Hybrid Cluster NEMO

HPC and Virtualization

Chair in Communication Systems

Computing Center

University of Freiburg

2015

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Konrad Meier

konrad.meier@rz.uni-freiburg.de

Overview

UNI FREIBURG

1. bwForCluster ENM (NEMO)
2. Motivation
3. Hybrid HPC
4. Virtualization Concept
5. OpenStack Challenges
6. Virtualization for ATLAS
7. Summary

1. bwForCluster ENM (NEMO)

UNI FREIBURG

- Located at Freiburg University, Computer Department
- Will be available Q2 2016 with > 10000 cores
- Prototype installation (codename "NEMO") started late 2014 as a testbed with 1248 cores
- OpenStack is deployed as a Infrastructure as a Service (IaaS) solution
- Shared by 3 diverse scientific user groups:
  Elementary Particle Physics, **N**euroscience, **M**icrosystem Engineering
- 19 Physics research groups in Baden Württemberg
  - 8 x KIT Karlsruhe
  - 5 x University of Freiburg
  - 4 x University of Heidelberg
  - 2 x University of Tübingen

1. Physics Research Groups

KIT Karlsruhe	Prof. Dr. Günter Quast, Prof. Dr. Thomas Müller, Prof. Dr. Ulrich Husemann, Prof. Dr. Wim deBoer	Institut für experimentelle Kernphysik
KIT Karlsruhe	Priv.-Doz. Dr. Thomas Kuhr	Institut für experimentelle Kernphysik
KIT Karlsruhe	PD Dr. Stefan Gieseke	Institute for Theoretical Physics
KIT Karlsruhe	Prof. Dr. M. Margarete Mühlleitner	Institute for Theoretical Particle Physics
KIT Karlsruhe	Prof. Dr. Frans R. Klinkhamer	Institute for Theoretical Physics
KIT Karlsruhe	Prof. Dr. Ulrich Nierste	Institute for Theoretical Particle Physics
KIT Karlsruhe	Prof. Kirill Melnikov	Institut for Theoretical Particle Physics
KIT Karlsruhe	Prof. Dr. Matthias Steinhauser	Institute for Theoretical Particle Physics
University of Freiburg	Prof. Dr. Stefan Dittmaier	Theoretical Particle Physics and Quantum Field Theory
University of Freiburg	J.Prof.Dr. Harald Ita	Quantum Fields and Particle Phenomenology
University of Freiburg	Prof. Dr. Markus Schumacher	Experimental Particle Physics
University of Freiburg	Prof. Dr. Karl Jakobs	Experimental Particle Physics
University of Freiburg	Prof. Dr. Gregor Herten	Experimental Particle Physics
University of Heidelberg	Prof. Dr. André Schöning	Experimental Physics
University of Heidelberg	Prof. Dr. Hans-Christian Schultz-Coulon	ATLAS group
University of Heidelberg	Prof. Dr. Tilman Plehn	Institut für Theoretische Physik
University of Heidelberg	Prof. S. Hansmann-Menzemer, Prof. U. Uwer	Physikalisches Institut
University of Tübingen	Prof. Dr. Werner Vogelsang	Institute for Theoretical Physics
University of Tübingen	Prof. Dr. Barbara Jäger	Institute for Theoretical Physics

2. Motivation

- Traditional “bare metal” paradigm:

  “Maximum Usage of Resources”

- Problems:
  - Complex software environments are expensive to provide
  - Conflicting requirements: CentOS 6.5 ↔ CentOS 7.1
  - Software maintenance:
    - Software is often provided by the cluster-operator
    - Software environment is not fix and can change without the user knowing
- Solution:
  - Virtualization provides separation between hardware resources and software environment

