HTCondor EU WS 2022 - Highlights

11-14 October 2022 (Cuneo, IT)

J. Flix / C. Beyer / T. Hartmann

PIC
port d'informació
científica

Workshop in Numbers

- **30 Attendees**
  - 20 from Europe (inc. 7 from Italy)
  - More biased towards experienced admins than past workshops
  - Travel is hard, many institutes only sending one person
  - Mix of new faces and old friends

- **18 Institutes from 12 Countries**

- The **Organizing Committee**: *Chris Brew, Christoph Beyer, Helge Meinhard, Todd Tannenbaum, Catalin Condurache, Antonio Puertas-Gallardo, Gabriele Fronzé, Greg Thain, Josep Flix, Michel Jouvin, Matthew West*

https://indico.cern.ch/event/1174979/

Great Venue!

AP creates Capabilities,
Ships them to servers
(with authorized services)
and ships them to client.
HT
PATh

Workshop group photograph

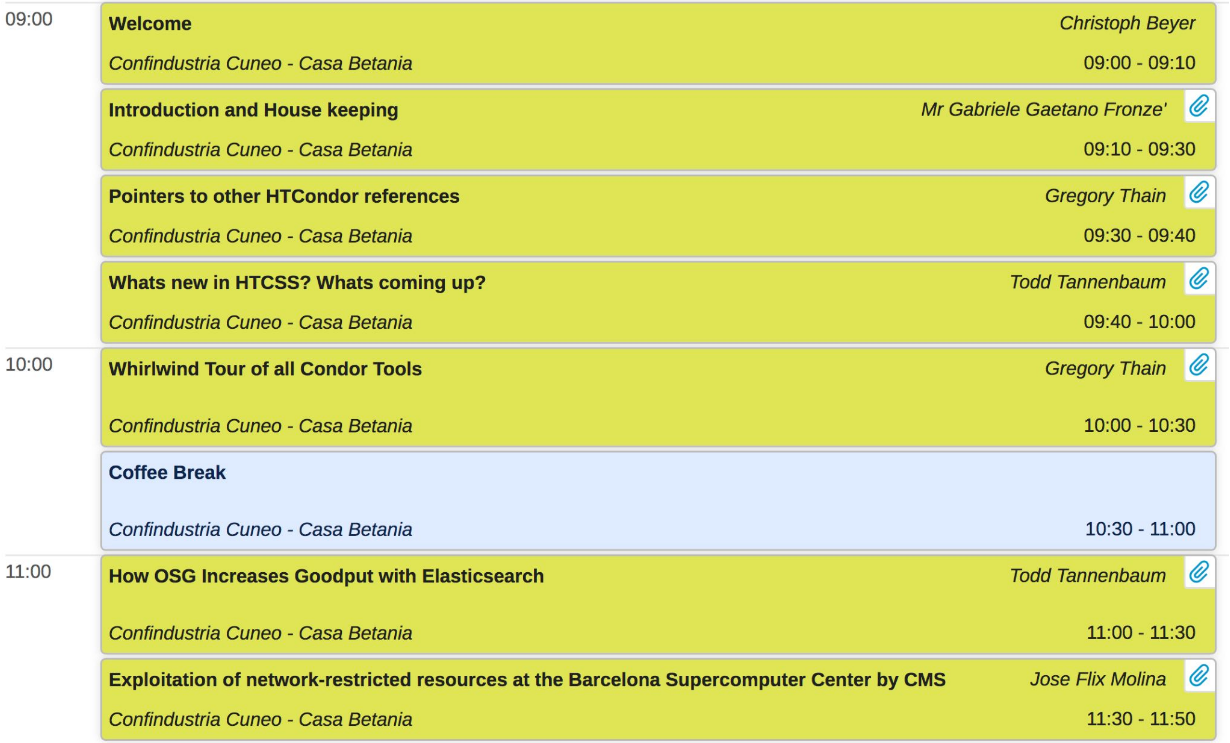
Presentations

- **41 Contributions by 15 Presenters**
  - Scheduled Duration 15h25m
    - HTC Team – 12h0m
    - Users – 3h25m
  - Discussion 1h25m (open) – lots included in the other sessions
  - Office Hours 50m
  - 30m Show and tell session

- Interesting **wide range** of talks, **high quality** presentations.

- **New style** of themed sessions

Tuesday morning



09:00	**Welcome**	*Christoph Beyer*	
	Confindustria Cuneo - Casa Betania	09:00 - 09:10	
	Introduction and House keeping	*Mr Gabriele Gaetano Fronze'*	
	Confindustria Cuneo - Casa Betania	09:10 - 09:30	
	Pointers to other HTCondor references	*Gregory Thain*	
	Confindustria Cuneo - Casa Betania	09:30 - 09:40	
	Whats new in HTCSS? Whats coming up?	*Todd Tannenbaum*	
	Confindustria Cuneo - Casa Betania	09:40 - 10:00	
10:00	**Whirlwind Tour of all Condor Tools**	*Gregory Thain*	
	Confindustria Cuneo - Casa Betania	10:00 - 10:30	
	Coffee Break		
	Confindustria Cuneo - Casa Betania	10:30 - 11:00	
11:00	**How OSG Increases Goodput with Elasticsearch**	*Todd Tannenbaum*	
	Confindustria Cuneo - Casa Betania	11:00 - 11:30	
	Exploitation of network-restricted resources at the Barcelona Supercomputer Center by CMS	*Jose Flix Molina*	
	Confindustria Cuneo - Casa Betania	11:30 - 11:50	

