Karlsruhe Institute of Technology

Towards High Performance Cloud Computing (HPCC)

Marcel Kunze, Karlsruhe Institute of Technology (KIT)

Research Group Cloud Computing

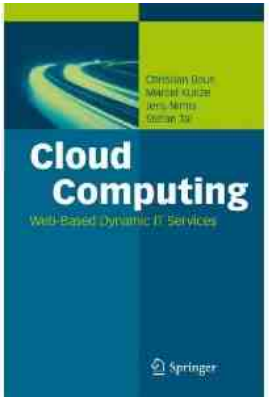
Contents

- Cloud Computing and HPC as a Service
- Amazon Compute Cluster Instances
- Virtualization of Infiniband
- Outlook

Cloud Computing: Definition

"Building on compute and storage virtualization, ***cloud computing*** provides scalable, network-centric, abstracted IT infrastructure, platforms, and applications as on-demand **services** that are billed by consumption."

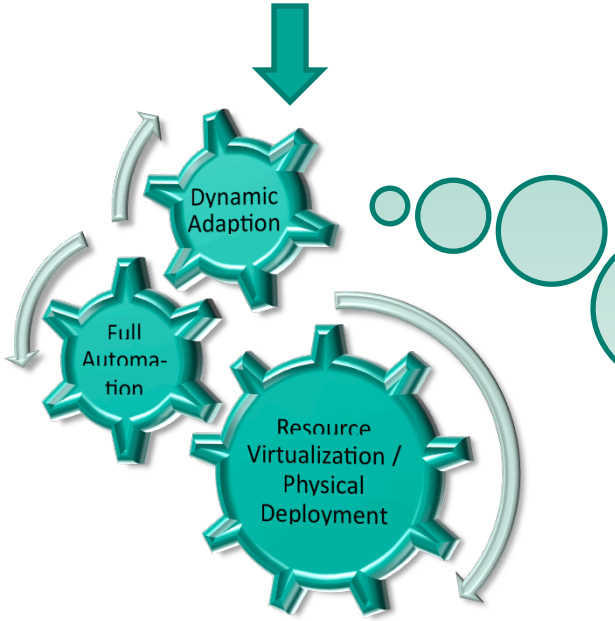
C.Baun, M.Kunze, J.Nimis, S.Tai: Cloud Computing, Springer 2011



HPC as a Service

Traditional HPC Architecture …

- is characterized by very specific computer clusters designed for special applications
- offers pre-defined operation systems and user environments only
- serves one single application at a given time
- provides restricted user access
- provides management privileges exclusively to administrators

