ERAC - Efficient and Robust Architecture for the big data Cloud

Evangelos Tasoulas (vangelis@simula.no), Feroz Zahid (ferozz@ifi.uio.no)
[simula . research laboratory] UiO : University of Oslo

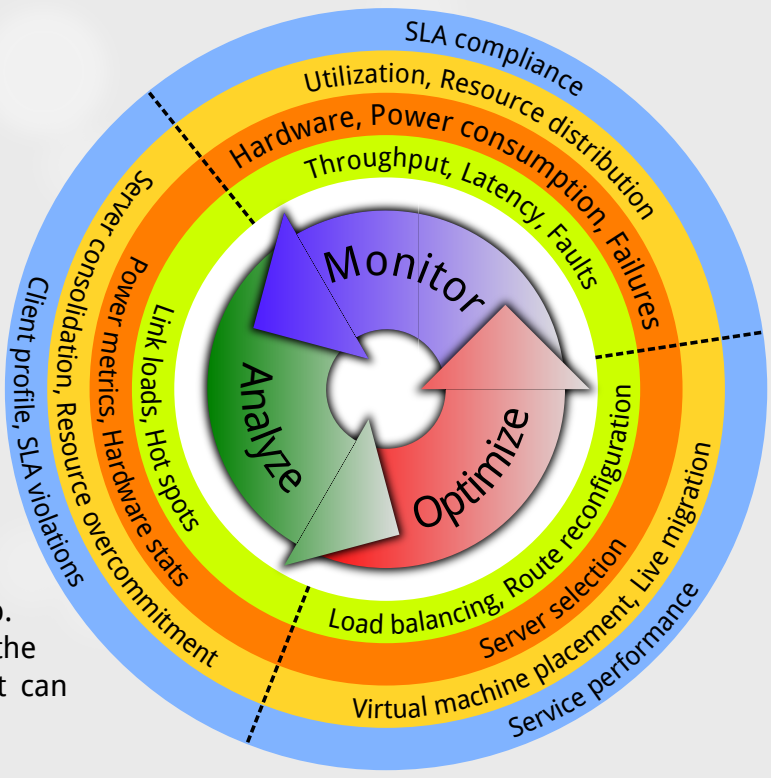
Abstract

The primary objective of the ERAC project is to provide the knowledge and solutions that enable an elastic, scalable, robust, flexible, secure, and energy efficient cloud architecture that matches both the expectations of the Social Networks (SN) and the Internet of Things (IoT) in terms of services, functionality, and the efficiency requirements of the cloud providers. In the project we shall research, develop, build, and demonstrate cloud technologies for the Future Internet.

Project Overview

ERAC investigates methods for reducing overheads associated with the cloud management on high-performance interconnection network*. We address specific challenges associated with reducing downtime of Virtual Machines (VM) during live migration, and preserving active connections after a VM has been migrated. The underlying network fabric is optimized for High Performance Computing offering high-level of predictability and performance guarantees. The project intents to enable adaptive and predictable provisioning of data center resources, to devise methods for high granularity service differentiation, and to provide robustness to the cloud computing data centers using strong fault tolerance techniques at both the network and the virtualization layer. Equipped with efficient migration technology and a self-adaptive network architecture, ERAC aims to optimize resource utilization by server consolidation and efficient orchestration in a cloud data center. The implementation of the prototypes for ERAC involves designing methods for each part of the monitor-analyze-optimize loop. After running the current configuration, monitors will collect statistics about various states of the components. The collected data will be analyzed and a new optimized configuration is proposed that can potentially make the system more efficient. This new configuration is applied and the loop continues.

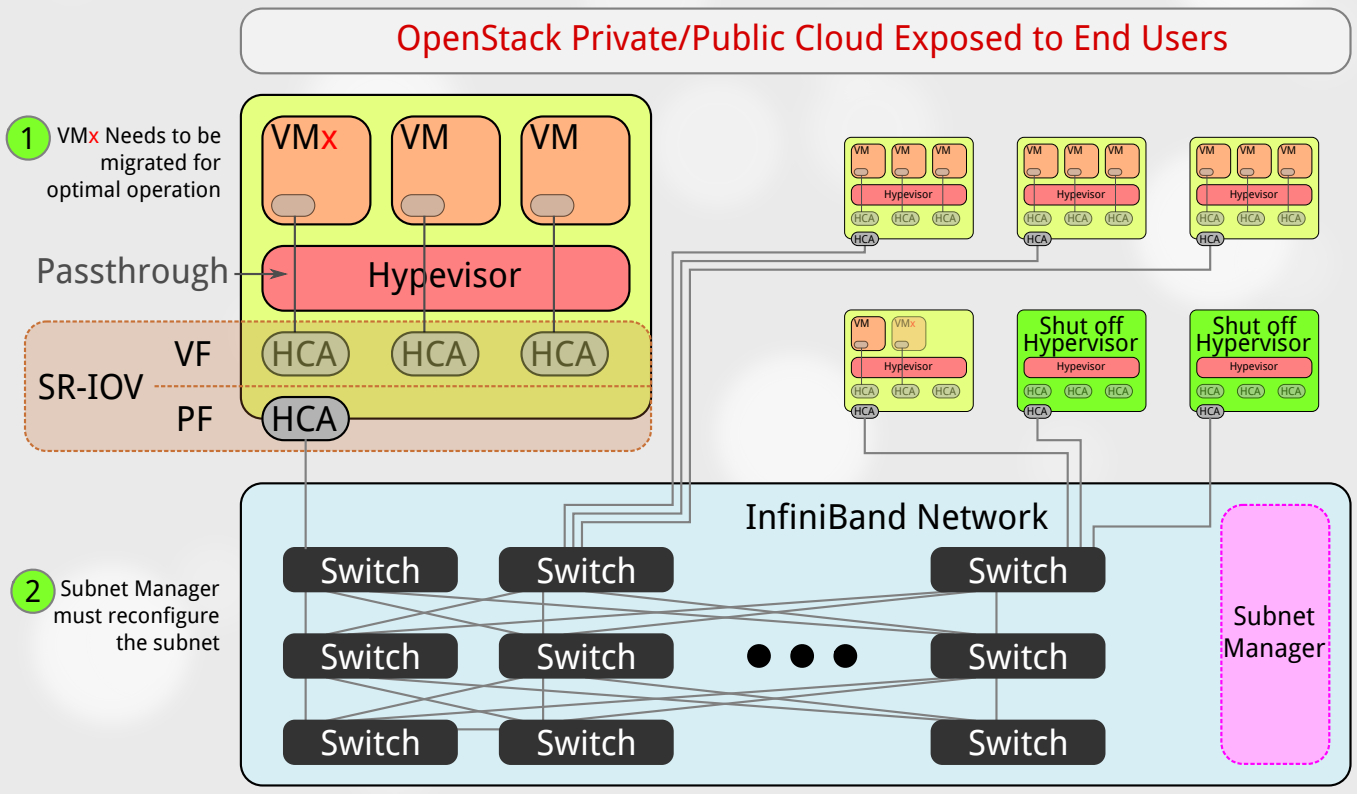
Architecture	Test-bed
Applications	
OS	
Orchestration	OpenStack
Virtualization	KVM
Storage/Servers	HCAs, CPUs, ...
Networking	Infiniband



