servers from practically anywhere without severe limitations on functionality. It is possible to remotely power up and down the compute nodes, get the console output of the server via KVM or obtain valuable sensor information such as fan speed or cpu temperatures. The management cards conform to the *IPMI 2.0* specification, thus allowing non-interactive remote control by means of open source tools such as *ipmitool*.

As a second pillar of monitoring we use the Simple Network Management Protocol (SNMP) supported by almost all of the installed network connected devices. This includes reading out the environmental sensors of the water cooled racks or air condition parameters. SNMP not only enables us to read sensor information, but also to control parameters, e.g. the fan speeds of the water cooled racks, or perform remote commands, e.g. open the doors of the racks. The sensor information is gathered and archived on a dedicated server. This server provides status information of the cluster as well as sensor information via a web interface. For the purpose of data visualization we use rrd-tool, displaying the sensor values with a decreasing time resolution, i.e. the newest information is resolved with the largest resolution. The monitoring also implements emergency handling scripts. Crucial parameters are regularly checked against user defined policy tables, according to which appropriate actions are undertaken to resolve the issue or preserve the system from further damage. Since our air condition system is not redundant emergency handling is of utmost importance and not merely an academic issue. Even though the cluster components have intrinsic safety measures to prevent hardware damage, several operation critical IT infrastructure components of our institute are located in the same server room. The emergency shutdwon scripts already proved invaluable in the case of an air condition failure, which led to substantial heat up of the server room within short time. In addition to the software driven precautions a circuit breaker connected to room temperature sensors was installed, physically interrupting the cluster's power supply if the temperature thresholds are exceeded.

Contact

Professor Hartmut Wittig
Institut für Kernphysik
Johannes Gutenberg-Universität Mainz
Johann Joachim Becher-Weg 45
D-55099 Mainz
Germany
☎ +49 6131 39-26808
FAX +49 6131 39-27079
✉ wittig@kph.uni-mainz.de
☞ http://www.kph.uni-mainz.de/T/230.php

The Wilson HPC Cluster

Institut für Kernphysik
Johannes Gutenberg-Universität Mainz

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Overview

Computer simulations play an increasingly important role in physics, and have established themselves as a third pillar besides the more traditional disciplines of experimental and theoretical physics. In particle and nuclear physics, Monte Carlo simulations are becoming the standard tool in order to understand the forces between quarks inside atomic nuclei at the quantitative level. The theoretical foundation is Quantum Chromodynamics (QCD), which describes the interactions between quarks and gluons. The formulation of QCD on a discrete space-time lattice makes it amenable to large-scale numerical simulations, similar to Monte Carlo simulations applied in condensed matter physics.

The physics projects carried out on this HPC cluster focus on predictions of properties of the fundamental particles which all known matter in the universe is made of. The results will complement particle physics experiments like the ones at the LHC at CERN. The aim is to assess where our theoretical understanding of nature has its limits and where possibly so far unobserved new worlds of matter are waiting to be discovered and understood.

For our research projects, physical quantities that describe the low-energy properties of Quantum Chromodynamics have been identified which are expected to show a clean signal of *new physics* beyond the Standard Model of Particle Physics, and the objective is their high precision determination from theory. The cleanest way to achieve this is based on numerical simulations using state-of-the-art high performance computers. New ideas on how the quality of the numerical simulations can be improved significantly will be developed, implemented and applied.

Traditionally, the types of computing platforms that are used in simulations of lattice QCD are

- commercial supercomputers,
- custom-made platforms and
- clusters based on commodity hardware.

Currently, one of the largest high performance computing (HPC) cluster installations dedicated exclusively to simulations of lattice QCD is operated by the Institute for Nuclear Physics at the University of Mainz. The cluster comprises 280 compute nodes, each equipped with two AMD QuadCore processors, and a switched Infiniband network.