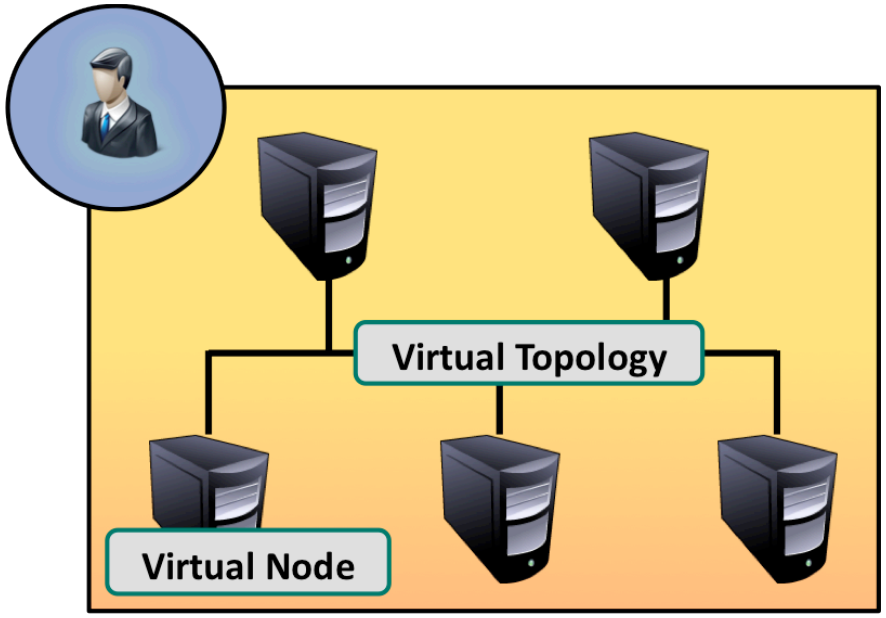


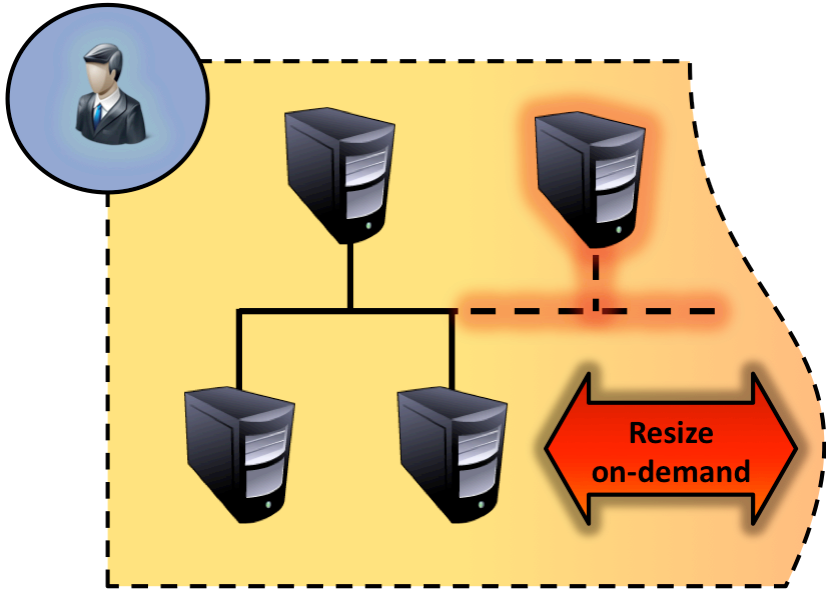
Concept of HPCaaS

- Capability of using clustered servers and storage as resource pools, fully automated management
- Individual cluster configuration on-demand
- Flexibility to serve multiple user groups and applications with varying requirements
- Customers gain resource management privileges

High Performance Cloud Computing (HPCC)

- Clusters of virtual machines
- Cluster management on the user level
- Performance guarantees
- Elasticity to consume exactly the HPC capacity required
- Dynamic pricing models and non-HPC workload to optimize utilization





Amazon Compute Cluster Instances

Blog: Behind the Cloud

Behind the Cloud | Main Blog Index

July 15, 2010

The Low-Latency Imperative and Amazon's New CCI for HPC

Today Purdue University's Coates Cluster, which is ranked at the #103 spot on the TOP500 supercomputer roll, was declared to the first native 10Gb Ethernet cluster system to be ranked on the honor roll, which means, of course, that the cluster of clusters before this one have all been employing the mighty InfiniBand to sate their low-latency imperatives.

There is little room for questioning that the purist side of the high performance computing community sees InfiniBand as the gold standard. Shortly after my surprise following the announcement regarding Amazon's new HPC-inspired Compute Cluster Instances, which have the power to place them at the equivalent of the #145 position on the TOP500 list, I figured that the word "InfiniBand" would follow—but it didn't. Amazon instead went with 10GbE, a decision that has ruffled a few feathers because it is seen by some as being still inferior on low-latency front.

Nicole Hemsoth is the managing editor of *HPC in the Cloud* and will discuss a range of overarching issues related to the intersection between high performance computing and the cloud. You can reach Nicole via email at editor@hpcinthecloud.com

More Nicole Hemsoth

High-CPU Instances

Instances of this family have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications.

- High-CPU Medium Instance 1.7 GB of memory, 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each), 350 GB of local instance storage, 32-bit platform
- High-CPU Extra Large Instance 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform

Cluster Compute Instances

Instances of this family provide proportionally high CPU with increased network performance and are well suited for High Performance Compute (HPC) applications and other demanding network-bound applications. Learn more about use of this instance type for HPC applications.

- Cluster Compute Quadruple Extra Large 23 GB memory, 33.5 EC2 Compute Units, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet

Cluster GPU Instances

Instances of this family provide general-purpose graphics processing units (GPUs) with proportionally high CPU and increased network performance for applications benefitting from highly parallelized processing, including HPC, rendering and media processing applications. While Cluster Compute Instances provide the ability to create clusters of instances connected by a low latency, high throughput network, Cluster GPU Instances provide an additional option for applications that can benefit from the efficiency gains of the parallel computing power of GPUs over what can be achieved with traditional processors. Learn more about use of this instance type for HPC applications.

- Cluster GPU Quadruple Extra Large 22 GB memory, 33.5 EC2 Compute Units, 2 x NVIDIA Tesla "Fermi" M2050 GPUs, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet

EC2 Compute Unit (ECU) – One EC2 Compute Unit (ECU) provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Source: aws.amazon.com, 2011

HPL Benchmark Results on Amazon EC2 Instances

- Execution of the HPL benchmark using Intel MPI, MKL, and ICC on various Amazon EC2 systems
  - m1.large: 2 cores, 7.5GB RAM, 4ECU

