a detailed view of our infrastructure, encompassing multiple HTCondor pools running in federation, aggregating hundreds of thousands of CPU cores from all over the world. Additionally, we will describe our High Availability setup, based on distributed (and in some cases replicated) infrastructure, deployed between the CERN and Fermilab centres, to ensure that the infrastructure can support critical CMS operations, such as experimental data taking. Finally, the present composition of this combined set of resources (WLCG, CERN, OSG and HPC) and their roles will be explained.

Desired slot length

Speaker release

Yes

Authors: PEREZ-CALERO YZQUIERDO, Antonio (Centro de Investigaciones Energéticas Medioambientales y Tecnológicas); MASCHERONI, Marco (Univ. of California San Diego (US)); HALEEM, Saqib (National Centre for Physics (PK))

Presenter: PEREZ-CALERO YZQUIERDO, Antonio (Centro de Investigaciones Energéticas Medioambientales y Tecnológicas)

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **11**   Type: **not specified**

Operations and Monitoring of the CMS HTCondor pools

Thursday 23 September 2021 15:30 (30 minutes)

The CMS Submission Infrastructure team manages a set of HTCondor pools to provide the vast amount of computing resources that are required by CMS to perform tasks like data processing, simulation and analysis. A set of tools that enables automation of regular tasks and maintenance of the key components of the infrastructure has been introduced and refined over the years, allowing the successful operation of this infrastructure. In parallel, a complex monitoring system that includes status dashboards and alarms have been developed, enabling this effort to be performed with minimal human intervention. This contribution will describe our technology and implementation choices, how we monitor the performance of our pools in diverse critical dimensions, and how we react to the alarms and thresholds we have configured.

Desired slot length

Speaker release

Yes

Authors: PEREZ-CALERO YZQUIERDO, Antonio (Centro de Investigaciones Energéticas Medioambientales y Tecnológicas); MASCHERONI, Marco (Univ. of California San Diego (US)); HALEEM, Saqib (National Centre for Physics (PK))

Presenter: HALEEM, Saqib (National Centre for Physics (PK))

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **12**

Type: **not specified**

What's new in HTCondor? What is upcoming?

Monday 20 September 2021 15:20 (35 minutes)

Presenter: TANNENBAUM, Todd (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **13**　　　　　　　　　　　　　　Type: **not specified**

What to measure and why

Monday 20 September 2021 15:55 (40 minutes)

Speaker release

Desired slot length

Author: LIVNY, Miron (CHTC, U Wisconsin-Madison)

Presenter: LIVNY, Miron (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **14**   Type: **not specified**

What new / upcoming for the HTCondor-CE?

Friday 24 September 2021 15:00 (30 minutes)

Speaker release

Desired slot length

Author: LIN, Brian (CHTC, U Wisconsin-Madison)

Presenter: LIN, Brian (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: 15   Type: **not specified**

Job Router Transforms

Friday 24 September 2021 15:30 (30 minutes)

Speaker release

Desired slot length

Authors: LIN, Brian (CHTC, U Wisconsin-Madison);   KNOELLER, John (University of Wisconsin--Madison)

Presenter: KNOELLER, John (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **16**   Type: **not specified**

Using SciTokens in HTCondor 9

Tuesday 21 September 2021 16:00 (35 minutes)

Speaker release

Desired slot length

Author: BOCKELMAN, Brian (CHTC, U Wisconsin-Madison)

Presenter: BOCKELMAN, Brian (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **17**   Type: **not specified**

Introducing the HTCondor 9.0 Series for Users

Tuesday 21 September 2021 15:30 (30 minutes)

Speaker release

Desired slot length

Author: KOCH, Christina (CHTC, U Wisconsin-Madison)

Presenter: KOCH, Christina (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: 18   Type: **not specified**

Introducing HTCondor 9.0 for Admins

Wednesday 22 September 2021 15:00 (40 minutes)

Speaker release

Desired slot length

Author: THAIN, Gregory (University of Wisconsin-Madison)

Co-author: FREY, James (University of Wisconsin Madison (US))

Presenter: THAIN, Gregory (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **19**   Type: **not specified**

Self-Checkpointing Jobs in HTCondor

Thursday 23 September 2021 16:00 (35 minutes)

Speaker release

Desired slot length

Author: KOCH, Christina (CHTC, U Wisconsin-Madison)

Presenter: KOCH, Christina (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **20**

