Karlsruhe Institute of Technology

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

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

HPCaaS - High Performance Computing as a Service

Status and Outlook

Viktor Mauch, Marcel Kunze, Jan Stoess, Marius Hillenbrand

Introduction

What is High Performance Computing (HPC)?

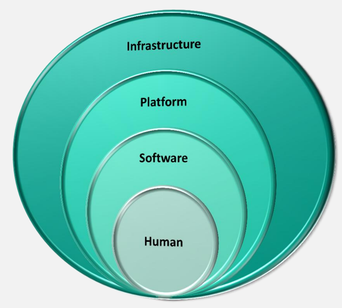
- HPC uses computer clusters to solve advanced computational problems
- Operation Area:
  - Parallel computing (MPI Jobs)
  - Data-intensive, distributed application (thousands of nodes, petabytes of data)
- Strong requirements concerning computing power, storage, and (particularly for parallel computation) communication networks

INFINIBAND™ TRADE ASSOCIATION

- Typically: **InfiniBand** Fabrics are deployed, > 60% of the Top 100 supercomputers
  - High bandwidth, up to effective 32 Gbit/s (between nodes)
  - Low latency, < 1µs
  - Future-proof development and outlook
  - Supported by most IT vendors: Intel, IBM, Cisco, Oracle, Voltaire, Mellanox, QLogic, …

What is Cloud Computing?

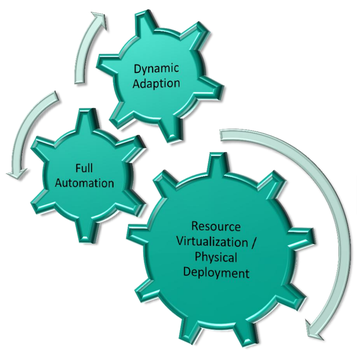
- Abstracted IT resources and services on-demand over the internet
- Dynamically adapted to the needs of the customers
- Settlement depends on usage, only actually used resources / services must be paid
- Combination of virtualized computing infrastructure and management via web-based services
- Fully automated system with a minimum of maintenance and costs
- Illusion of unlimited resources, available anytime
- *"Everything" as a Service (XaaS)* philosophy:
  - *IaaS*: virtual / physical computing resources
  - *PaaS*: development / execution environment
  - *SaaS*: Applications, Server Services
  - *HaaS*: manpower on-demand



Motivation for HPCaaS

Traditional HPC Architecture has restrictions:

- **Is characterized by very specific computing clusters designed for one or just a few special applications**
- **Has pre-defined operating systems and user environments**
- **Serves a single application at a given time**
- **Provides restricted user accounts**
- **Depends on the maintenance of the administrators**



Solution: Concept of HPCaaS

- **Clustered servers and storage as resource pools**
- **Fully automated allocation**
- **Individual cluster configuration on-demand**
- **Flexibility to serve multiple users and applications**
- **Customers have full administrative rights over the provided infrastructure**

Challenge: Provide InfiniBand Support for automated systems to deliver HPC cloud computing services!