T/V	N	NB	P	Q	Time	Gflops
WR01C2R4	4096	128	1	2	5.05	**9.084e+00**

  - m1.xlarge: 4 cores, 15GB RAM, 8ECU

T/V	N	NB	P	Q	Time	Gflops
WR01C2R4	8192	128	2	2	13.64	**2.687e+01**

  - m2.2xlarge: 4 cores, 34.2GB RAM, 13ECU

T/V	N	NB	P	Q	Time	Gflops
WR01C2R4	16384	128	2	2	78.80	**3.721e+01**

  - cc1.4xlarge: 8 cores, 68.4GB RAM, 26ECU

T/V	N	NB	P	Q	Time	Gflops
WR01C2R4	16384	128	2	4	38.61	**7.595e+01**

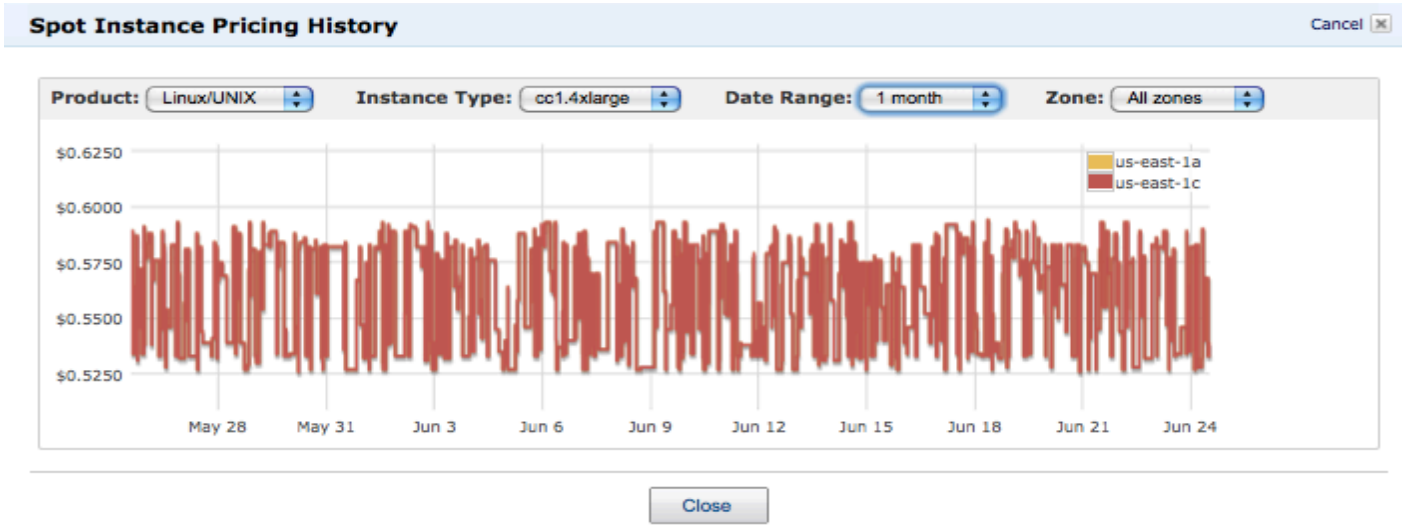
Current HPCaaS offered by Amazon (cc1.4xlarge)

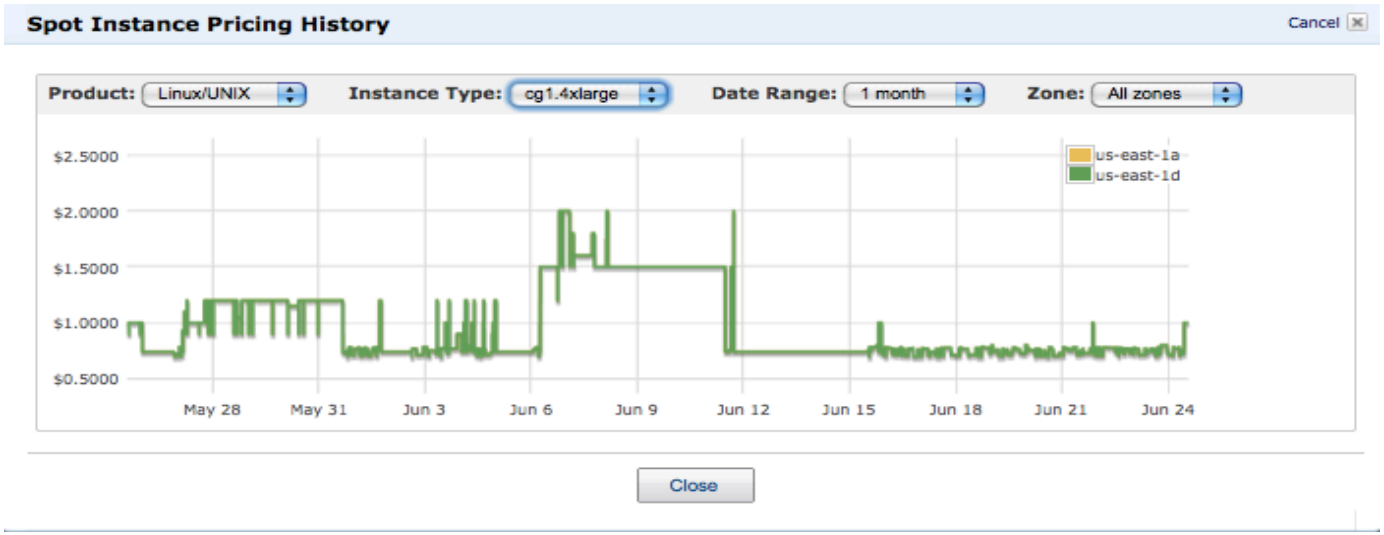
23 GB of memory
8 cores (Intel Nehalem Xeon X5570)
1690 GB of instance storage
64-bit platform
10 Gigabit Ethernet

Cost:
1,60 $ / hour (Spot price ca. 0.50$ / hour)
4290 $ / year
6590 $ / 3 years

- **Getting HPC resources from the spot market is pretty cheap**
  - Works marvelous for development, testing and smaller campaigns
  - No guarantee for continuous operation: If the market price exceeds your maximum offer, the machine is stopped
  - Checkpointing is of importance (Snapshots)

HPC CPU Spot Market Prices





TOP500 SUPERCOMPUTER SITES

ISC

Karlsruhe Institute of Technology

http://www.top500.org/

PROJECT | LISTS | STATISTICS | RESOURCES | NEWS

Home ▸ Lists ▸ November 2010

TOP500 List - November 2010 (1-100)

R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

Power data in KW for entire system

next

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00
2	DOE/SC/Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
...						
230	Telecommunication Company China	Cluster Platform 3000 BL460c G6, Xeon E5540 2.53 GHz, GigE / 2010 Hewlett-Packard	7848	41.88	79.42	
231	Amazon Web Services United States	Amazon EC2 Cluster Compute Instances - Amazon EC2 Cluster, Xeon X5570 2.95 Ghz, 10G Ethernet / 2010 Self-made	7040	41.82	82.51	