Tuesday morning highlights

- **Update** on Condor resources worldwide
- **New terminology: HTCondor** → **HTCSS** aka **HTC**ondor **S**oftware **S**uite
  - Schedd → Access Point [AP]
  - Startd → Execution Point [EP]
  - HTCondor Pool = Central Manager + Execution Point(s) [EP]
  - HTCondor Compute Entrypoint [HTCondor-CE - HTCE]
  - Stable series → LTS [Long-Term Support]
  - Developer series → feature release
- **New tool syntax:** htcondor ‹CMD› ‹VERB› ‹Options,Flags,...› → htcondor job status 123.45
- New in **9.0 LTS**
  - Condor_watch_q command gives more 'human readable' output
  - New tools and mechanisms to support jobs that checkpoint
  - Security model changes - by default secure now
  - Authorisation through tokens (IDTokens, SCITokens)
- New in **10.0**
  - ARM and PowerPC architectures support based on Alma Linux

Tuesday morning highlights

- General **HTCSS enhancements**
  - Can make "+CustomAttributes" first class with knobs
    - +SiteJobType = "analysis" → SiteJobType = analysis
  - New container universe (universe = container, container_image = /cvmfs/my/image/dir/)
  - GPU support enhanced (multi-usage and heterogeneous GPU environments)

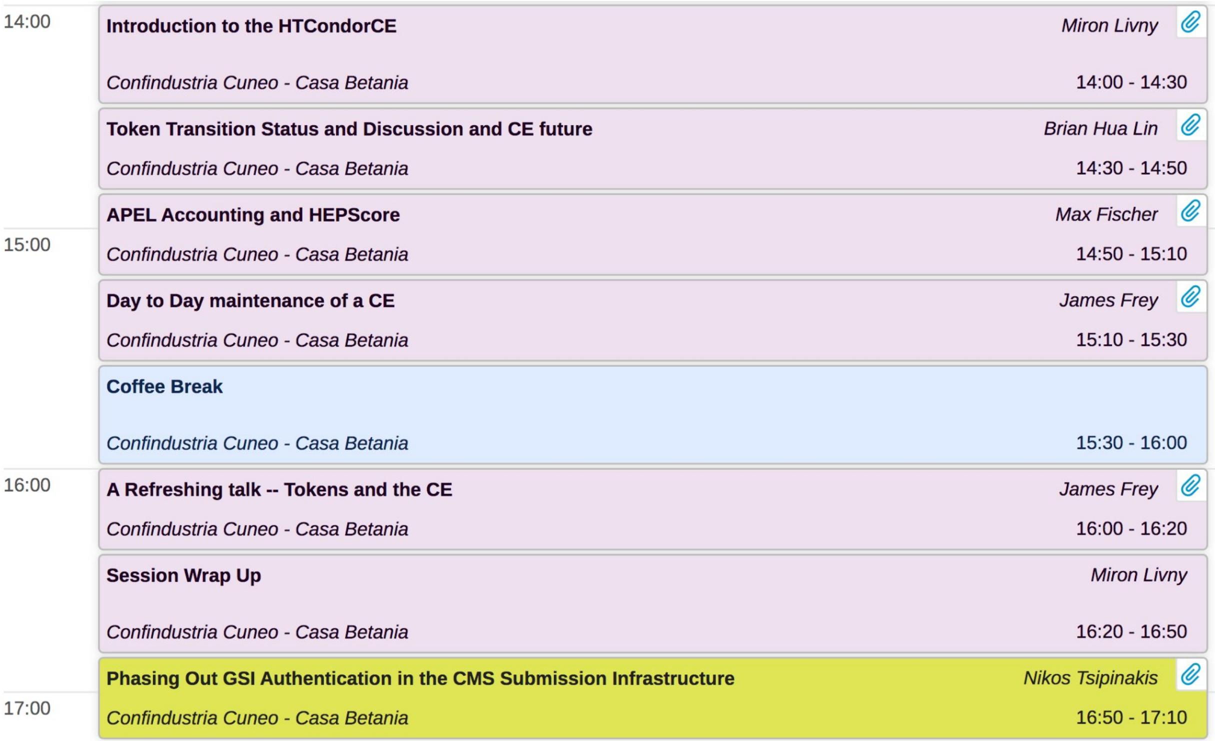
- "Whirlwind Tour of HTCondor Tools" replaced "Beginner tutorial" - great **overview** of 'all' Condor functions

- OSG increases goodput by analyzing **job history in Elasticsearch**
  - condor_adstash: writing completed job ads to ElasticSearch
  - Looking for users that have a bad time
  - Looking for broken sites (work in progress)
  - Looking for correct checkpointing (work in progress)

