WWU
MÜNSTER

Machine Learning and GPU Accelerated Visualization with JupyterHub

Dr. Markus Blank-Burian

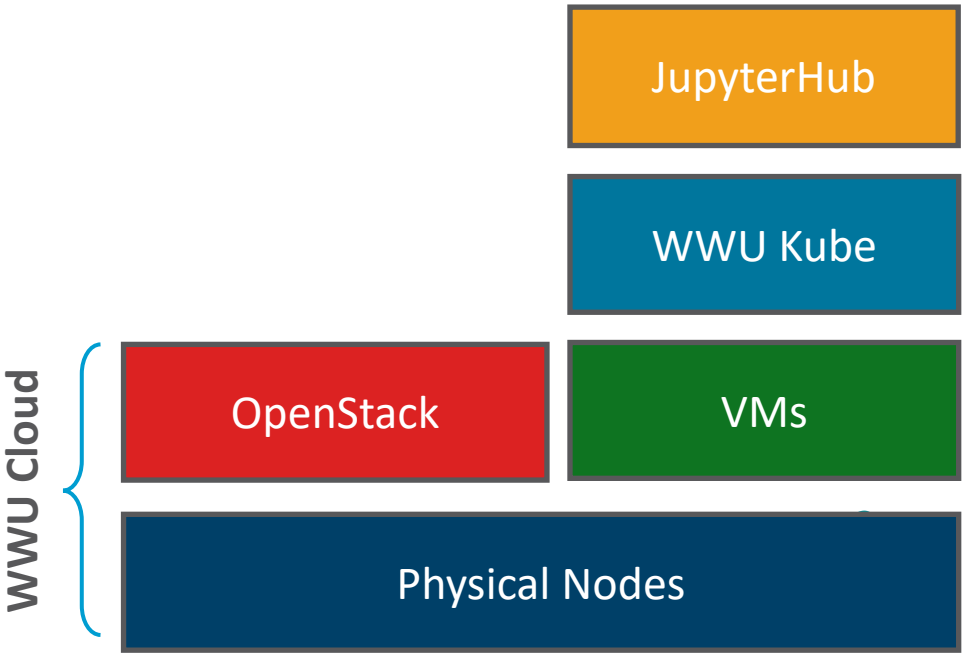
IT Department, University of Münster, Germany (WWU IT)

living.knowledge

IT

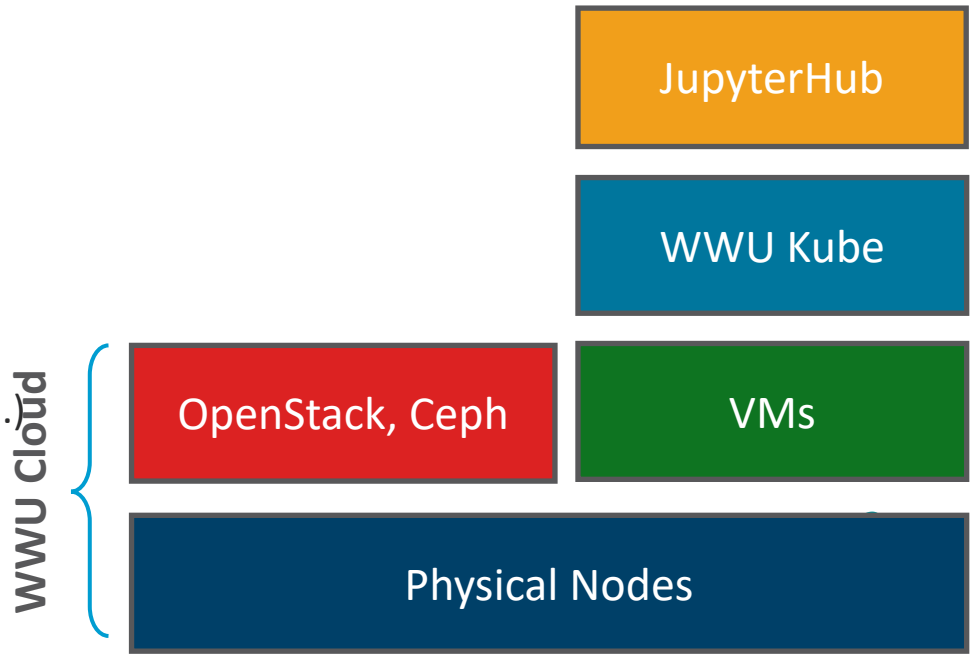
WWU Cloud: Private IaaS Cloud

- “Cloud” like in AWS, GCP, Azure, Alibaba: Infrastructure as a Service (IaaS)
- Users can create virtual infrastructure through Website or API
  - Virtual machines, disks, networks, router, network shares
- Intended primarily for services due to high cpu overcommit ratio
- Currently free of charge for employees
  - “Projects” must be requested with a detailed description and desired quotas