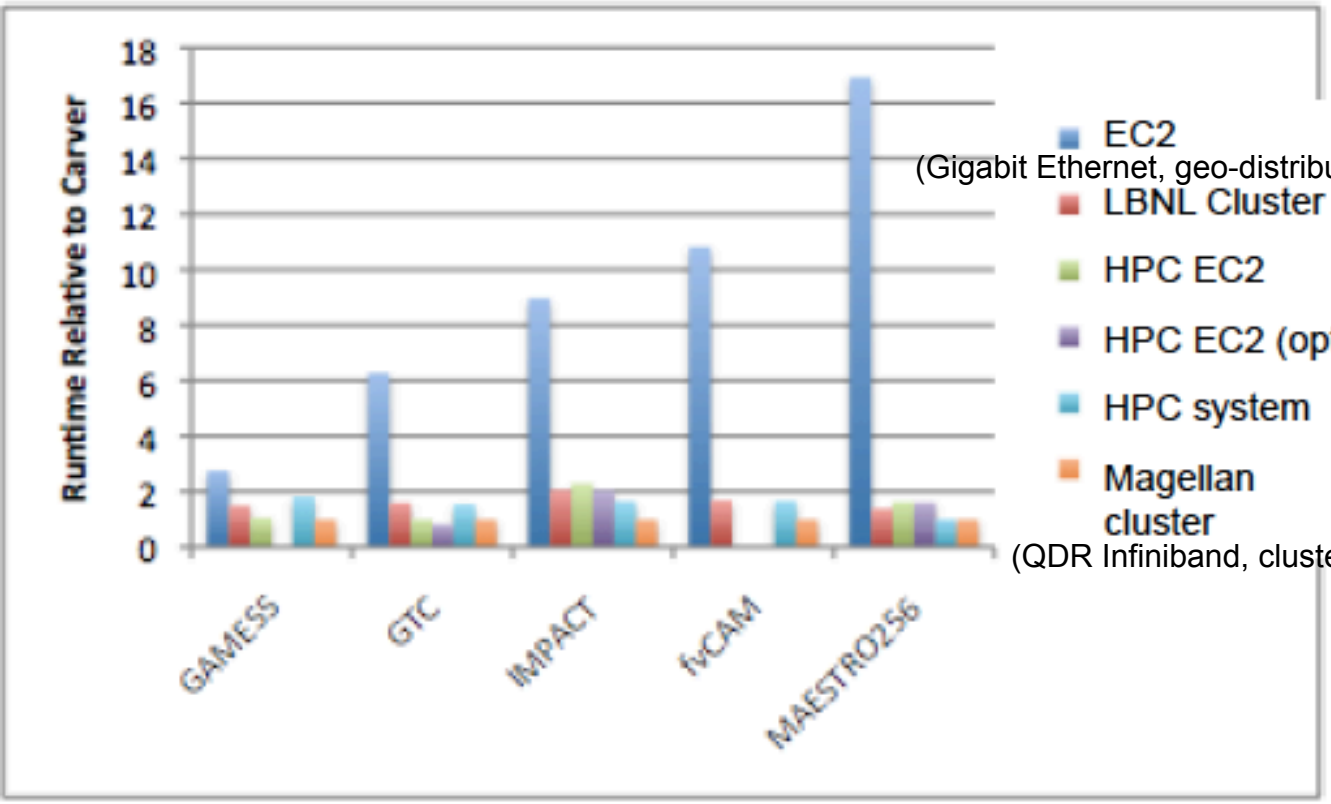
AWS advertising campaign to launch **Linpack Benchmark:**

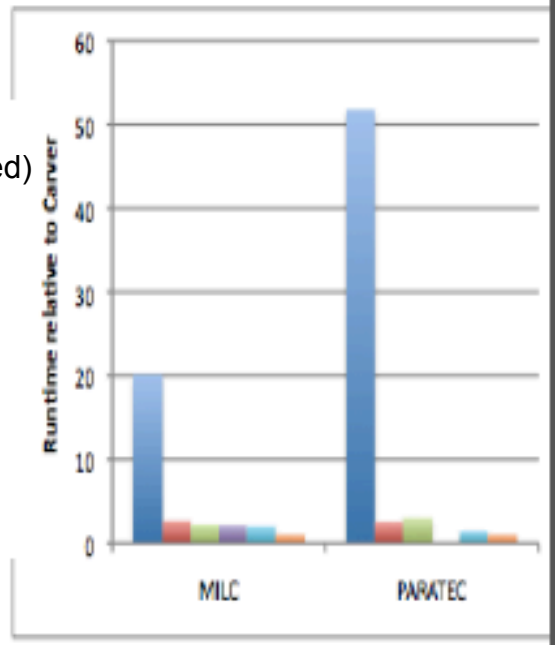
880 instances (7040 cores) reached 42 TFlop/s

