HK 70: Instrumentierung XIV

Zeit: Freitag 14:00–16:00 Raum: HG ÜR 6

HK 70.1 Fr 14:00 HG ÜR 6

Scheduling of Virtualized Machines for ALICE HLT — Stefan Boettger[1], Udo Kebschull[1], and •Volker Lindenstruth[2] for the ALICE-HLT-Collaboration — [1]Kirchhoff-Institut für Physik, Heidelberg — [2]Frankfurt Institute for Advanced Studies

For the ALICE experiment at CERN a computing farm (ALICE HLT) is used for on-line processing of events. There are phases where no or few data is available for processing, leaving the computing power of this special-purpose cluster unused. Our goal is to make these computing resources available to 3rd party physics applications whenever possible, thereby making the HLT a general-purpose cluster. To achieve this goal the resource constraints of the on-line event processing have to be satisfied and a provisioning scheme of unused resources to 3rd party applications is needed. OS-virtualization has been evaluated and found to be an enabling technology for this aim. Common scheduling algorithms have been studied and turn out to be insuffiencient in providing the flexibility needed in an on-line environment. We therefore propose a scheduling solution for virtual machines which extends job-scheduling algorithms with the preemption and migration features offered by virtualization. To comply with the on-line processing requirements a policy-based capacity management has been added to our solution to free additional resources when needed. First results with our prototypical implementation show that our framework is capable of maximizing the cluster resource utilization.

HK 70.2 Fr 14:15 HG ÜR 6

Automatic Run-Configuration of the ALICE High Level Trigger — •Timm Steinbeck for the ALICE-HLT-Collaboration — Frankfurt Institute for Advanced Studies, University Frankfurt

The ALICE High Level Trigger (HLT) uses a pipelined and component based approach for data reconstruction and analysis. Processing components push data to the next step in the processing chain via a common interface. Data flow components transport data between nodes and merge different parts of data belonging to the same event. In order for this to work, a configuration for a processing chain has to be created before the start of a run. A repository of XML files is used to automate this, with each file holding the necessary configuration for one component, including its parents components that provide its input data. The ALICE Experiment Control System (ECS) provides a number of configuration parameters to the HLT, including an identifier for the trigger menu with the algorithms to run, a list of participating detectors, and a list of active input DDLs providing data from the detectors to DAQ and HLT. From these parameters an HLT configuration is determined fully automatically including determination of the full parent hierarchy from the top-level trigger and output components to the components receiving the data from the detector, without any manual intervention or configuration.

Work on the ALICE High-Level Trigger has been financed by the German Federal Ministry of Education and Research (BMBF) as part of its program "Förderschwerpunkt Hadronen- und Kernphysik - Großgeräte der physikalischen Grundlagenforschung".

HK 70.3 Fr 14:30 HG ÜR 6

Time synchronization and measurements of a hierarchical DAQ network — •Frank Lemke[1], Sebastian Manz[2], and Wenxue Gao[1] for the CBM-Collaboration — [1]ZITI University of Heidelberg, Germany — [2]KIP University of Heidelberg, Germany

The Data Acquisition (DAQ) system for the Compressed Baryonic Matter (CBM) experiment at the Facility for Antiproton and Ion Research (FAIR) in Darmstadt will introduce different challenges. Expected raw data rates of about 1 TB/s require online filtering and preprocessing. Detectors run in a self triggered time stamped mode depending on precise time distribution and synchronization. A readout chain has been developed, composed of a Read-Out-Controller (ROC) interfacing front-end electronics, a Data-Combiner-Board (DCB) combining data and an Active-Buffer-Board (ABB) buffering, reorganizing and transferring data via PCI express to the memory of a cluster computing node. The chain communicates via optical fibers at 2.5 Gbps. A compact control system is embedded in the ROC. A link protocol has been defined, involving three traffic classes, Data Transfer Messages, Detector Control Messages and Deterministic Latency Messages, which allow precise time distribution and synchronization within the DAQ system. The optical link is also used for clock distribution, the recovered link receive clock processed in a jitter cleaner PLL can be used as local system clock and as transmit clock. A DMA engine is implemented in the ABB to transfer data from the event buffer to the host node. Performance tests running Linux 2.6 in the end node delivered 224 MB/s bandwidth. Prototype systems are in a reliable status.

HK 70.4 Fr 14:45 HG ÜR 6

Cluster Self-Test and Self-Installation — •Jörg Peschek for the ALICE-HLT-Collaboration — Kirchhoff-Institute for Physics, Heidelberg University — Frankfurt Institute for Advanced Studies, Frankfurt University

Experimental and theoretical research strongly depends on the ability to provide sufficient compute power. Furthermore different kinds of physical research may take advantage of different types of underlying hardware. Additionally an efficient cluster should be able to grow with new and more powerful hardware becoming available, like graphic cards or FPGA-Pre-Processors. The requests for a heterogeneous cluster increase the effort in cost and man power needed for cluster administration. Therefore scalable self healing is a quality a cluster must provide to keep affordable in scientific context. Helpful for a solution are board management controllers (BMC), an embedded system included in most server mainboards

Presented is the concept to handle new nodes in a cluster or nodes that show up a problem. The node should be fully tested and installed taking advantage of BMC and the node itself. It will be pointed out what is necessary to perform decentralized self administration. Furthermore report functionality to the central management solution is outlined. It should become clear, that this kind of administration can be performed in a productive cluster. Parts of the concept are already used in the High Level Trigger (HLT) cluster for the ALICE experiment at CERN.