Tuesday morning highlights

- Communication of HTCondor elements via shared file system has been developed by the condor team and implemented at PIC Tier-1
- **Functionality** and **scale tests** have been run, linking the Barcelona Supercomputing Center (BSC) resources with the CMS (ITB) Global Pool
  - Simple workflow (no input): SIM workflow executed manually from the CMS ITB schedds and the outputs transferred into PIC storage using a Data Transfer Service (DTS) service deployed in PIC
- The infrastructure and developed services were proven capable to sustain a scale of **~15k CPU cores** in BSC's center (~10% of BSC!) and **500 MB/s aggregate output rate**
- This model needs to be **consolidated** and further **integrate** it into CMS WM and operations
  - More workflows would be suitable when the **DTS stage-in** feature will be commissioned

Tuesday afternoon



14:00	**Introduction to the HTCondorCE** *Confindustria Cuneo - Casa Betania*	*Miron Livny*	14:00 - 14:30
	Token Transition Status and Discussion and CE future *Confindustria Cuneo - Casa Betania*	*Brian Hua Lin*	14:30 - 14:50
	APEL Accounting and HEPScore *Confindustria Cuneo - Casa Betania*	*Max Fischer*	14:50 - 15:10
15:00	**Day to Day maintenance of a CE** *Confindustria Cuneo - Casa Betania*	*James Frey*	15:10 - 15:30
	Coffee Break *Confindustria Cuneo - Casa Betania*		15:30 - 16:00
16:00	**A Refreshing talk -- Tokens and the CE** *Confindustria Cuneo - Casa Betania*	*James Frey*	16:00 - 16:20
	Session Wrap Up *Confindustria Cuneo - Casa Betania*	*Miron Livny*	16:20 - 16:50
	Phasing Out GSI Authentication in the CMS Submission Infrastructure *Confindustria Cuneo - Casa Betania*	*Nikos Tsipinakis*	16:50 - 17:10
17:00			

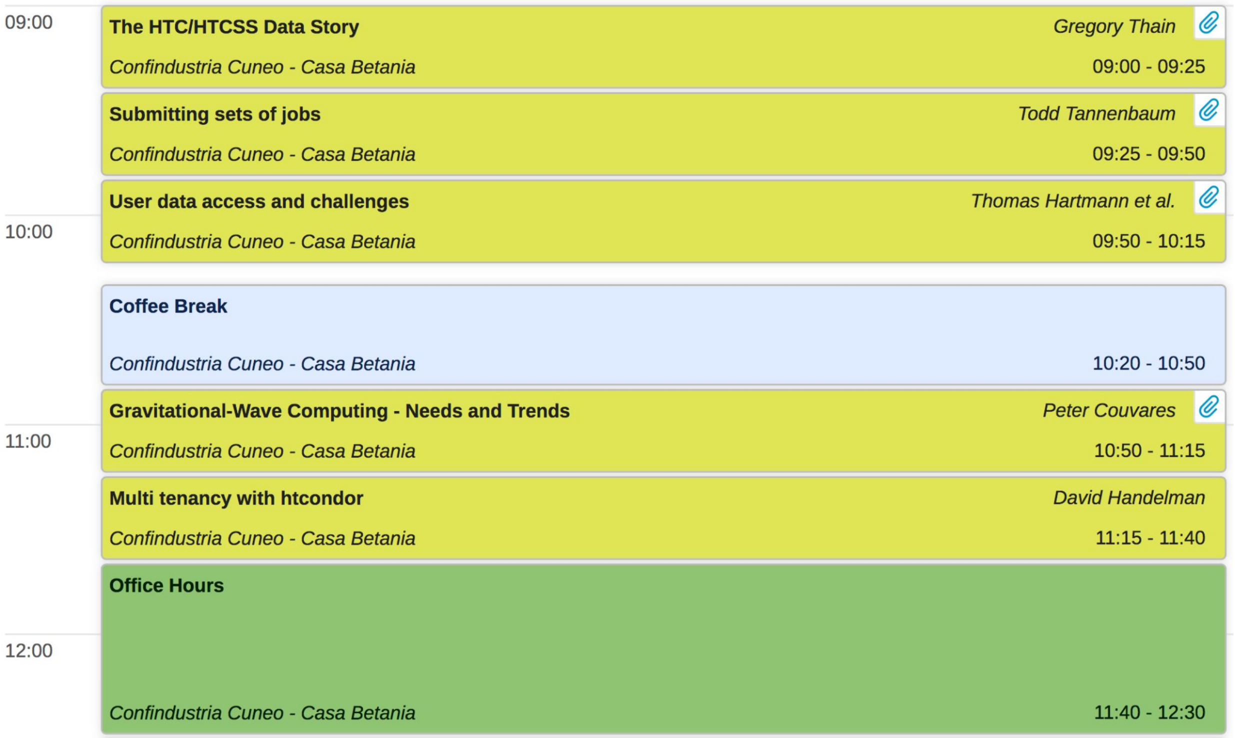
Tuesday afternoon highlights

- **Compute Entry Point (CE):** bridging the Grid into a HTCondor pool
  - New-style syntax in CE routes with more powerful macros and conditions (old syntax getting phased out in 10.1.x.?) → detailed info here
- **Token transition**
  - HTCondor 10.0.0 binaries will no longer depend on the Grid Community Toolkit
  - No more GSI authentication on the CE
    - X.509 proxy delegation still works, e.g. for use by the job for remote storage access
- Simple and advanced strategies on how to tweak **APEL accounting**
  - Support for multiple & future benchmarks and averaged and per node performances
- Day to Day Maintenance of an HTCondor-CE was a great roundtrip on how to administer and do the **house keeping on your HTCE**