Communication and Performance in AWS

- **AWS communication only offers Ethernet in virtual machines**
  - Most HPC in-house clusters use physical machines
  - Typical HPC bandwidth: 40 Gigabit Infiniband
  - Typical HPC latency: Micro-seconds
- **AWS instances are usually located in different locations / racks, yielding a large latency (a few dozen milli-seconds)**
  - Paravirtualization
  - Bandwidth: 1 Gigabit Ethernet (1GE)
- **AWS Compute Cluster Instances (CCI) may be grouped and are located in a close context with low latency**
  - Hardware virtualization (HVM)
  - Bandwidth: 10 Gigabit Ethernet (10GE)
- **A science benchmark has been produced at NERSC/Berkeley**
  - AWS clusters based on normal instances are much slower (except BLAST)
  - AWS clusters based on CCI are competitive in all fields

HPC Commercial Cloud Results





- **Commercial HPC clouds catch up with clusters if set up as shared cluster**
  - **High speed network and no over-subscription**

U.S. DEPARTMENT OF ENERGY | Office of Science

Keith Jackson, Lavanya Ramakrisha, John Shalf, Harvey Wasserman

Slide presented by Kathy Yelick, ISC Cloud

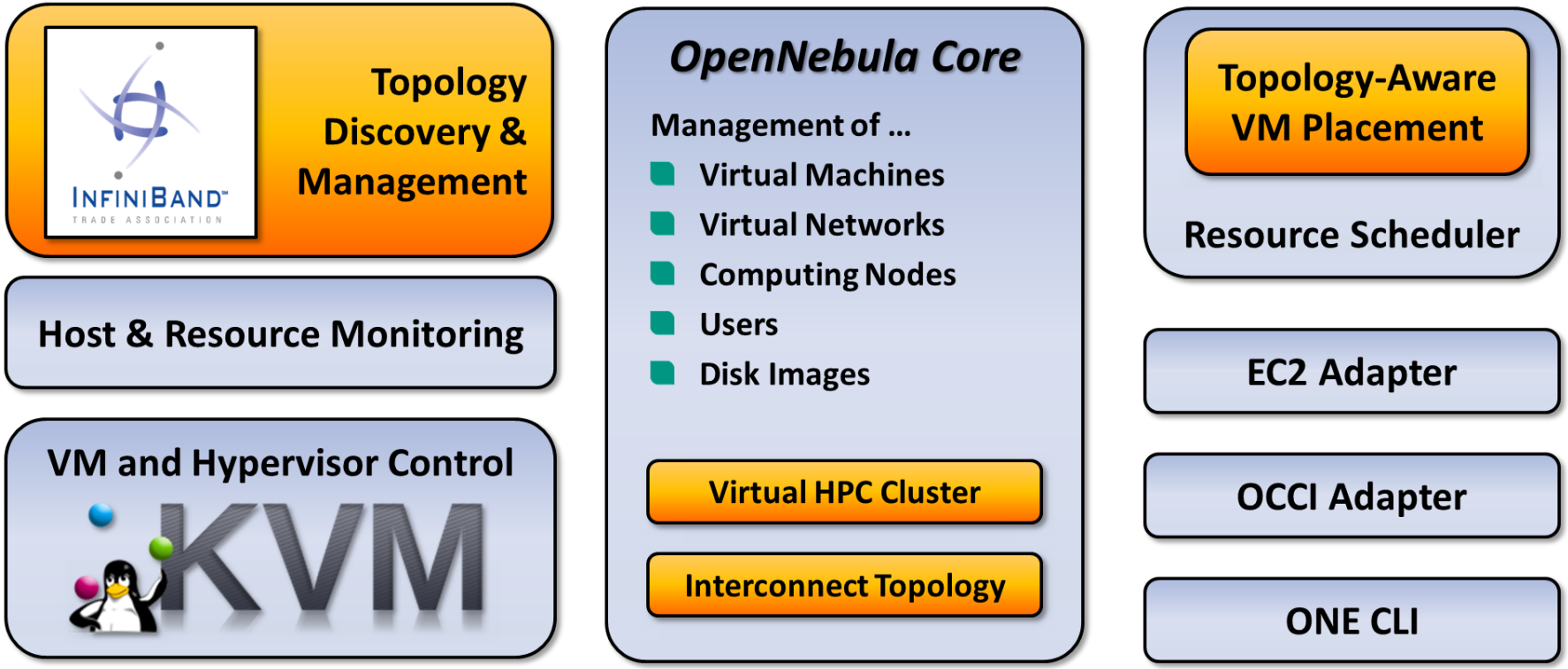
HPCaaS at KIT

- Challenge: **Development of a HPCaaS system with Infiniband support**

- 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: PCI pass-through of Infiniband interface
- Setup of a testbench on the basis of OpenNebula with extensions