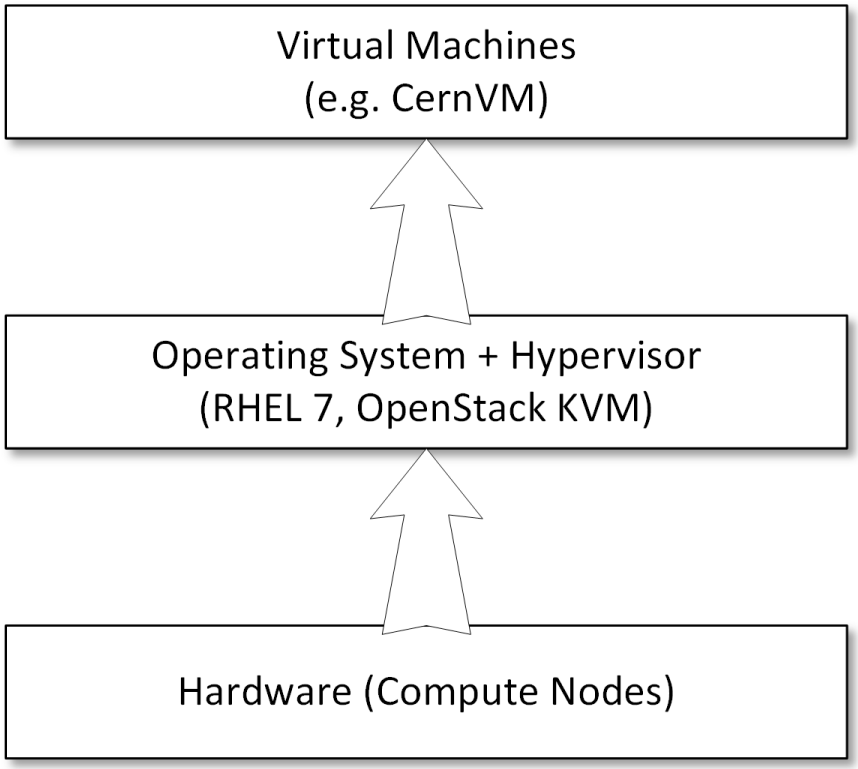
3. Hybrid HPC in NEMO

UNI FREIBURG

- Provide classic HPC (“bare metal”) and virtualization on the same cluster
- No hardware partitioning between virtualization and classic HPC nodes
- OpenStack as virtualization management framework
- Allow users to provide own VM images with required software stack
- VM scheduling is integrated into HPC scheduler
- Implemented as part of the bwForCluster NEMO

3. Hybrid HPC

- Layer Model:



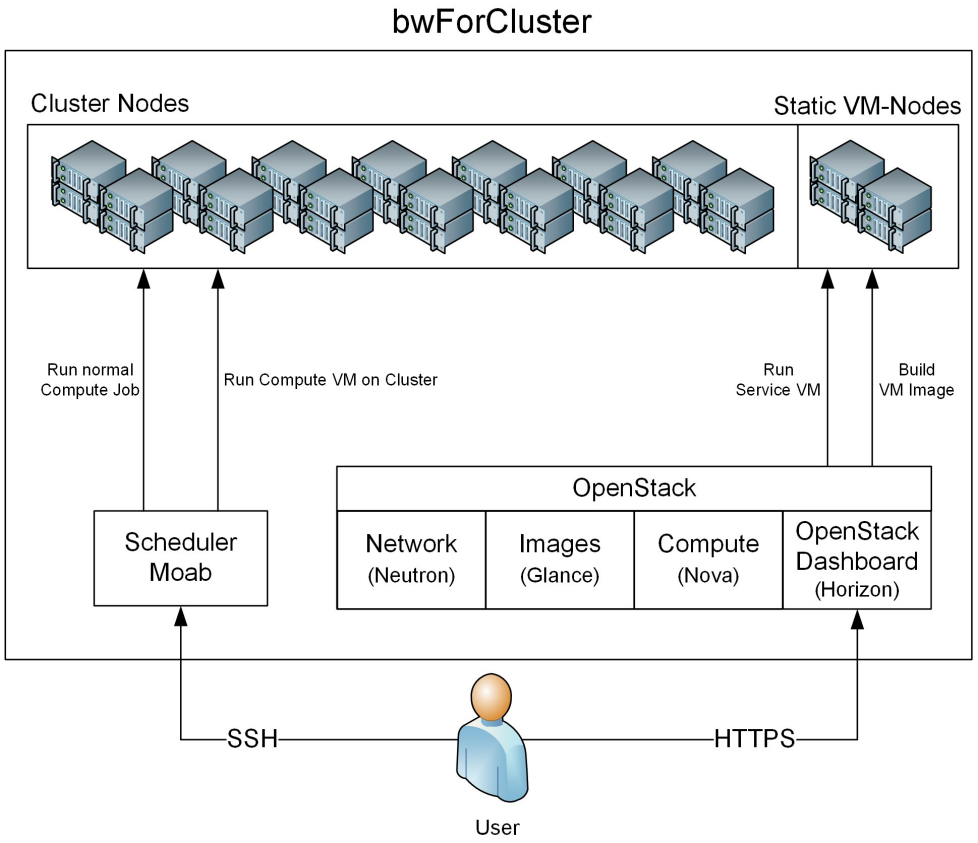
Virtualization provides:
- Virtual Research Environments

Bare metal provides:
- Resources for “classic HPC”
- Direct hardware access (Infiniband)