Tuesday afternoon highlights

- HTCondor-CE doesn’t support **token refresh** - Does it need to? [discussion time]
- Moving to **tokens** in the **CMS Submission Infrastructure**
  - CMS Glideins are migrated to use IDTokens and SCITokens
  - Meeting industry standards
  - Retiring the Globus toolkit
  - Timeline: March 2023 - HTCondor GSI End Of Life
- Upcoming option to feed a WLCG token into Condor and get an **equivalent ersatz-token** that condor interprets as the original one
  - No need for a CE admin to have access to each supported VOs' IAMs
- **Session wrap up** - 30 minutes at the end of each day gave a lot of room for discussions and questions

Wednesday morning



09:00

10:00

11:00

12:00

The HTC/HTCSS Data Story *Gregory Thain*
Confindustria Cuneo - Casa Betania 09:00 - 09:25

Submitting sets of jobs *Todd Tannenbaum*
Confindustria Cuneo - Casa Betania 09:25 - 09:50

User data access and challenges *Thomas Hartmann et al.*
Confindustria Cuneo - Casa Betania 09:50 - 10:15

Coffee Break
Confindustria Cuneo - Casa Betania 10:20 - 10:50

Gravitational-Wave Computing - Needs and Trends *Peter Couvares*
Confindustria Cuneo - Casa Betania 10:50 - 11:15

Multi tenancy with htcondor *David Handelman*
Confindustria Cuneo - Casa Betania 11:15 - 11:40

