AI-DAYS@HES-SO 2025 – GENEVA & LAUSANNE – JANUARY 27- JANUARY 29, 2025

WORKSHOP DAY JANUARY 27

Session 1: Compute Infrastructures for IA applications in the wild

Location: Movie theater 6

With the advent of Chatbots, LLMs and other generative IA technologies, as well as other progresses in the IA field, there is an explosion of the demand for compute force. IA is no longer computer science: it is computational science. As such, it can no longer be done with casual, self-managed equipment. More advanced compute infrastructures are required both to satisfy user needs (in terms of compute power, GPU Ram capacity) and to ensure a decent utilization of the increasingly costly resources.

Content and topics

The purpose of this workshop is to gather people in charge of compute infrastructure (on-prem, cloud or hybrid) destined to support AI workloads (both training and inference). Being "in charge" means someone who does typically performs one or more of the tasks below

- Prepares and follows purchase orders for servers including GPUs.
- Racks the servers, ensure they are connected with adequate network / IO performance
- Install the OS, the drivers, some software stack
- Binds the servers with the login system, typically LDAP
- Monitor the utilization of the server over time
- Ensure there are no abuse
- Insert the server into a cluster (e.g. SLURM, K8s)
- Designs and implement cluster architecture
- Helps AI engineers to run their workloads according to the policies
- Provisions cloud GPUs resources
- Manages cloud platforms accesses
- Secures the cluster
- Dimensions the cluster and plan its evolution over the time

The workshop is organized around keynote presentations, but above all around **short presentations (~18min) during which participants will present one of the following** (not limited to) :

- A technology under test or underutilization in their group (e.g. code, Determined.ai, etc.)
- A picture of how workloads are mapped onto hardware in their group
- A particularly painful aspect of IA-infra there are experiencing in their group
- A project for improving the situation in their group

At the end of the day, a panel/group discussion is planned to discuss how to become more efficient at these tasks going forward.

The workshop has no specific registration, and walk-ins are welcome.

Workshop committee

Last Name	First Name	Institution	e-mail address
Rumley	Sébastien	HEIA-FR	sebastien.rumley@hefr.ch
Gambin	Dorian	HEIG-VD	dorian.gambin@heig-vd.ch
Stadelmann	Marc	ZHAW	marcandre.stadelmann@zhaw.ch
Kucharavy	Andrei	HE-VS	andrei.kucharavy@hevs.ch
Menetrey	Jämes	UNINE	james.menetrey@unine.ch
Marques Reis	Henrique	HE-Arc	henrique.marquesreis@he-arc.ch

AI-DAYS@HES-SO 2025 – GENEVA & LAUSANNE – JANUARY 27- JANUARY 29, 2025

WORKSHOP DAY JANUARY 27

Schedule: 8h30-12h20 + 13h15-17h15

8:30:00 AM	0:05	Opening remarks	Sébatien Rumley	HEIA-FR, HES-SO / Swiss Ai center
8:35:00 AM	0:30	The Alps research infrastructure at CSCS: enabling world-class ML research in Switzerland	Fawzi Mohamed	The Swiss National Supercomputing Centre (CSCS), ETH Zurich
9:05:00 AM	0:18	SCITAS: On-premise and Cloud Infrastructure driving HPC & AI Scientific Computing at EPFL	Gilles Fourestey	SCITAS/EPFL
9:23:00 AM	0:18	Picterra's Infrastructure: Scaling ML for Geospatial Imagery Analysis	Julien Rebetez	CTO @ Picterra
9:41:00 AM	0:19	Securing AI Infrastructure: Strategies & Common Pitfalls	Terry Vogelsang	Kudelski Security
10:00:00 AM	0:30	Is your infrastructure ready for AI workloads ?	Jean-Baptiste Thomas	Principal Field Solutions Architect – Pure Storage EMEA
10:30:00 AM	0:20	Break - Offered by the Swiss Ai Center		
10:50:00 AM	0:18	Enabling ressources optimisation thru Monitoring of Mixed GPU Setups	Martin Roch-Neirey	HEIA-FR, HES-SO
11:08:00 AM	0:18	Enabling headache-free heterogeneous GPU resource sharing for a small cluster	Abele Mälan, Romain De Laage	UNINE
11:26:00 AM	0:18	Efficient GPU Resource Sharing with Kubernetes and Coder	Dorian Gambin	HEIG-VD
11:44:00 AM	0:18	Unlocking Performance: vGPU Acceleration on Cisco UCS X-Series with VMware vSphere	Jérémy Gamba	HEFR
12:02:00 PM	0:18	GPU infrastructure at UNIL, an attempt at continuous adaptation	Emmanuel Jeanvoine	UNIL - Scientific Computing and Research Support unit
12:20:00 PM	0:55	Lunch - Offered by PureStorage and Exoscale		
1:15:00 PM	0:30	GPU hyperspecialisation	Antoine Coetsier	Exoscale
1:45:00 PM	0:18	Building a SLURM Cluster with Existing Heterogeneous Hardware	Marc Stadelman	Centre for Artificial Intelligence, ZHAW School of Engineering
2:03:00 PM	0:19	Slurm & heterogenous users: avoiding main pitfalls	Ljiljana Dolamic	armasuisse S&T, CYD Campus
2:22:00 PM	0:19	OpenOnDemand	Adrien Albert, Yann Sagon	UNIGE
2:41:00 PM	0:19	Where to ask for help when building or operating a GPU cluster?	Marco Merkel	HPC-DoItNow
3:00:00 PM	0:30	Break - Offered by E4		
3:30:00 PM	0:20	OAR: a Versatile Resource and Job Management System to Tame Complexity	Olivier Richard	Polytech Grenoble - INP, UGA
3:50:00 PM	0:30	Private LLM for industrial users	Daniele Cremonini	E4
4:20:00 PM	0:05	Mini-break		
4:25:00 PM	0:50	Panel/group discussion (TBC)		
5:15:00 PM		End		