Provisioning:
- PXE Boot
- DNBD copy on write

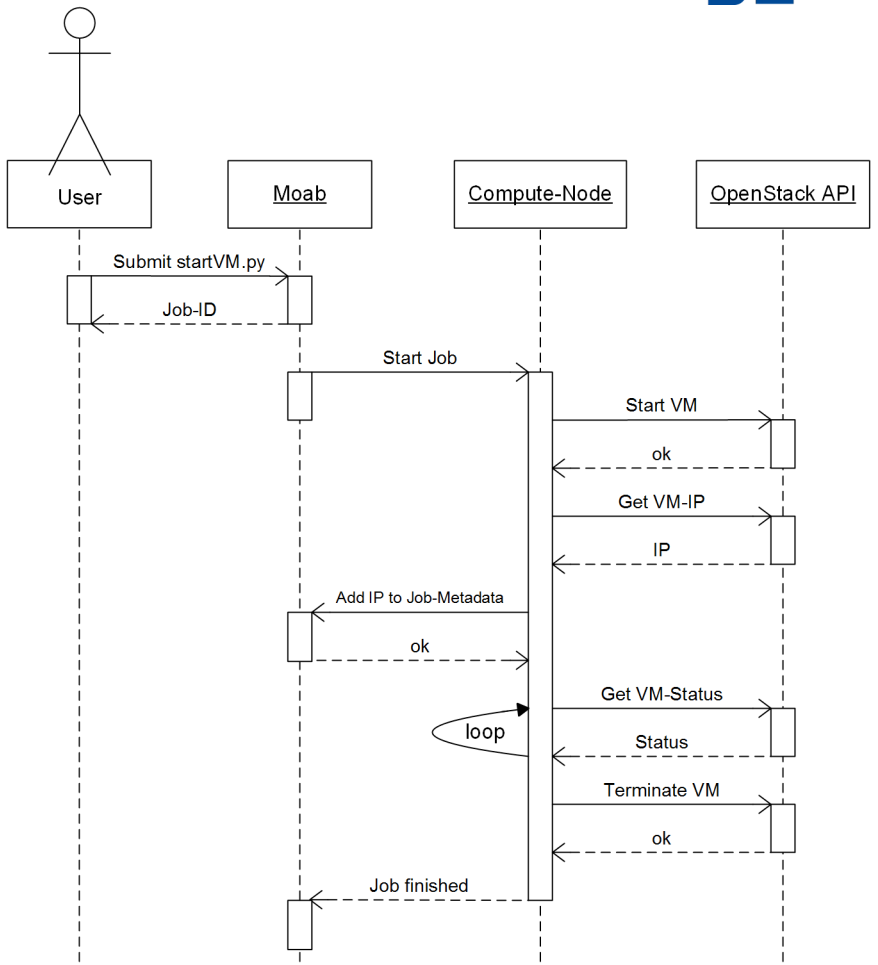
4. Virtualization Concept

- Interactive „Static VM-Nodes“ to run:
  - Build VM Images
  - Cluster services (e.g. Monitoring)
- Graphical OpenStack Webinterface provides easy access for testing/debugging and VM image creation
- For computation, VM images are started via the standard job submission procedure



4. Virtualization Concept

- Integration into Scheduler
  - The integration is transparent to the scheduler
  - A VM is like any other cluster job
  - User can monitor and control the VM with standard scheduler tools (Job state, VM IP, cancel job)
  - Accounting and Fairshare are working



5. OpenStack Challenges

- Storage
  - Shared storage does not scale
    → use local disks for VMs
  - Image must be copied to the node
    → provide images over NFS mount
- Network in NEMO
  - New cluster network concept was required
    - The cluster network is not a black box
  - Private 10.x.0.0/16 Networks in VLANs
  - OpenStack Neutron configuration:
    - Neutron ml2 + linux bridge plugin
  - NAT over hardware firewall (FortiGate 1500D)
    - Max throughput 80 Gbit/s

5. OpenStack Challenges

- Scaling 500+ VMs
  - High load on management servers: neutron, nova
    - Timeouts if load is to high
      → new management Servers for nova, neutron
      → Increase number of worker threads to 64
  - High number of TCP Connections
    - Again: Timeouts and Errors
    - → Increase the number of allowed connections for: rabbitmq, mariadb, neutron, nova
  - Neutron Security Groups implementation not designed for high number of VMs in one Project

    → Disable Security Groups (Firewall via Hardware Firewall)

6. Virtualization for ATLAS

- Complex Tier 2 and Tier 3 structure with dCache in Freiburg
- Idea: extend the Tier 2 resources with VMs
- First successful tests with a static number of VMs
  1. VMs are started manually by an administrator
  2. VMs are booting via iPXE the ATLAS-System
  3. VMs are automatically integrated into Slurm
- VMs successfully running for 4 weeks (400 Cores total)
- ToDo:
  - Dynamic allocation of resources (no static VMs)
  - New Virtual Machine Image (not PXE-Boot)

7. Summary

- Hybrid Cluster Model provides
  - Efficient resource usage for classic HPC
  - Virtual Research Environments
  - Platform for certified software stacks (e.g. CernVM)
- OpenStack integration requires:
  - Requires dedicated servers for management
  - Network planing
  - Storage planing
  - Detail optimization for scaling
- Future challenges: Dynamic Cluster Computing
  - Dynamically add virtual resources to the cluster