Office Hours
Confindustria Cuneo - Casa Betania 11:40 - 12:30

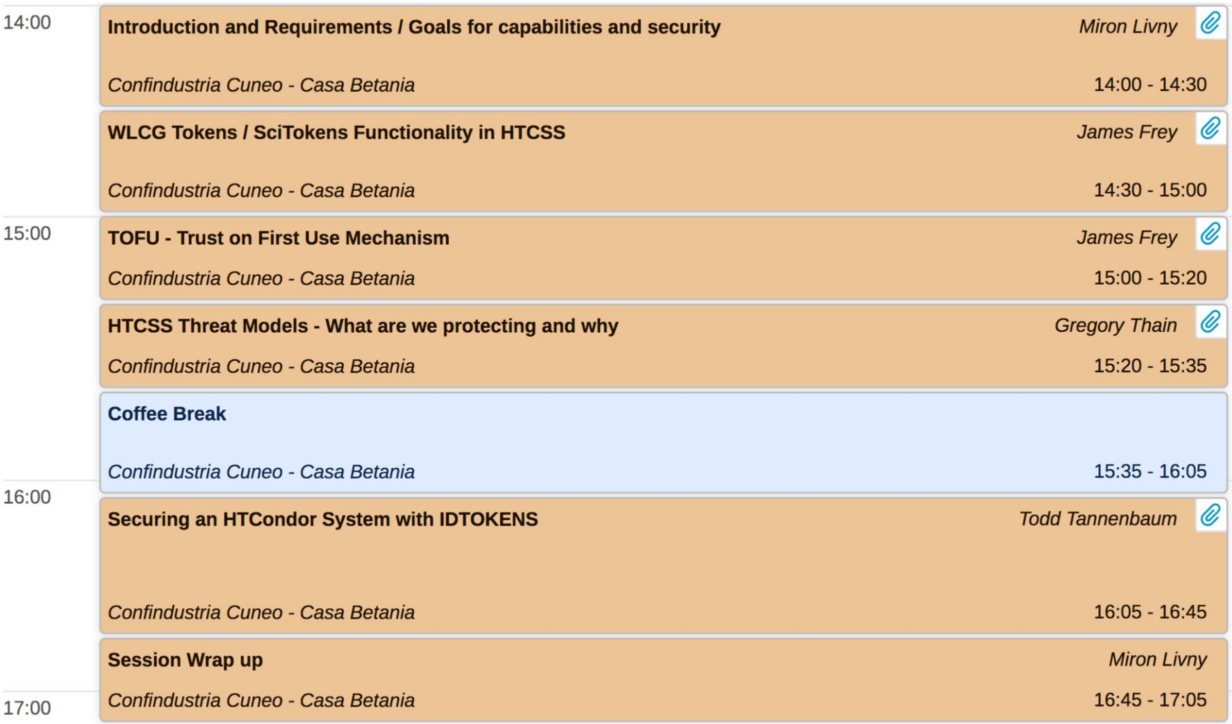
Wednesday morning highlights

- **Common Quests**
  - Shared file systems are implying unmanaged traffic
- **HTCondor Data Story**
  - Data access is one of the main pieces of the puzzle that adds up to a successful job especially in HEP environments
  - Explicit file transfer is a bit more effort to set up than shared FS
  - Built-in file transfer extendable by plug-ins means managed traffic
- **Submitting Multiple Jobs**
  - Job cluster consume much less resources on the AP and provide a single “handle” for the job management by the user and the admin alike
  - **New**: Jobsets, either defined in the submit file (jobset = XXX): “htcondor jobset ‹submit,list,...› set.def”
- **Job Monitoring Plans @ DESY**
  - DESY heavily uses shared FS (dCache, NFS, GPFS, CVMFS etc.)
  - Huge parallel file transfers are stressing APs and EPs
  - Often difficult to tell which job/user is causing issues
  - Moving for job profiling from cAdvisor to a job ad view (work in progress)
    - Discussion on job wrappers vs. sidecar jobs vs. ...

Wednesday morning highlights

- **Gravitational-Waves Computing** - Needs and Trends: covering what’s hard
  - Increasingly heterogeneous resource demands - tricky for capacity planning and scheduling
  - Balancing “low-latency/online” vs “batch/offline” HTC CPU demand
  - Lots of reactive, manual user priority tweaking - when an urgent science goal needs to get done faster
  - HTCondor has no concept of sites at all, but admins and users need to reason about them constantly
- **Multi tenancy with HTCondor**
  - United four condor pools, struggling with mix of fast response and lower prio jobs

Wednesday afternoon



14:00

Introduction and Requirements / Goals for capabilities and security — *Miron Livny*
Confindustria Cuneo - Casa Betania — 14:00 - 14:30

WLCG Tokens / SciTokens Functionality in HTCSS — *James Frey*
Confindustria Cuneo - Casa Betania — 14:30 - 15:00

15:00

TOFU - Trust on First Use Mechanism — *James Frey*
Confindustria Cuneo - Casa Betania — 15:00 - 15:20

HTCSS Threat Models - What are we protecting and why — *Gregory Thain*
Confindustria Cuneo - Casa Betania — 15:20 - 15:35

Coffee Break
Confindustria Cuneo - Casa Betania — 15:35 - 16:05

16:00

Securing an HTCondor System with IDTOKENS — *Todd Tannenbaum*
Confindustria Cuneo - Casa Betania — 16:05 - 16:45

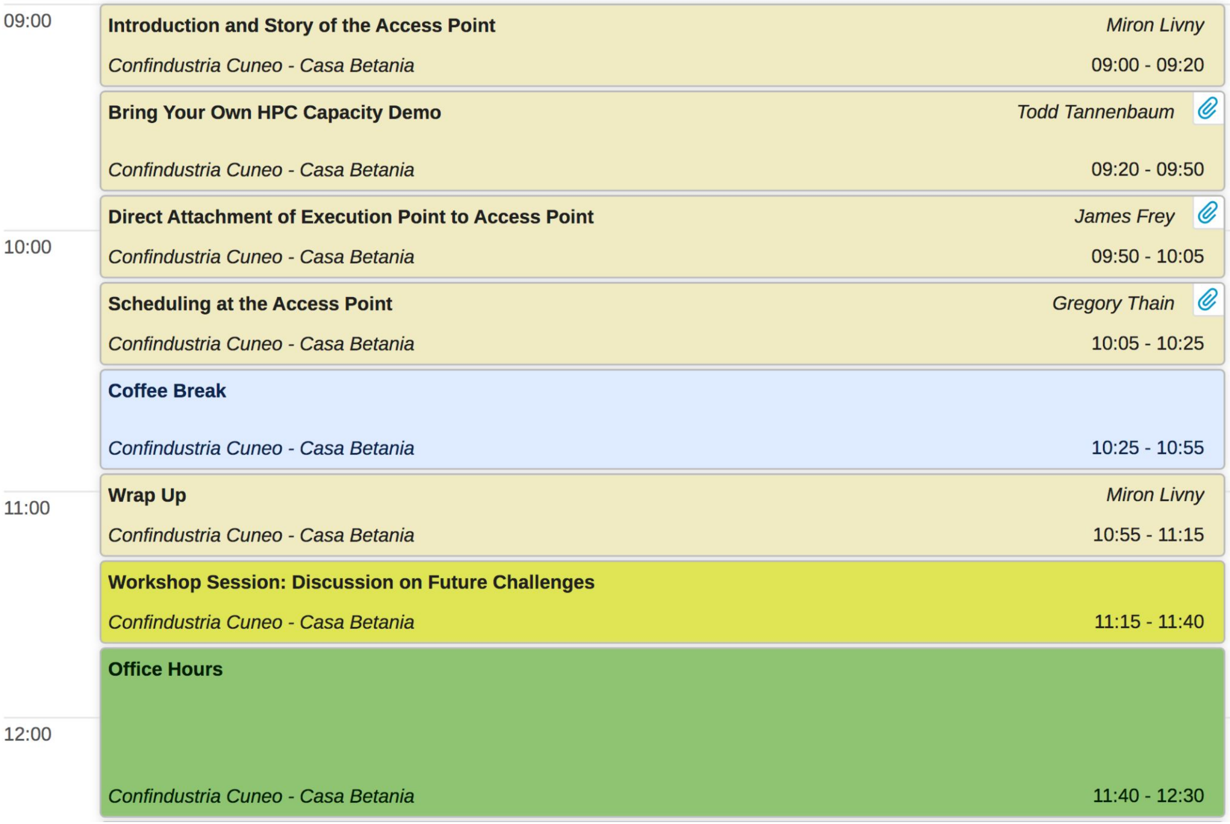
Session Wrap up — *Miron Livny*
Confindustria Cuneo - Casa Betania — 16:45 - 17:05