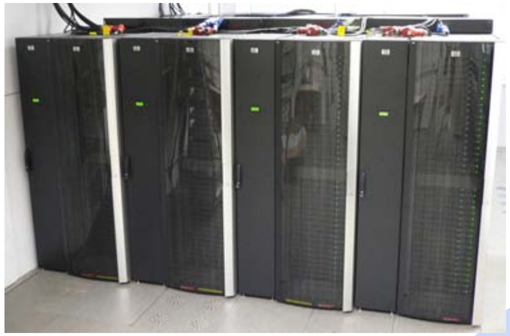


Figure 1: One row of water cooled Racks.

The peak performance of the entire system amounts to 22 TFlops, while 17 TFlops are achieved for the Linpack benchmark. The total memory of the system amounts to 2.24 TBytes. The efficient cooling of the installation is provided via water-cooled server racks.

The Wilson HPC Cluster

The Wilson cluster consists of seven fully equipped racks, each hosting 40 Hewlett-Packard DL165 servers, and two switches dedicated to administration and interactive usage of the nodes. Obviously such highly populated racks demand efficient cooling, i.e. we need a cooling capacity of approximately 12 kW per rack. Water cooled racks operating at a modestly low intake temperature provide a very high cooling efficiency at moderate costs. In addition to the water cooled racks, one air cooled rack is installed containing a HP DL385 server managing the system, an Infiniband switch for fast communication between the compute nodes and the storage system with a capacity of 9 TB configured as a RAID 5 with one hot spare. The Infiniband network consists of a Voltaire ISR 2012 switch providing 288 ports (see picture on the front cover) at 20 Gbps, connected to host channel adapters using copper cables.

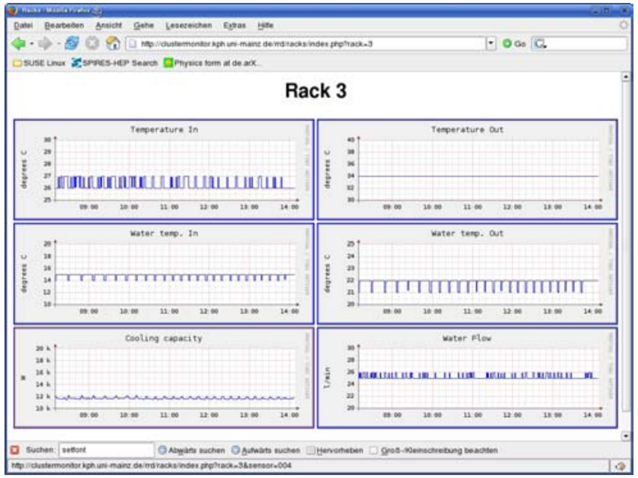
The DL165/DL385 servers have two AMD QuadCore processors and 8/16 GB of RAM, with Red Hat Enterprise Linux 5.1 as the operating system.

Administration and Monitoring

For the installation of the OS on the 280 compute nodes we used the cluster Management Utility (CMU) supplied by HP. CMU allows for a very efficient installation using the so called cloning mechanism of nodes. One first creates a *golden image* which in turn is automatically distributed over the entire cluster, including appropriate local reconfigurations, e.g. IP addresses and hostnames. The distribution is done in a cascading manner, such that each cloned node acts as a distributor. Furthermore one can define network entities, typically all nodes in a rack, reducing the overall network load, thus significantly improving the performance. The distribution of the golden image over the entire cluster takes one hour. Furthermore, CMU provides a parallel shell which is very useful for administrative purposes and online sensor information from the compute nodes.

Monitoring is an essential part of maintaining large clusters such as Wilson. Already in the design phase great emphasise was put on manageability of the various cluster components.

All compute nodes are equipped with baseboard management cards (*Lights-Out 100i*), which allow for an almost complete remote control of the servers. The management cards enable the administrator to use and manage the



http://clustermonitor.kph.uni-mainz.de/rrd/racks/index.php?rack=3

Suchen: setfont

http://clustermonitor.kph.uni-mainz.de/rrd/racks/index.php?rack=3&sensor=004

Figure 2: Web interface of our monitoring displaying sensor information of rack three.