Type: **not specified**

Campus Research and Facilitation

Friday 24 September 2021 16:45 (35 minutes)

Speaker release

Desired slot length

Author: MICHAEL, Lauren (CHTC, U Wisconsin-Madison)

Presenter: MICHAEL, Lauren (CHTC, U Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **21**
Type: **not specified**

How to write a custom file xfer plugin

Friday 24 September 2021 16:00 (30 minutes)

Speaker release

Desired slot length

Author: COATSWORTH, Mark (University of Wisconsin-Madison)

Presenter: COATSWORTH, Mark (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **22** Type: **not specified**

New GPU architectures MIGs, multiple jobs per GPU, etc.

Monday 20 September 2021 16:50 (30 minutes)

Speaker release

Desired slot length

Author: KNOELLER, John (University of Wisconsin-Madison)

Presenter: KNOELLER, John (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **23**　　　　　　　　　　　　　　　　　Type: **not specified**

Welcome and logistics

Monday 20 September 2021 15:00 (20 minutes)

Speaker release

Desired slot length

Presenter:　MEINHARD, Helge (CERN)

Session Classification:　Workshop session

Track Classification:　Miscellaneous

Contribution ID: **24**　　　　　　　　　　　　　　　　Type: **not specified**

Workshop wrap-up

Friday 24 September 2021 17:50 (10 minutes)

Speaker release

Desired slot length

Presenter:　MEINHARD, Helge (CERN)

Session Classification:　Workshop session

Track Classification:　Miscellaneous

Contribution ID: 25   Type: **not specified**

Q / A / Discussion: HTCondor Python Bindings

Thursday 23 September 2021 17:10 (20 minutes)

Prior to this Q&A session, if you are unfamiliar with the HTCondor Python Bindings, please work through the online Python Bindings tutorials. Use the Binder link at the following URL to launch an interactive Jupyter notebook in your web browser with the tutorials already loaded:

https://htcondor.readthedocs.io/en/v9_0/apis/python-bindings/tutorials/index.html

Speaker release

Desired slot length

Author: PATTON, Jason (University of Wisconsin-Madison)

Presenter: PATTON, Jason (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: 26   Type: **not specified**

Q / A / Discussion: Negotiator policy and configuration

Thursday 23 September 2021 17:30 (20 minutes)

Links to material to consult beforehand will be published here.

Speaker release

Desired slot length

Author: THAIN, Gregory (University of Wisconsin-Madison)

Presenter: THAIN, Gregory (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **27**   Type: **not specified**

Q / A / Discussion: Using IDTokens for authentication in HTCondor 9

Thursday 23 September 2021 17:50 (20 minutes)

Links to material to consult beforehand will be published here.

The attached slide deck was designed to (hopefully) be self-explanatory to an HTCSS administrator. Bring forth any questions to the discussion!

Speaker release

Desired slot length

Author: TANNENBAUM, Todd (University of Wisconsin Madison (US))

Presenter: TANNENBAUM, Todd (University of Wisconsin Madison (US))

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials

Contribution ID: **28**   Type: **not specified**

Open stage - Show Us Your Toolbox, followed by office hours

Wednesday 22 September 2021 17:05 (1 hour)

This session is intended to serve as an opportunity for administrators to show the audience how they do their work with and on HTCondor - what are the most useful tools for them to perform their work? Why are they so useful? What do they look (and feel) like?

In case of interest, the session could be split into breakouts at some point in time.

This session will not be recorded. We would appreciate a 'sanitized' (if needed) slide by the contributors for the records, though.

The "open stage" will be followed by breakouts for office hours - see the 'Videoconference' link in Indico for the links.

Speaker release

Desired slot length

Author: TANNENBAUM, Todd (University of Wisconsin Madison (US))

Presenter: TANNENBAUM, Todd (University of Wisconsin Madison (US))

Session Classification: Workshop session

Track Classification: HTCondor and HTCondor CE office hours

Contribution ID: **29**   Type: **not specified**

Synthetic populations for personalized policy

Tuesday 21 September 2021 16:50 (30 minutes)

Public policy design generally targets ideal households and individuals representing average figures of the population. However, statistics only make sense when referring to large numbers, less so when we are trying to represent real people belonging to the actual population. In fact, referring to the characteristics of the average citizen, the policy maker loses the capacity to represent the diversity of the population at large, negatively affecting minorities and under-represented people. Statistics over the population are usually given as univariate figures. Typically, knowing that e.g. in a certain area live 55% women and 30% university educated people do not give a high quality information for the distribution and we may actually misinterpret what the real issue is.