* The ERAC project also includes cloud security and trust related research activities, handled by the University of Stavanger.

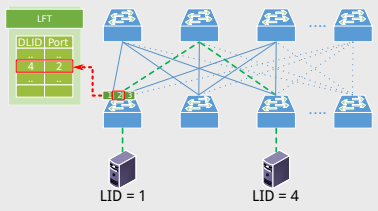
System Architecture

The ERAC test-bed consists of a virtualized High Performance Computing (HPC) infrastructure based on InfiniBand (IB) lossless interconnet network. For efficient elasticity and industrial relevance, ERAC components will be deployed, managed, monitored and (re)configured using OpenStack Cloud Computing Platform. The high level system architecture highlighting how the components in ERAC interacts with each other is given below.



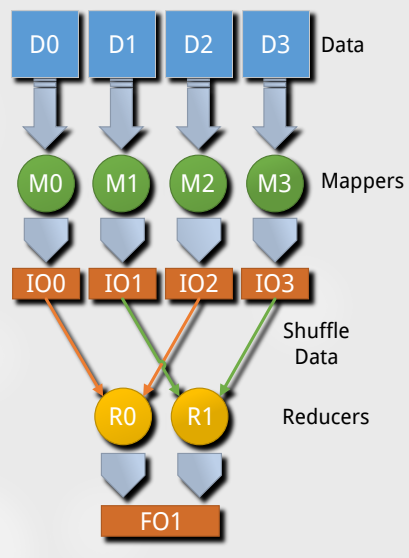
1 + 2 coordinate to optimize resource utilization for the whole datacenter

The underlying IB fabric imposes unique research issues at both the virtualization and the network layer. Addressing schemes and the deterministic routing in IB networks make it challenging to reconfigure and optimize the fabric on-the-fly, to gain performance and improve fault-tolerance. In IB, the network interface cards are called Host Channel Adapters (HCAs). Each HCA port on an end node and all switches are addressed using local identifiers (LIDs). Routing is based on linear forwarding tables (LFTs) stored in switches. The LFTs are calculated by the Subnet Manager (SM) according to the routing algorithm in use. When a packet arrives at a switch, its output port is determined by looking up the destination LID in its forwarding table (as shown in the figure below). LFTs need to be recalculated when a fabric reconfiguration is required, which is a substantial overhead in large networks. Additionally, in virtualized environments, VMs need direct access to the HCAs in order to fully benefit from the high performance network. This is achieved using a passthrough technique called Single Root I/O Virtualization (SR-IOV). SR-IOV introduces the notions of Physical PCI Functions (PF) and Virtual Functions (VFs). VFs are light weight instances of a PF (they cannot reconfigure the PF). VM live migrations need to account for reconnecting VFs, imposing further challenges.



Implications on Big Data

Map-Reduce is a programming paradigm for processing large data sets in scalable, parallel, distributed and fault-tolerant computing clusters. For efficient Map-Reduce performance, the network should be highly available and resilient to avoid overheads. For example, a map job may have to be reassigned on a different node due to the failure of a network switch in the cluster resulting in degrading overall algorithm execution time. Furthermore, latency of the network has an impact on the performance of Map-Reduce, particularly for the jobs in which nodes require higher degree of coordination or message passing between each other. However, for most types of Map-Reduce workloads, the shuffle operations between the map and the reduce phase are the most network intensive task since data is shuffled to the reducers over the network resulting in bursty traffic. If bursts are not handled efficiently in the network, dropped packets will affect the performance of the Map-Reduce as packets will need to be retransmitted over the network. ERAC's efficient network architecture aims to improve performance for Big Data applications by optimizing the network fabric, while considering the bursty nature of the traffic.



Summary

Results drawn from this project could help improve the performance of current network architectures of (virtual) HPC cloud data centers. Service providers will benefit from increased efficiency of the clouds, which may financially transfer into more revenues and reduced end user prices. For the cloud users, improved performance will aid clouds to become more efficient and flexible for deploying applications that require low latency, high throughput and Quality of Service (QoS) guarantees, resulting in improved handling of data-intensive loads and traffic peaks for a variety of computational domains.

For more information please visit: https://www.simula.no/department/cloud