17:00

Wednesday afternoon highlights

- Introduction on **capabilities and tokens**
- **Tokens in HTCSS**
  - WLCG- & SciTokens: OpenID Connect derived JWTs describing file- and job-based authorizations
    - Asymmetric - anyone with issuer's public key can verify
  - HTCSS internal IDTokens are symmetric - Issuer's private key required to verify
    - Easy to deploy encrypted daemon to daemon authorization and communication
- **TOFU** = Trust On First Usage as part of the new Condor security model
  - New trust model for nodes + encryption by default
    - ssh-like approving on first sight (reinforced by pool idtokens)
  - Code will likely appear in 9.1
- Different **threat models** were discussed and how they influence the secure setup for a Condor pool

Thursday morning



Introduction and Story of the Access Point — *Miron Livny*
Confindustria Cuneo - Casa Betania — 09:00 - 09:20

Bring Your Own HPC Capacity Demo — *Todd Tannenbaum*
Confindustria Cuneo - Casa Betania — 09:20 - 09:50

Direct Attachment of Execution Point to Access Point — *James Frey*
Confindustria Cuneo - Casa Betania — 09:50 - 10:05

Scheduling at the Access Point — *Gregory Thain*
Confindustria Cuneo - Casa Betania — 10:05 - 10:25

Coffee Break
Confindustria Cuneo - Casa Betania — 10:25 - 10:55

Wrap Up — *Miron Livny*
Confindustria Cuneo - Casa Betania — 10:55 - 11:15

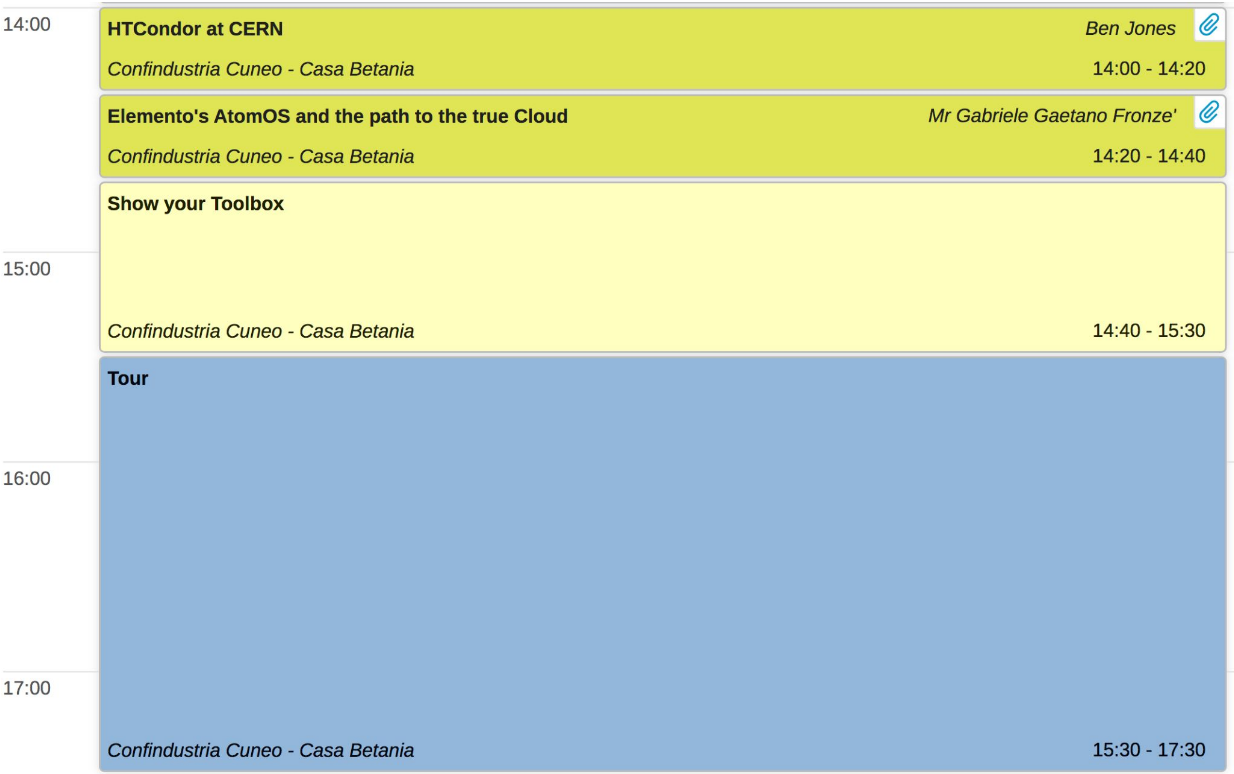
Workshop Session: Discussion on Future Challenges
Confindustria Cuneo - Casa Betania — 11:15 - 11:40