WWU Cloud: Technical Details

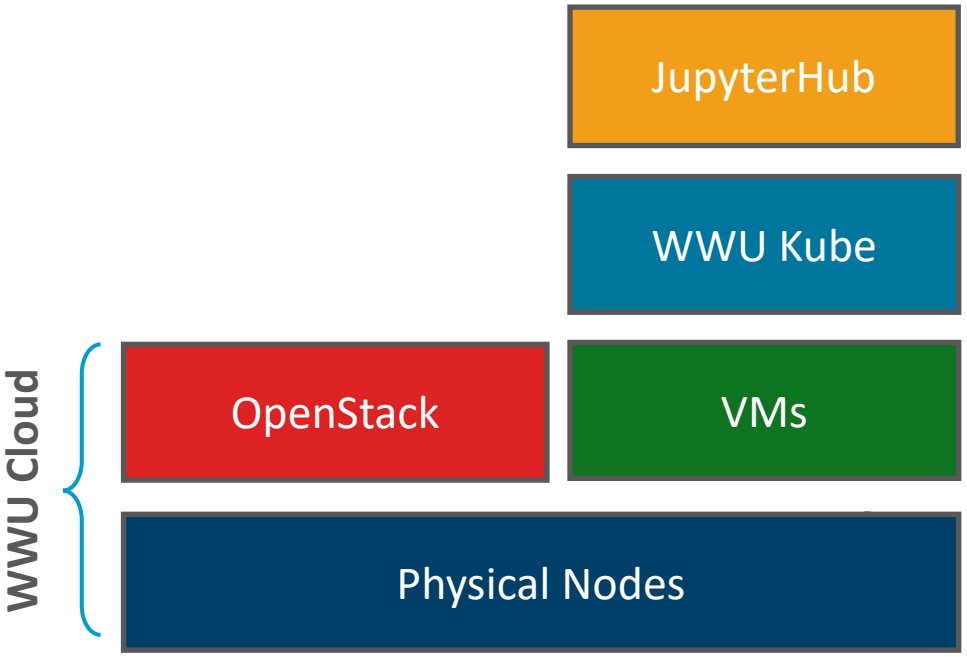
- Hyperconverged hardware: No distinction between storage and compute nodes => optimal resource usage
- Ceph as block storage for volumes (3x replicated)
- CephFS as shared filesystem for large data (8+3 ErasureCoding)
- OpenStack as cloud environment for managment of virtual hardware
- OVN on top of OVS as virtual network provider for OpenStack
- Kubernetes as container orchestration engine
  - Coordinates all services (Ceph, OpenStack, OVN, Databases, Logging, Metrics, ...)
- All services are highly available => no single point of failure
- Node rollout via ClusterAPI + Metal3.io (currently transitioning from Ansible)
- Multiple data centers as independent regions