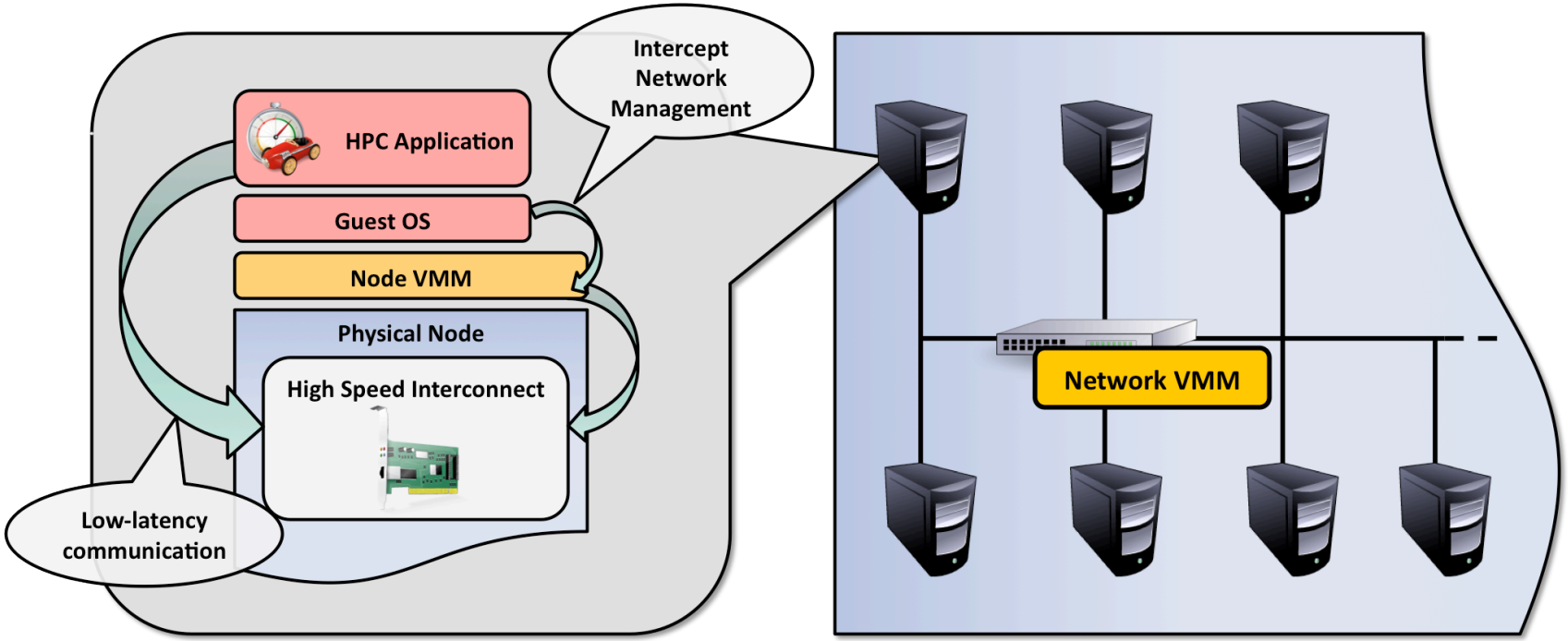
Cloud Management – OpenNebula



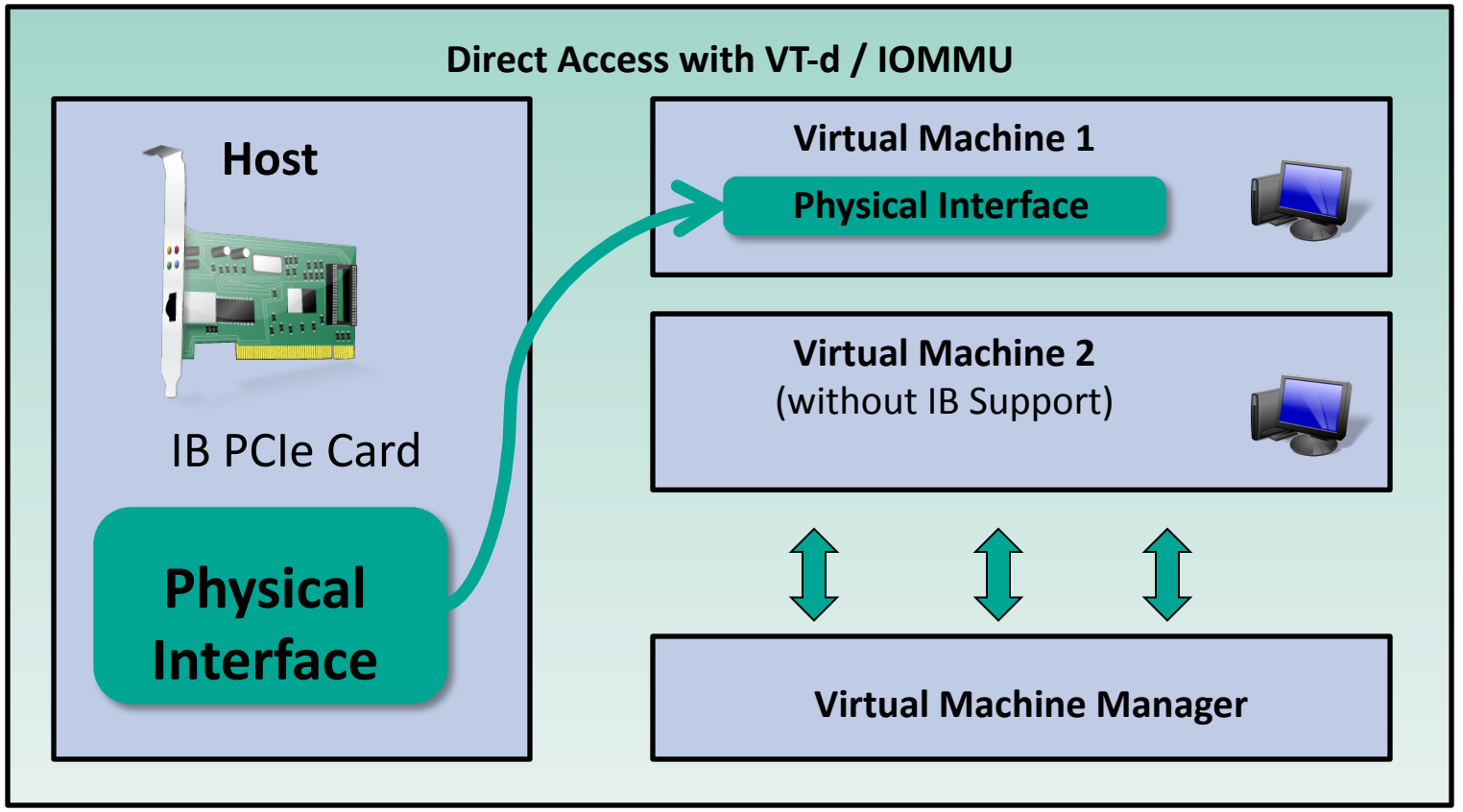
- ONE Extensions…
  - Allow to allocate complete virtual clusters with InfiniBand Support (pass-through)
  - Control the InfiniBand Subnetmanager to configure partitions dynamically