Office Hours
Confindustria Cuneo - Casa Betania — 11:40 - 12:30

Thursday morning highlights

- **HPC Annex** is a new self-service HPC system tool that enables user to easily submit into a HPC pool
  - Very interesting live demo on how to create the annex using 'htcondor annex create example debug@stampede'
- Direct Attachment of Execute Point to Access Point - a method of enabling user to 'bring your own capacity' into an existing Condor pool by attaching it directly to the AP
- **Scheduling at the Access Point** - the Distributed nature of HTCondor requires distribute identifiers that are not supported very good today but work in progress
- Workshop Session: Discussion on Future Challenges
  - Lively discussion about near future challenges the community is facing
  - Big concern power consumption (energy prices in Europe are skyrocketing)
  - Energy saving is a big thing (resources not in use should not consume any power)
  - Running a Condor pool below a given max energy consumption that might change over time will be necessary for some of us

Thursday afternoon



14:00

HTCondor at CERN — *Ben Jones*
Confindustria Cuneo - Casa Betania — 14:00 - 14:20

Elemento's AtomOS and the path to the true Cloud — *Mr Gabriele Gaetano Fronze'*
Confindustria Cuneo - Casa Betania — 14:20 - 14:40

Show your Toolbox

15:00

Confindustria Cuneo - Casa Betania — 14:40 - 15:30

Tour

16:00

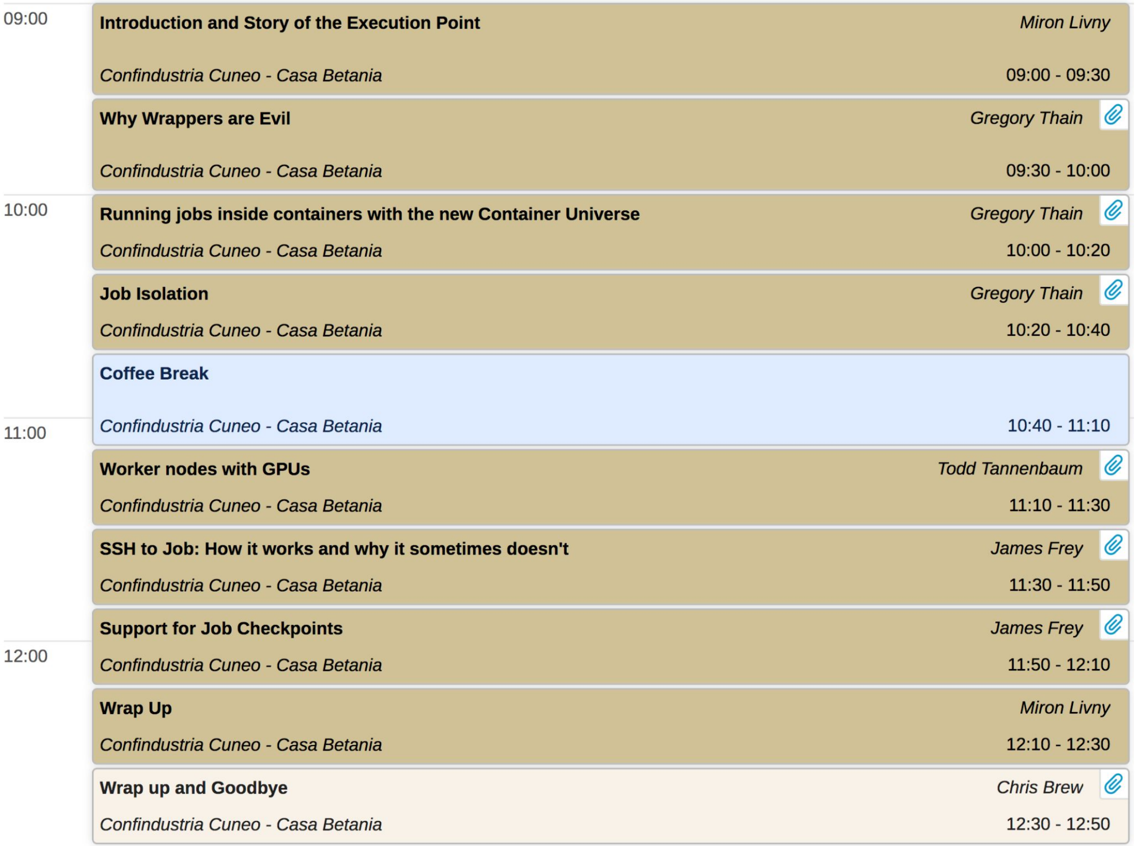
17:00

Confindustria Cuneo - Casa Betania — 15:30 - 17:30

Thursday afternoon highlights

- **HTCondor @CERN**
  - EL9 for next platform
  - ARM (some, maybe)
  - Move pool auth from GSI (probably to kerberos)
  - Interactivity / responsiveness expectations different with ”users” vs ”production”
  - User analysis is 10-20% of jobs, but 80% of support overhead
  - Increasing interest around “analysis facilities” and meta schedulers e.g. DASK
  - Faster response times for interactive & short jobs work in progress currently mostly reserved resources (some EP only start short jobs etc.)
- **Atomos - first distributed LINUX hypervisor**
  - Aggregating local and remote cloud resource in a common interface
- Show your toolbox was again a good success with different people off the audience showing some of their work and tools
  - People in the room shared screen through zoom session (!)

