EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

ML Cloud

The Cluster of Excellence "Machine Learning - New Perspectives for Science" at the University of Tübingen offers a position as

HPC Systems Engineer – Machine Learning Infrastructure & Operations (m/f/d, E13 TV-L, 100%)

The position is available in the team of the Machine Learning Science Cloud and is permanent.

About Us

The ML Cloud team designs, runs and safeguards high-performance computing infrastructure tailored to machine learning workloads. Our state of the art systems are distributed across four data centers and are operated for the Tübingen AI Center, the Cluster of Excellence 'Machine Learning: New Perspectives for Science' and the Hertie Institute for AI in Brain Health. Every day, researchers unleash thousands of compute jobs on our systems, training frontier-scale neural networks and running experiments.

The Opportunity

Our clusters keep growing – more GPUs, more storage, and more services around them. We are looking for an HPC Systems Eingineer who enjoys developing and running production infrastructure end to end: from bare-metal compute and the Slurm scheduler to the virtualization layer on which central service run, the storage backends and the automation that holds it all together. You will keep our systems fast and reliable, extend them as demand grows, and turn manual work into reproducible, code-driven operations that the whole team can rely on. You will be joining a motivated team of HPC specialists who share responsibility for the whole stack, work closely, and have real room to shape how our infrastructure is built and run. You will also have the opportunity to shape two new directions: AI agents in our administrative workflows, and agent-native access to the clusters for our users.

Your Tasks

- Design and operate HPC clusters across four data centers – spanning inter-/intra-site networks, SLURM scheduling, parallel filesystems (e.g., Lustre/Ceph), and GPU accelerators – ensuring high availability and sustained throughput for research workloads.
- Build and manage provisioning and configuration of heterogeneous compute, storage and network nodes (e.g. image-based provisioning, node lifecycle, Layer 2 and Layer 3 networking)

- Run and extend our Proxmox virtualization clusters hosting the management plane and infrastructure services (HA, live migration, backup)
- Operate and scale our storage platforms: Ceph clusters for S3-compatible object storage for research data, including capacity planning, tiering, replication and backup
- Automate operations as code with Ansible and Git-based workflows – provisioning, configuration, roll-outs and recurring maintenance
- Build and operate monitoring, logging and alerting, and keep systems patched, hardened and reliable through regular maintenance windows
- Plan and carry out hardware roll-outs, capacity extensions and migrations together with vendors and data center staff
- Bring AI agents into our admin workflows and build agent-native access to the cluster for users.
- Support and advise users on efficient cluster usage (job scheduling, data handling, storage layout, access workflows)

Your Profile

- Masters degree in Computer Science or a related field
- Solid Linux administration experience in a production environment (RHEL/Rocky/Ubuntu)
- Hands-on HPC experience: Slurm, parallel file systems (Weka, Lustre, Ceph), GPU workloads (driver/CUDA, DCGM, Fabric Manager, NVLink/NVSwitch, MIG) and high-speed networks (InfiniBand, 400G Ethernet)
- Solid data-center networking skills: Ethernet/IP (VLANs, routing, bonding/LACP), InfiniBand fabric management (subnet manager, fabric diagnostics), and firewall/VPN configuration
- Experience operating Ceph and S3-compatible object storage (e.g. RADOS Gateway) at scale
- Experience with virtualization for the management plane and infrastructure services (Proxmox, including clustering and backup)
- Strong automation skills with Ansible (roles, inventories, CI-driven roll-outs) and scripting in Bash and Python
- containers (Apptainer/Singularity/Docker), monitoring stacks (Prometheus/Grafana), or database and service operations
- A plus: Hands-on LLM tooling (tool calling and APIs, agent frameworks, retrieval and evaluation)
- Independent, structured working style and good communication in English; German is a plus
- A collaborative, user-facing mindset – comfortable supporting and advising researchers

What We Offer

- Technically deep, architecturally open work that directly enables cutting-edge machine-learning research
- A small, senior team with flat hierarchy
- Flexible working hours and the option to work partially from home
- Working in an English-speaking, international team of HPC experts
- Professional development, conference attendance and real influence on our roadmap

Interested? We'd love to hear from you!

Disabled candidates will be given preference over other equally qualified applicants. The University of Tübingen is committed to equal opportunities and diversity and actively promotes gender equality. The position is divisible.

Please submit your application, including CV and relevant transcripts by December 6, 2026 via our application portal careers.tuebingen.ai. For questions, please contact Simon Kreuzer at simon.kreuzer@uni-tuebingen.de. The position is available from now on. Hiring is done by the Central Administration of the University of Tübingen.