Node Virtualization

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



InfiniBand Virtualization: PCI Pass-Through



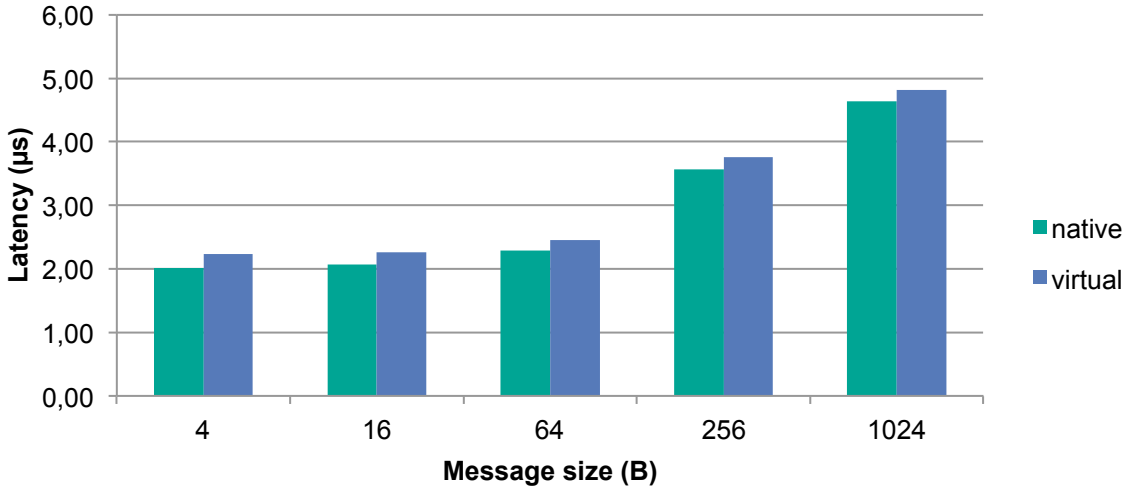
Single-Tenancy wrt. Infiniband

- 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 near native performance

Evaluation: Current State

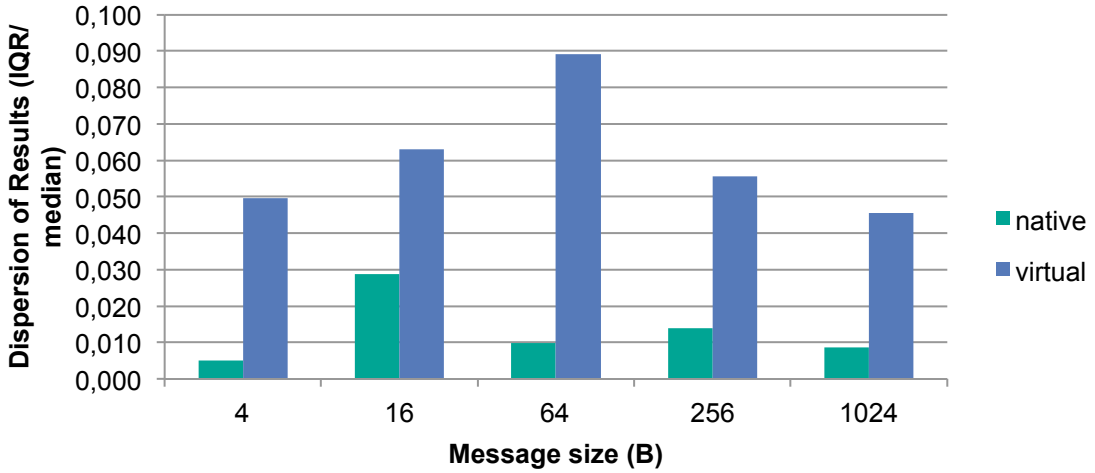
- Readily available virtualization
  - Linux / KVM (CentOS with current OFED and KVM Versions)
  - PCI pass-through (InfiniBand HCA exclusive to one VM)
- Small test cluster
  - Started with 2 blades, now 4 blades
  - Quad-core Intel Xeon E5520 CPUs
  - Mellanox ConnectX-2 InfiniBand adapters
- Benchmarks
  - SkaMPI: Special Karlsruhe MPI Benchmark
  - High Performance Linpack

SKaMPI – PingPong (SendRecv)



- Latency:
  Almost no difference between physical and virtual environment

- Jitter increases in the virtual environment
- Needs optimization of the hypervisor