Friday morning

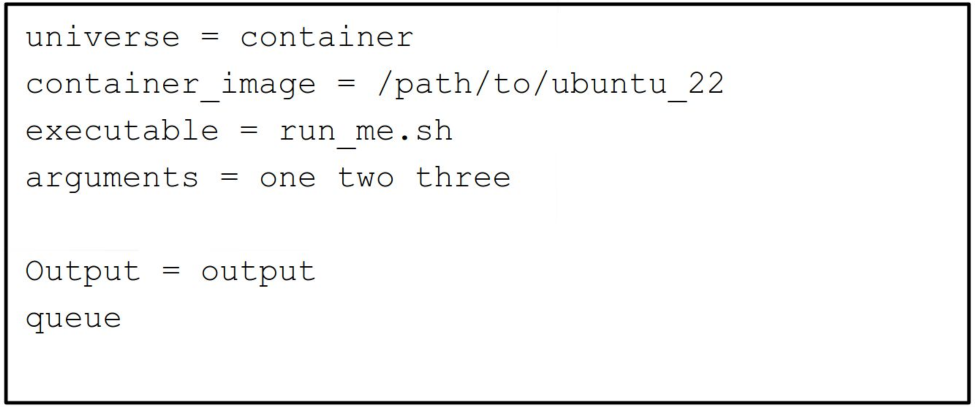


09:00	**Introduction and Story of the Execution Point**	*Miron Livny*	*Confindustria Cuneo - Casa Betania*	09:00 - 09:30
	Why Wrappers are Evil	*Gregory Thain*	*Confindustria Cuneo - Casa Betania*	09:30 - 10:00
10:00	**Running jobs inside containers with the new Container Universe**	*Gregory Thain*	*Confindustria Cuneo - Casa Betania*	10:00 - 10:20
	Job Isolation	*Gregory Thain*	*Confindustria Cuneo - Casa Betania*	10:20 - 10:40
	Coffee Break		*Confindustria Cuneo - Casa Betania*	10:40 - 11:10
11:00	**Worker nodes with GPUs**	*Todd Tannenbaum*	*Confindustria Cuneo - Casa Betania*	11:10 - 11:30
	SSH to Job: How it works and why it sometimes doesn't	*James Frey*	*Confindustria Cuneo - Casa Betania*	11:30 - 11:50
	Support for Job Checkpoints	*James Frey*	*Confindustria Cuneo - Casa Betania*	11:50 - 12:10
12:00	**Wrap Up**	*Miron Livny*	*Confindustria Cuneo - Casa Betania*	12:10 - 12:30
	Wrap up and Goodbye	*Chris Brew*	*Confindustria Cuneo - Casa Betania*	12:30 - 12:50

Friday morning highlights

- Job Wrappers are evil - Problems and Alternatives
  - Job wrapper scripts hide job exit codes in one way or another
  - HTCondor then can't differentiate a setup/cleanup problem from a bona-fide job problem (and we want to treat these differently)
  - Alternative move the wrapper logic into the submit file

- The new Container Universe
  - Successor off Docker Universe
  - Users just ask for a container and image



```
universe = container
container_image = /path/to/ubuntu_22
executable = run_me.sh
arguments = one two three

Output = output
queue
```

Friday morning highlights

- Job Isolation
  - Protecting jobs from each other, machines from jobs, jobs from machines
    - Namespace isolations and cgroup resource controls
  - cgroups V1 are supported, V2 work in progress
    - glide-owned job sub-cgroups upcoming

- GPU support in HTCondor
  - On the EP: ‘use feature:gpus’
  - New in HTCSS 10 -> support for different GPUS in one EP, NVIDIA Multi-Instance GPU (MIG)

- SSH_TO_JOB
  - Sometimes hard to debug, firewalls, container, job wrapper are common issues

- Self-Checkpointing Jobs with HTCondor
  - preserving progress through interruptions - especially for longer-running jobs
  - Checkpoint: Periodically write state to a file on disk
  - Restart: Code can both find the checkpoint file and can resume from it
  - Talk with link to examples, longer version and demo

Conclusions & Questions

- Very **good feedback** from attendees
- Seems like the most **productive meeting** so far in this series
- The **development** in the ‘batch-system-world’ is rather **gaining pace** than slowing down
- **New challenges** not only in terms of higher data rates everywhere
- Higher **electricity prices** do not help and need additional attention
- Will try to continue the series of workshops as **in-person-meetings**