WWU Kube: Multi Tenant Kubernetes Cluster

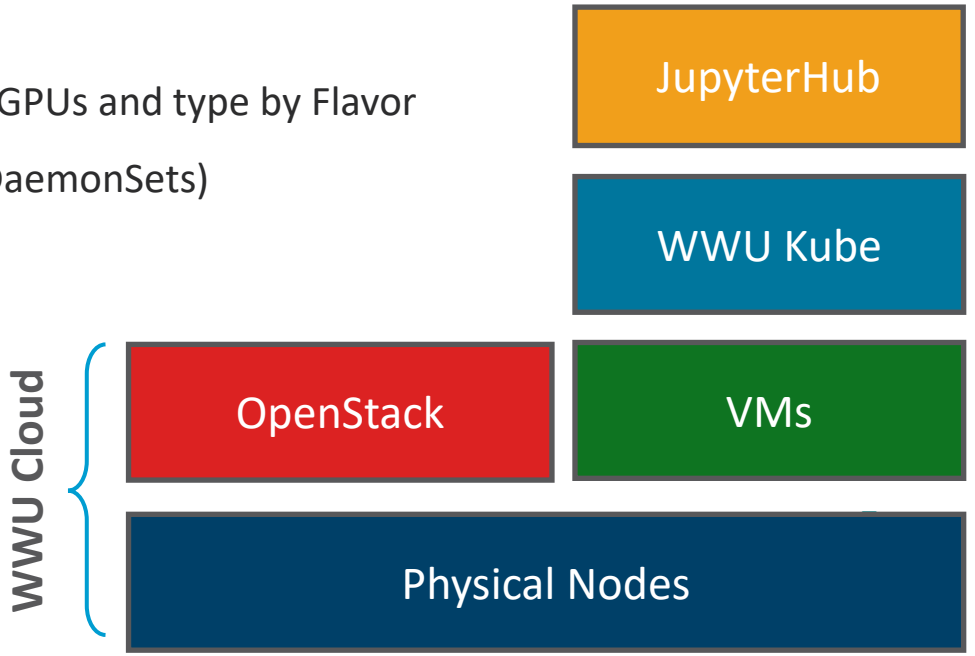
- Large multi tenant cluster, operated by WWU IT
- Hosted inside VMs from WWU Cloud
- Intended for service hosting
- Management of node updates, platform updates, security by WWU IT

  => Service administrators can focus on applications
- Technical Details:
  - Cilium networking: Fast and secure
  - Istio: Service mesh for advanced networking and security (mTLS, Routing)
  - Falco: Security via eBPF
  - Monitoring: Prometheus + Thanos, Alertmanager, Grafana, Loki

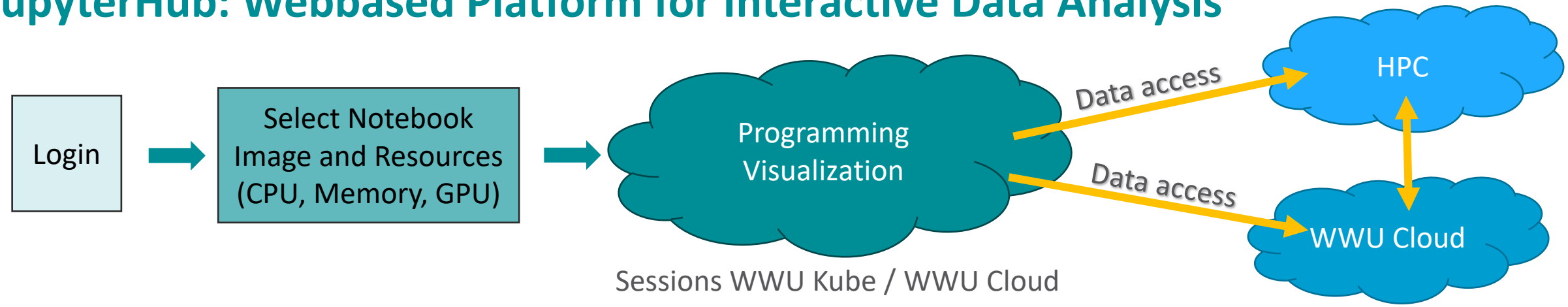


GPUs in WWU Cloud and WWU Kube

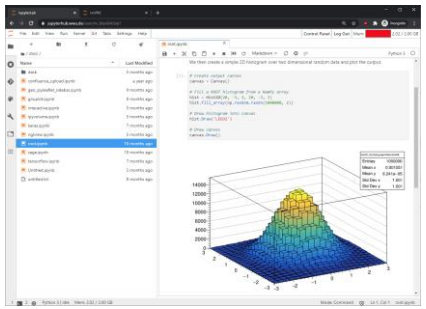
- Some nodes with virtualized GPUs (M10, T4) via NVIDIA Grid
  - Special drivers and licenses needed for GPU virtualization
- Nodes are annotated (resource traits) with GPU type => VMs can select number of GPUs and type by Flavor
- WWU Kube detects GPU types in VMs and labels nodes automatically (via NVIDIA DaemonSets)
- Pods can specify GPU type via resource requests and nodeSelector
- Cluster automatically creates new VMs with GPUs on demand (cluster autoscaler)
- OpenStack Cyborg (planned):
  - Allows creation of multiple virtual GPU types on single GPU
  - Optimize resource utilization together with cluster autoscaler