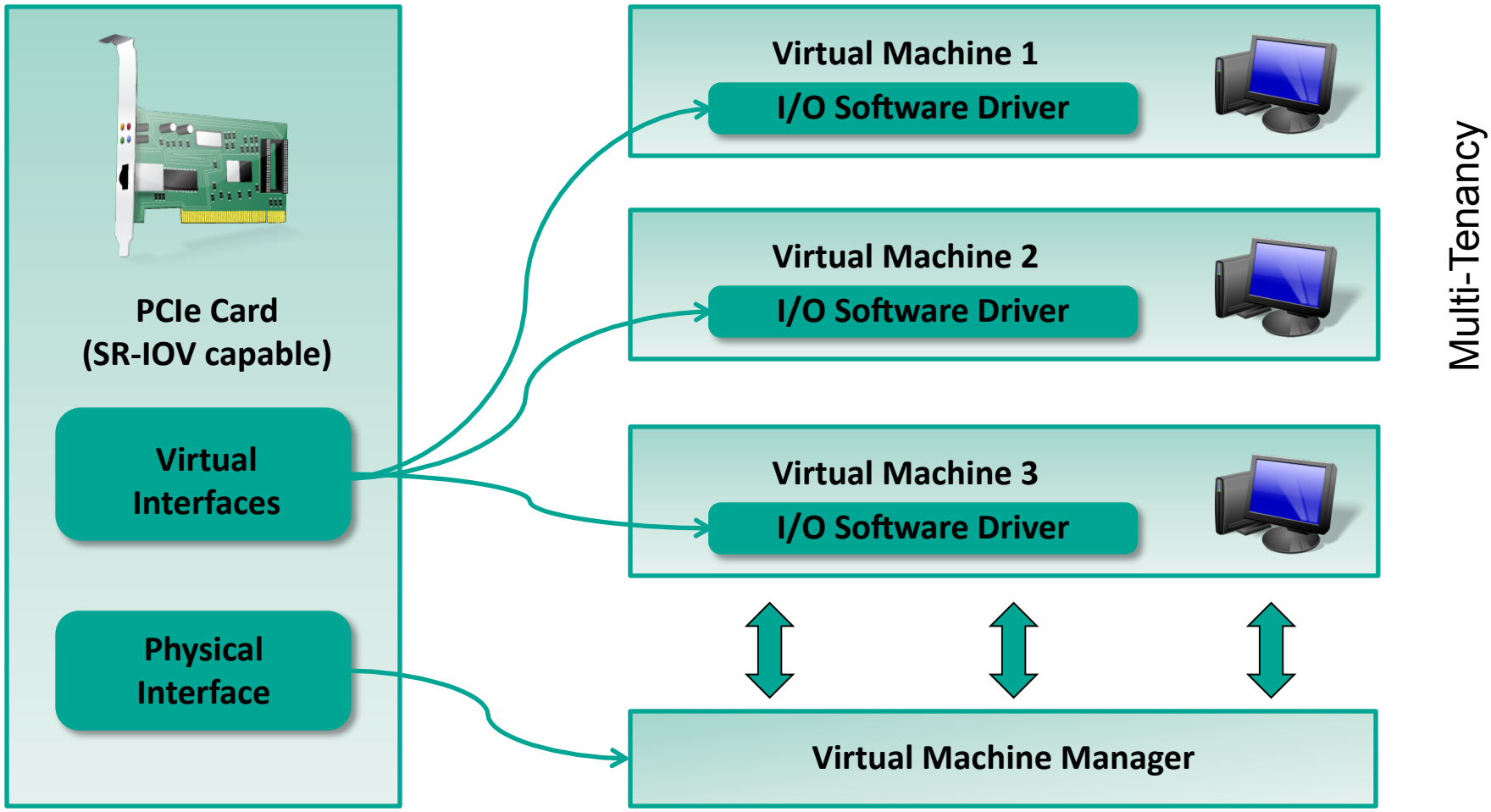


High Performance Linpack (HPL)

- Standard HPL with OpenMPI and gcc
- Assess overall performance for HPC workload
- Vary # processors, vary distribution over nodes (2 blades – 4core CPU)

Processes	Native (GFlops)	Virtual (GFlops)	Overhead %
1 / 0	1.26	1.23	**2.49 %**
2 / 0	1.85	1.78	**3.99 %**
1 / 1	2.50	2.41	**3.79 %**
4 / 0	2.21	2.19	**0.86 %**
2 / 2	3.64	3.49	**4.05 %**
4 / 4	4.24	4.25	**2.05 %**

Future: Single Root I/O Virtualization (SR-IOV)



Current InfiniBand Development

- First SR-IOV supported IB Host Channel Adapters (HCAs) are already available by Mellanox® Technologies:
  Model Type: **ConnectX®-2/3**

- SR-IOV supported Drivers for the OFED Software Stack and Firmware are currently in development and will be available in the second half of 2011

Summary and Outlook

- PCI-Passthrough already works, see ISC'11 Poster:
  Towards High-performance Cloud Computing for x86/InfiniBand Clusters
- The SR-IOV standard will allow us to use InfiniBand at near-native performance in a multi-tenant environment
- Further steps: Integration of a parallel filesystem like e.g. Lustre
- Cloud Frameworks like Nimbus or OpenNebula have to be extended to manage and monitor tightly coupled computing nodes
- **High Performance Cloud Computing is a feasible and useful concept especially to support midrange applications!**

OpenNebula


©johnlund.com