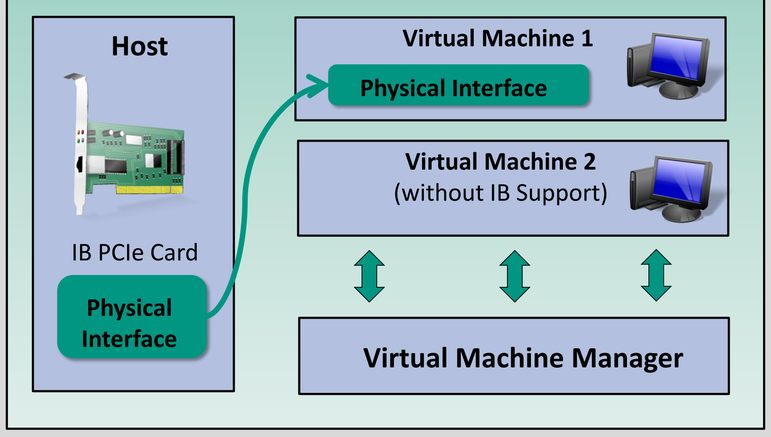
Spectrum of Technical Solutions

- **Limits of Software-only I/O virtualization:**
  - **Increased I/O latency:** VMM must process and route every data packet and interrupt, leads to higher application response time
  - **Scalability limitations:** software-based I/O processing consumes CPU cycles, reduces the processing capacity
- **Solution I: PCI Pass-Through**
  - VT-d (Intel) / IOMMU (AMD) chipset specification allows to pass-through a IB PCIe Adapter to single VM
  - VMM does not have to manage I/O traffic
  - Direct access with native performance
- **Solution II: Single Root - I/O Virtualization**
  - Extension to the PCI Express specification suite
  - Physical I/O resources are virtualized within the PCIe card, each card presents multiple virtual I/O interfaces
  - Almost native performance
  - Virtual Functions (VFs):
    - Provide all the functionality which is necessary for communication
    - VM interfaces directly with a VF without VMM intervention
  - Physical Function (PF):
    - VMM interfaces with PF to configure and manage I/O resource sharing among the multiple VMs
- **Workaround: Physical Resource Deployment**
  SINA - https://savannah.fzk.de/projects/sina
  - User-friendly web frontend
  - Controls the PXE server setup
  - Manages computing nodes, user accounts and install routines
  - Provides user functionality to allocate nodes, reboot them and deploy specific operating system install routines
  - Direct access to hardware may not be available in virtualized environments (e.g. InfiniBand)
  - All allocated resources run with native speed

Using InfiniBand in Virtualized Environments

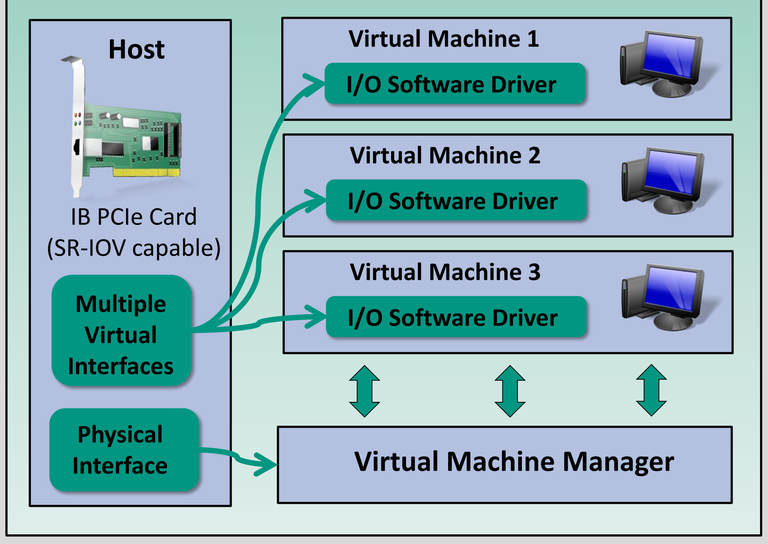
PCI Pass-Through

Direct Access with VT-d / IOMMU

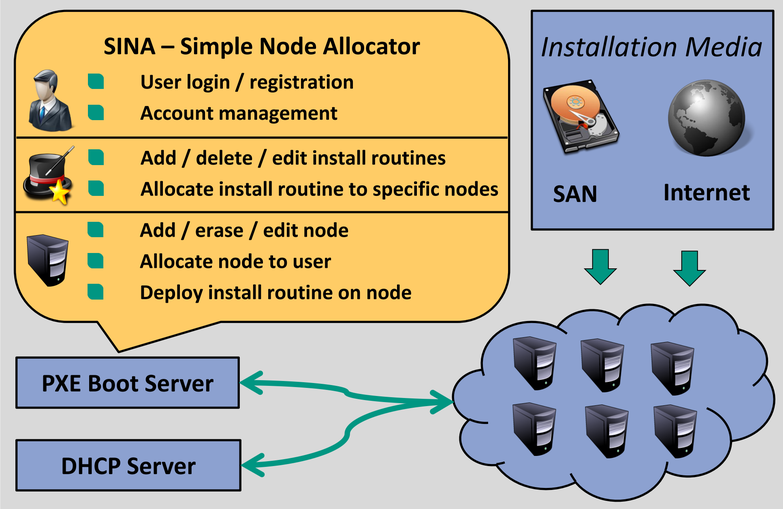


Single-Root - I/O Virtualization

DMA and Device Sharing with Intel® VT-d and SR-IOV



Physical Resource Deployment with SINA



Current Development and Outlook

Goal at KIT:
Development of an HPCaaS Prototype System

- PCI Pass-Through and Physical Deployment already work
- First SR-IOV supported IB Host Channel Adapters (HCAs) are already available by Mellanox® Technologies: Model Type: **ConnectX®-2**

Mellanox TECHNOLOGIES

OPENFABRICS ALLIANCE

- SR-IOV supported Drivers for the OFED Software Stack and Firmware are currently in development and will be available end of 2010
- Next steps: Create & manage Isolated domains within a IB fabric for multi-tenancy
  - Using special IB switches with isolation support
  - Dynamic configuration of the IB subnet manager

➡ **Enable customers to instantaneously reserve complete HPC computing clusters according to their needs!**