Virtualized High Performance Computing

Scientific and Engineering Agility

What is High Performance Computing?

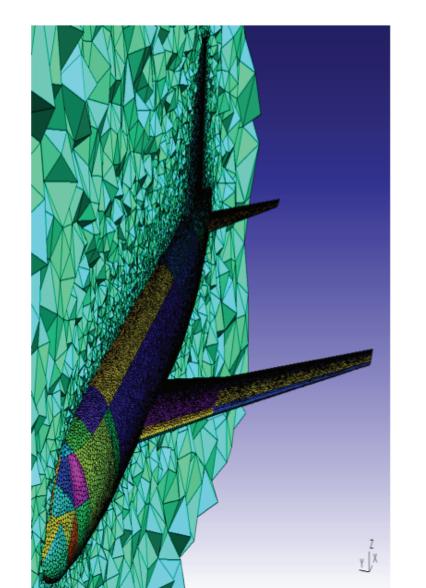
Applications & Systems

HPC applications are compute intensive, often scientific or engineering-related, often parallel, and usually involve large datasets.

Many VMware customers have HPC workloads in areas like Life Sciences, Electronic Design Automation (chip design), Computer Aided Engineering (aeronautics, automotive), Financial Services, Digital Content Creation (movie making).

HPC applications are often either throughput applications - many processes running concurrently and independently, or distributed parallel - many processes running concurrently and exchanging data repeatedly to cooperatively solve a problem too large to fit within a single host.

HPC systems are clusters - groups of a dozen to hundreds of thousands of hosts - that are administered and used as a single, large compute resource for running many HPC applications.



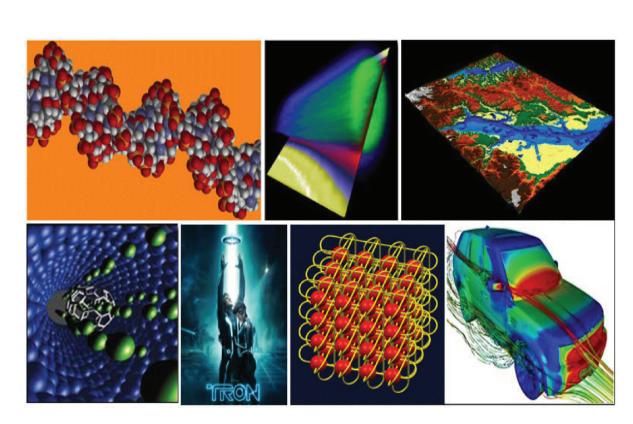
Source: geuz.org/gmsh

HPC Applications

- Science and Engineering:
 - Atmosphere, earth, environment
 - Bioscience, biotechnology, genetics
 - Physics applied, nuclear, particle, condensed matter;
 - Electrical engineering, circuit design, microelectronics
 - Mechanical engineering from prosthetics to spacecraft

Commercial:

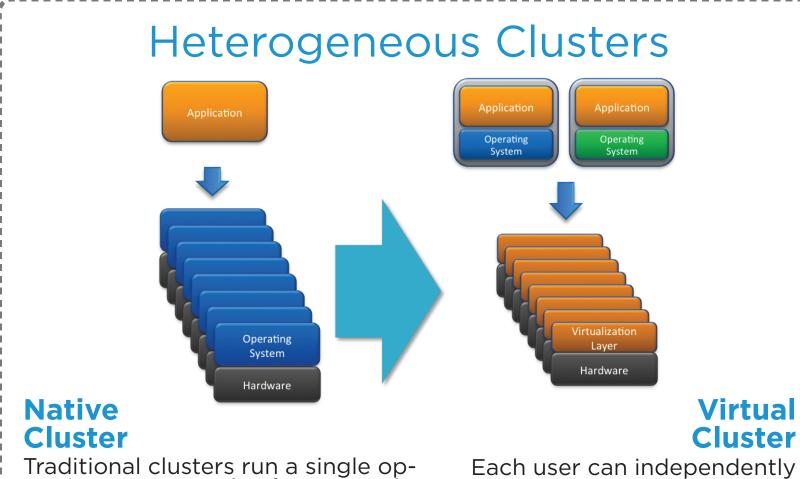
- Oil exploration
- Pharmaceutical design
- Financial and economic modeling
- Advanced data visualization



Various HPC Application Areas

Source: https://computing.llnl.gov/tutorials/parallel_comp/ by Lawrence Livermore National Laboratory

Primary Benefits of Virtualization



erating system and software stack on all nodes with no user choice.