One way to improve the representation of the diversity is to recur to multivariate distributions in spatial modelling, e.g. creating high quality aggregates for specific use cases.

Using real data to give these representations poses important privacy concerns, because knowing the combination of features in certain areas might give away the identity of some citizens.

In recent years, the performances of supercomputers skyrocketed, and at the same time the access for data scientists to high performance computing technologies has been democratized, offering to policy makers the unprecedented opportunity for creating tailored policy using a completely synthetic population.

Policy simulation models can take as input synthetic individuals that resemble the actual ones but are stripped out of their identities, as they are synthetic by design. Synthetic individuals are created inter-linking census data, behavioural surveys and other available data sets and the result is a synthetic population with average statistics similar to the actual one by design, to the point that one is not able to tell if an individual belongs to the real or to the synthetic population, with the advantage of being relieved from most privacy concerns.

In this context, we have generated the synthetic population of France, based on Census data from INSEE (French Institute of Statistics and Economic Studies) and other data sets available.

The main data sets involved: information at individual level, such as age, sex, level of education, household composition, etc.; information at household level, such as sociodemographic characteristics as well as information about the dwelling and its location, characteristics, category, type of construction, comfort, surface area, number of rooms etc.; information about mobility to the workplace, including their main socio-demographic characteristics, as well as those of the household to which they belong; information about the mobility to education facilities.

Additional datasets included the map of the census tracks used by the INSEE, and data from the cadastre about properties cross linked with geographic data form the French Geographic Institute (IGN) and OpenStreetMap to create as detailed a map with the distribution of dwellings by type. Data about the location of educational establishments was extracted from the Ministry of Education, while the location of economic activities was obtained by cross referencing the data from INSEE which covers 64 different economic activities with the buildings for the OpenStreetMap database.

By linking the datasets above it was possible in the first instance to create families and households and then to attribute them to individual buildings. This combinatorial optimization is known as the Variable Size Multiple Knapsack Problem. This problem can be tackled in different ways, no solution is perfect but there is always a trade-off between precision and computational intensity. Aiming at a better precision is only possible when the input data adds up useful information. Sometimes, the least computationally intensive solutions offer reasonable results as well. In our case,

having any additional attribute to houses, e.g. year when built, would make people positioning much more precise. Another source of uncertainty is that, in the absence of better information, we assumed that larger families would inhabit larger housing surfaces, which is obviously not always the case.

Notwithstanding these limitations, we modelled the synthetic population of 63 million people, in 35 million households allocated in 10 million houses in France including their travel to work and study places behaviour. The computations were performed in batch processing on the JRC Big Data Analytics Platform (BDAP), that uses HTCondor as a job scheduler and Docker Universe set up.

Around 35k jobs were performed, one for each French commune, each job taking 1 CPU. At our disposal were 20 servers of 40 CPUs each and 1TB RAM, and relatively unlimited storage space. The machine set was shared with other users.

The scripts were in Bash and Python, including libraries such as Numpy, Pandas, geoPandas and Shapely.

One of the challenges was to deal with very large CSV files in input (e.g. one of 12GB). Opening these files (in Pandas) required that the memory demand in the Condor submit file had to be so large (˜200GB) that machines were seldom allocated to our jobs.

The idea was to subset from the large files only the records that belong to the job that is performing, so to make a query for a certain value (zip code processed by the job) along a certain column (zip code column), and subset only those lines that correspond to that query and save the result in a new CSV.

A benchmark of several libraries used for subsetting was performed and eventually the winner was AWK, offering the best speed and inferior memory requirement.

Speaker release

Desired slot length

Authors: DI LEO, Margherita; HRADEC, Jiri (Joint Research Centre)

Presenter: DI LEO, Margherita

Session Classification: Workshop session

Track Classification: HTCondor user presentations

Contribution ID: **31**   Type: **not specified**

Q / A / discussion: HTCondor philosophy and architecture

Thursday 23 September 2021 16:50 (20 minutes)

Links to material to consult beforehand will be published here.

Speaker release

Desired slot length

Presenter: THAIN, Gregory (University of Wisconsin-Madison)

Session Classification: Workshop session

Track Classification: HTCondor presentations and tutorials