Karlsruhe Institute of Technology

System Architecture Group
http://os.ibds.kit.edu/

Steinbuch Centre for Computing (SCC)
http://www.scc.kit.edu/

Towards High-performance Cloud Computing for x86/InfiniBand Clusters

Marius Hillenbrand, Viktor Mauch, Jan Stoess

Limits of today's HPC clusters

- Administrative model: single-organization only
- Execution model: specialized HPC jobs
- Runtime model: Pre-defined, inflexible OS and libs
- Fluctuating demand; physical resources are underutilized or overloaded

Opportunity: HPC as a Service

Virtualized HPC Architecture based on Cloud model promises …

- Flexibility to serve multiple users and applications
- Individual HPC environment on-demand
- Fully automated resource allocation
- Cost savings, "Pay as you go" – principle

State of the Art

- Mainly GNU/Linux operating systems (>98% in **Top500**)
- Low-Latency interconnects: Ethernet (Gigabit 53%, 10G 2%); **InfiniBand** (43%) with **OS-bypass** protocol offloading
- Process-based job management

New Virtualization Challenges

- Handle interconnect interfaces and addressing schemes
- Preserve low latency; required for performance and scalability
- Provide standardized runtime (Linux) w/o additional OS noise
- Offer virtual **HPC clusters** instead of (single) virtual machines

Challenge: Virtualize cluster interconnects to provide HPC cloud computing

Approach

Node Virtualization

- Start with traditional KVM-based virtualization layer
- Examine lightweight OSs and virtualization layers (e.g. Kitten/Palacios)
- Develop novel OS layer with low-latency communication for virtual environments
- Focus on jitter and noise reduction

Network and Topology Virtualization

- Preserve advanced interconnect features like RDMA
- Implement effective isolation between customers through partitioning
- Offer advanced configurability and flexibility restricted to a customer's partition
- Virtualize and isolate management interfaces
- Provide recursive configurability

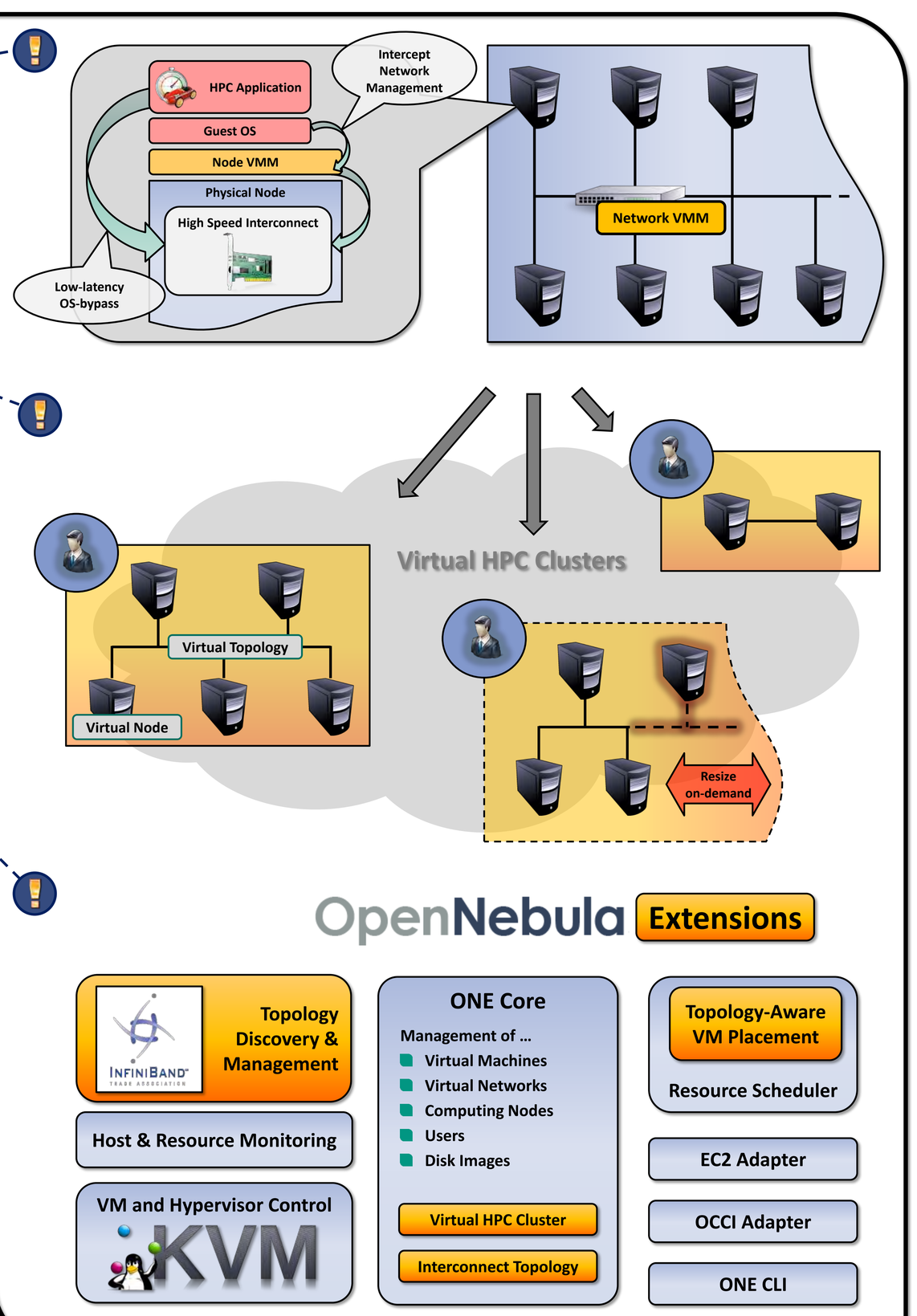
HPC Cloud Management

Offering

- Clusters of VMs instead of single VMs
- Performance guarantees
- Elasticity to consume exactly the HPC capacity required
- Dynamic pricing models and non-HPC workload to control utilization

Implementation

- Extensions to existing cloud computing framework **Open Nebula**
- HPC interconnect topology as a first-class abstraction
- Topology-aware VM placement
- Interfaces to Job Management Systems to exploit elasticity



Outlook

- **First HPCaaS Prototype** based on x86 computing nodes, KVM virtualization and InfiniBand network interconnect
- ***Mellanox*** ConnectX-2 IB HCAs provide SR-IOV functionality for multi-tenancy on single hosts (beta status)
- Management by the cloud computing framework ***OpenNebula*** (with extensions)

- Is slack stealing feasible to boost utilization?
- Will OS noise improve with lightweight virtualization layers?
- Will low-latency comm and RDMA enable cluster-wide scheduling?