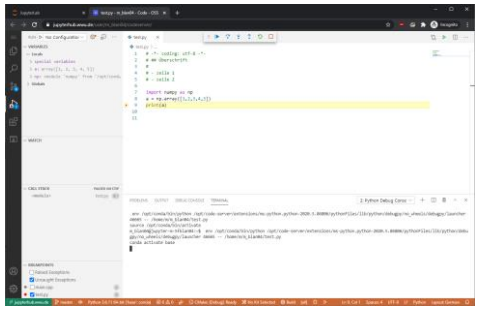
JupyterHub: Webbased Platform for Interactive Data Analysis



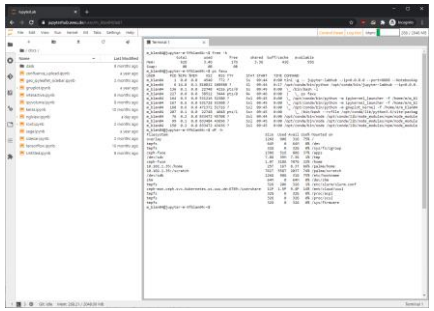
Sessions WWU Kube / WWU Cloud



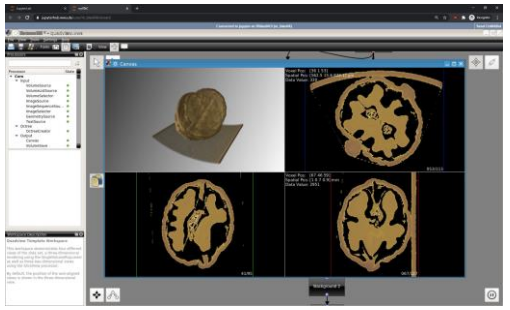
Notebooks



IDE



Terminal

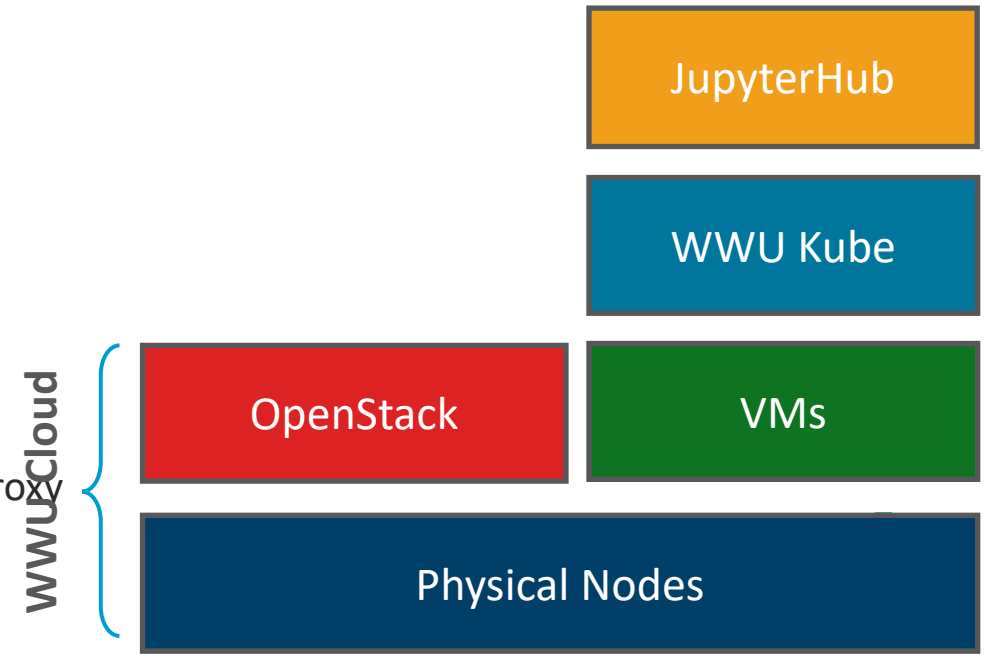


X11 Applications
with OpenGL

GPUs in JupyterHub

- Native access to CUDA within Jupyter session
- ML libraries with CUDA support pre-installed: Tensorflow, Keras, Torch

  => Notebooks can accelerate ML via libraries
- CUDA programs can be written and tested within VS Code IDE (code-server)
- X11 acceleration via OpenGL + VirtualGL + noVNC
- Technical Details:
  - VirtualGL uses single framebuffer for offscreen rendering
  - Websockets (for noVNS / web applications) are tunneled via jupyter-session-proxy



Thank you for your attention!