Realizing a Self-Adaptive Network Architecture for HPC Clouds

Feroz Zahid (feroz@simula.no) Advisors: Ernst Gunnar Gran, Tor Skeie

[simula . research laboratory] UiO : University of Oslo

Abstract

Clouds offer significant advantages over traditional cluster computing architectures including easy of deployment, rapid elasticity, and an economically attractive pay-as-you-go business model. However, effective use of cloud computing for HPC systems still remains questionable. When clouds are deployed on lossless interconnection networks, like InfiniBand (IB), challenges such as load-balancing, low-overhead virtualization, and performance isolation hinder full potential utilization of the underlying interconnect. In this work, we attack these challenges and propose a novel holistic framework of a self-adaptive IB subnet for HPC clouds. Our solution consists of a feedback control loop that effectively incorporate optimizations based on multidimensional objective function using current resource configuration and provider-defined policies. We build our system using a bottom-up approach, starting by prototyping solutions taking up individual research challenges associated, and later combining them together to build a working self-adaptive cloud prototype.

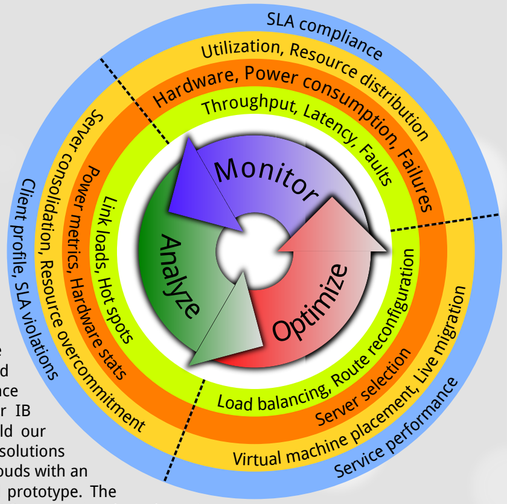
Project Overview

Architecture

- Applications
- OS
- Orchestration
- Virtualization
- Storage/Servers
- Networking

Test-bed

- OpenStack
- KVM
- HCAs, CPUs, ...
- Infiniband

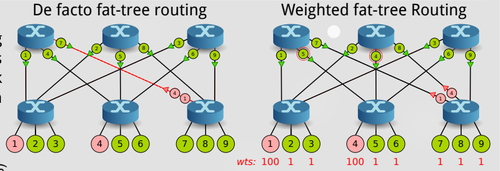


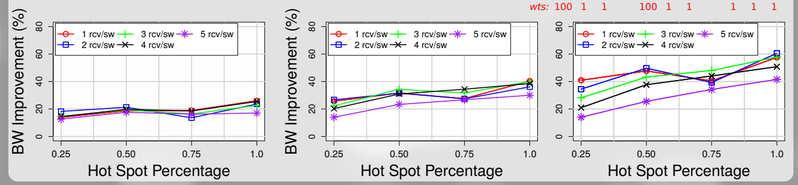
This work is a part of the ERAC project which investigates methods for reducing overheads associated with the cloud management on high-performance interconnection networks. We address the specific challenges associated with the optimization of the underlying network fabric for cloud applications, enabling high-level of predictability and performance guarantees. We design a holistic self-adaptive framework for IB subnets realizing HPC clouds based on the fat-trees. We build our system using a bottom-up approach, starting by prototyping solutions taking up individual research challenges associated with HPC clouds with an proactive plan to combine them later in an integrated cloud prototype. The implementation of the integrated cloud prototype in the ERAC project 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.

More details about the ERAC project are found at: https://www.simula.no/research/communication-systems/cloud

Efficient Load Balancing

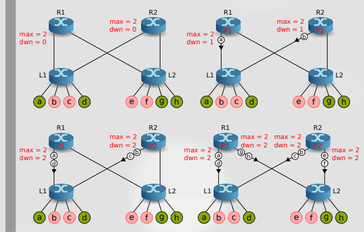
The weighted fat-tree routing considers node traffic characteristics to balance load across the network links more evenly, and with predictable network performance.



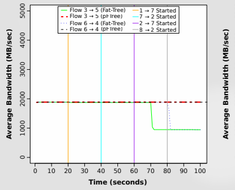


Tenant Performance Isolation

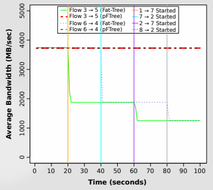
The pFTree algorithm utilizes several mechanisms to provide network-wide isolation of partitions belonging to different tenant groups.



Port selection in pFTree Routing

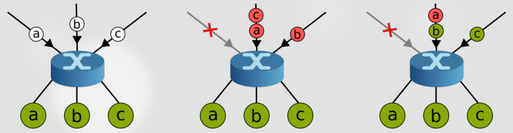


Non-oversubscribed Topology

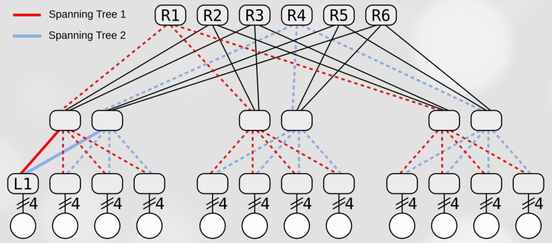


Oversubscribed Topology

Fast and Compact Network Reconfiguration

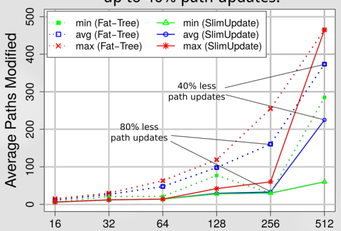


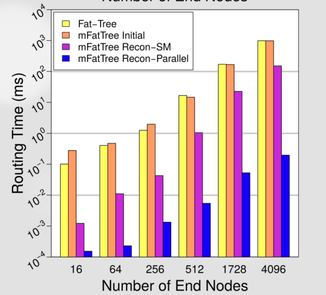
Minimal Route Update (MRU) ensures that the minimum required number of path updates are done on a reconfiguration event.



In metabase-aided reconfiguration method, routing is divided into two distinct phases: calculation of paths in the topology, and assignment of the calculated paths to the actual destinations. Metabase-aided routing substantially reduces the routing time on a reconfiguration event.

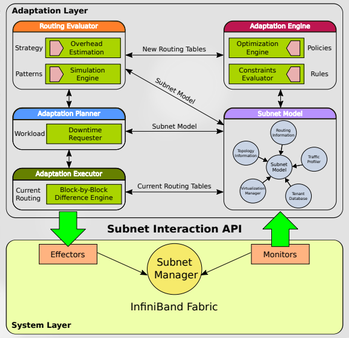
SlimUpdate Routing Algorithm saves up to 40% path updates.





The Big Picture - A Self-Adaptive IB Network

We now combine contributions of this work attacking different challenges together in the form of a holistic self-adaptive network architecture for HPC clouds. The architecture is supported by a domain specific language which user uses to write policies and actions in an Event-Condition-Action based rule format. The user-defined policies, together with the configurable system variables and monitored traffic profiles, are given as an input to a consolidated fat-tree routing algorithm to configure and adapt routing automatically.



Conclusion

In this work, we address several challenges impeding realization of efficient HPC clouds based on IB interconnect technology. We present prototype solutions to achieve efficient load-balancing, tenant performance isolation, fast and compact network reconfiguration, and improved routing for virtualized environments. Furthermore, based on our prototype solutions, we present the design of a self-adaptive network architecture for IB subnets.

For publication list, please visit: https://www.simula.no/people/feroz