Cluster choose their own operating

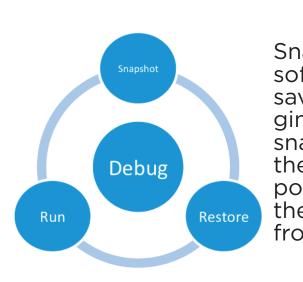
Dynamic Scheduling

Jobs scheduled onto a non-virtualized cluster cannot be moved after they are started. This can cause inefficiencies when other jobs are forced to wait even though sufficient free resources may currently be available. Using live migration in a virtual clus-

ter, workload can be dynamically rearranged to make room for waiting jobs.

Migration can also be used to reduce power use by consolidating workload during off-peak hours onto a subset of cluster nodes.

Software Development



Snapshots can be used to debug software more efficiently by saving a VM's state while debugging an application and using the snapshot to repeatedly restore the debugging session to the point prior to the bug, avoiding the need to rerun the application from the beginning.

Some MPI applications can be tested for correctness at or near scale on smaller numbers of real machines by running many VMs on each host in the test cluster to simulate a larger machine.



Workload Separation

Encapsulating applications within VMs rather than running them together within the same operating system instance has several benefits for HPC users.

Security: Applications can share the same hardware and be protected. Fault Isolation: Application or OS

crashes will not affect other users' applications. **Known Initial State:** Applications can start in a clean environment, unaffected by earlier applications that left their environments unusable.

> **Compliance and Reproducibility:** Compute environments can be archived to meet regulatory requirements and to facilitate reproducibility of scientific research results.

Reactive VMware ESX

■ n4np8

0.6

0.4

0.2

Application Resilience

With snapshots, virtualization can be used to checkpoint and restore HPC applications. In some cases, the state of an application's file storage can be saved as well, capturing the full state of the applica-

Checkpointing a single VM is straightforward. To capture the full state of an MPI job, network traffic must be quiesced before taking snapshots of each of the VMs in which MPI processes are running. Open MPI is one MPI implementation that supports the synchronization operation needed to implement this type of coordinated MPI checkpoint.



n4np4

■ n4np8

■ n4np16

■ n4np32

■ n4np64

Proactive resilience is a more sophisticated technique than checkpointing because it attempts to avoid failures rather than react after a failure has occurred. If the underlying hardware, BIOS, or hypervisor can predict probable future failures using event logs, sensor data, and statistical techniques, then live migration could be used to move VMs and their applications onto healthier systems, avoiding application

crashes. If the failing system is currently running a piece of an MPI application, then MPI traffic to and from that system must be guiesced and MPI communication links will need to be reestablished once the VM has been moved to a new system.