HK 70.5 Fr 15:00 HG ÜR 6

Complexity management for heterogeneous computer clusters in case of the ALICE High Level Trigger cluster — Camilo Lara, •Timo Breitner, Udo Kebschull, Stefan Boettger, and Marian Hermann for the ALICE-HLT-Collaboration — Kirchhoff-Institut for Physics, Heidelberg University, Germany

Heterogeneous computer farms are used for processing data in the ALICE HLT. A high complexity of managing a heterogeneous environment is the result of the increasing number of hardware and software releases and the resulting number of needed management resources for their management. For system management using the SysMES framework two object-oriented and Common Information Model based models have been developed: One of them describes the cluster environment whereas the other one describes the resources used for managing the former, e.g. monitors and rules. The next step for handling the mentioned complexity was the definition of relationships between those two models e.g. between an object which describes a CPU and another one describing a monitor object controlling the CPU's temperature. Our implementation provides a Java API used internally by the SysMES server to access the models in order to identify pairs (device, management resource) and to deploy the resources to their target devices. A relational database was chosen as a back-end, due to the requirements of our management system like transactionality and data integrity. For its integration with the object-oriented model and API a specifically designed object-relational mapping formalism has been developed.

HK 70.6 Fr 15:15 HG ÜR 6

Operation & Control Interfaces based upon Distributed Agent Networks — •Pierre Zelnicek[1], Udo Kebschull[1], and Volker Lindenstruth[2] — [1]Kirchhoff Institute of Physics, Ruprecht-Karls-University Heidelberg, Heidelberg, Germany — [2]Frankfurt Institute for Advanced Studies, Frankfurt, Germany

The majority of todays large scale compute clusters and software systems running on them are using operation and control interfaces (OCI) for monitoring and control. The majority of these OCI's are still based upon single node applications, which are limited by the physical system they are running on. In areas where hundred thousand and more statistical values have to be analyzed and taken into account for visualization and decision making this kind of OCI's are no option at

all. Furthermore, this kind of OCI's do not empower whole collaborations to control and operate cluster at the same time from around the world. Distributed agent networks (DAN) tend to have the possibility to overcome this limitations. A distributed agent network is per design a multi-node approach. Together with a web based OCI, automatic data propagation and distributed locking algorithms they provide simultaneous operation and control, distributed state tracking and visualization to world wide collaborations. The first compute cluster in the scientific world using this combination of technologies is the ALICE HLT at CERN.

HK 70.7 Fr 15:30 HG ÜR 6

Introducing high availability to non high available designed applications — •Pierre Zelnicek[1], Oystein Senneset Haaland[2], Udo Kebschull[1], and Volker Lindenstruth[3] — [1]Kirchhoff Institute of Physics, Ruprecht-Karls-University Heidelberg, Heidelberg, Germany — [2]Physic Institut, University of Bergen , Bergen, Norway — [3]Frankfurt Institut für Advanced Studies, University Frankfurt, Frankfurt, Germany

A common problem in scientific computing environments and compute clusters today, is how to apply high availability to legacy applications. These applications are becoming more and more a problem in increasingly complex environments and with business grade availability constraints that requires 24x7x365 hours of operation. For a majority of applications, redesign is not an option. Either because of being closed source or the effort involved would be just as great as re-writing the application from scratch. Neither is letting normal operators restart and reconfigure the applications on backup nodes a solution. In addition to the possibility of mistakes from non-experts and the cost of keeping personnel at work 24/7, these kind of operations would require administrator privileges within the compute environment and would therefore be a security risk. Therefore, these legacy applications have to be monitored and if a failure occurs autonomously migrated to a working node. The pacemaker framework is designed for both tasks and ensures the availability of the legacy applications. Distributed redundant block devices are used for fault tolerant distributed data storage. The result is an Availability Environment Classification 2 (AEC-2).

HK 70.8 Fr 15:45 HG ÜR 6

Online control package and event display for the upgraded COSY-TOF experiment. — •Ekaterina Borodina[1,2], Eduard Roderburg[1], and James Ritman[1] for the COSY-TOF-Collaboration — [1]Institut fuer Kernphysik I, Forschungszentrum Juelich GmbH, 52325, Juelich, Germany — [2]Moscow State Institute of Electronics and Mathematics, Russia

The new Straw Tube Tracker and Silicon Quirl Telescope detectors have been recently installed at the TOF (Time Of Flight) experiment at the COSY accelerator in the FZ-Juelich. These new detectors increase the number of channels of the COSY-TOF detector by about a factor of 3. Therefore, a new control package to adjust electronic parameters and to diagnose the proper functionality of all components is being developed.

The COSY-TOF online controlling is based on visualization of single events and the analysis of statistical distributions of detectors data. It consists of conversion software, which transforms the binary data stream from the DAQ to a detector oriented event format; visualization routines, which create event display, spectra, etc. and stores them in shared memory files; methods of IPC (Inter-Process Communications) for real time performance; Geometry package and GUI (graphical user interface). The event display, based on ROOT geometry classes, represents graphically detector states and events in different ways of visualization during online sessions. Examples of the event display and results from the last experiment will be presented.

Supported in part by FZ-Juelich.