n4np4

■ n4np8

■ n4np16

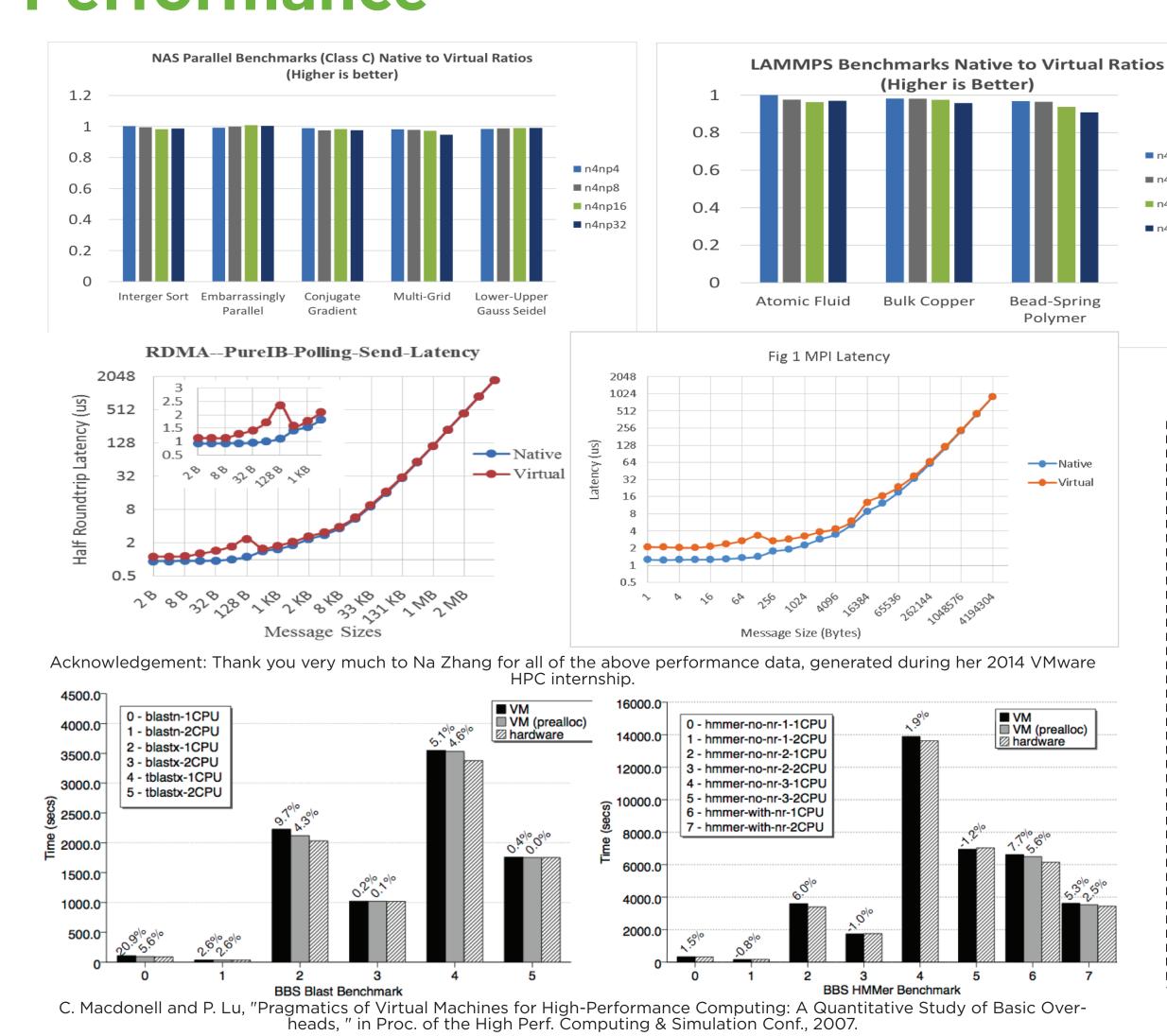
■ n4np32

n4np64

HPCC Virtual to Native Ratios

(higher is better)

Performance



Solution Architecture

High Performance Linpack Virtual to Native

Ratios (higher is better)

Secure Private Cloud

Because HPC is a market with very broad requirements, it is not possible to create a single reference architecture that meets all needs. However, the approach shown here has proven interesting to both commercial and academic customers. It can be implemented with either vCloud Automation Center (vCAC) or with OpenStack.

In this diagram, vCAC is used to create a private cloud with a self-service portal that allows research groups to "check out" a pre-configured virtual HPC cluster that has been sized to the requirements of their specific applications. This architecture has been deployed at one large customer site and we've had interest from others in this approach.

