IBPM: An Open-Source-Based Framework for InifiniBand Performance Monitoring

Michael Hoefling[1], Michael Menth[1], Christian Kniep[2], and Marcus Camen[2]

[1] University of Tuebingen, Chair of Communication Networks,
Sand 13, 72076 Tuebingen, Germany
{hoefling,menth}@informatik.uni-tuebingen.de
[2] science + computing ag, Hagellocher Weg 73, 72070 Tuebingen, Germany
{c.kniep,m.camen}@science-computing.de

Abstract. In this paper, we present a tool for performance measurement of InfiniBand networks. Our tool analyzes the network and presents a comprehensible visualization of the performance and health of the network. InfiniBand network operators can use the tool to detect potential bottlenecks and optimize the overall performance of their network.

1 Introduction

InfiniBand (IB) [1] is a communication technology used for interconnection in high-performance computing (HPC) data centers. With increasing computational needs, IB networks become more complex. Finding hot spots in the network, detecting congestion, and providing an overall health map of the network are features strongly needed to operate the network as efficiently as possible. Vendors of IB hardware offer proprietary software to manage IB networks. Mellanox' Unified Fabric Manager (UFM) [2] and QLogic's InfiniBand Fabric Suite (IFS) [3] allow network administrators and operators to monitor the current state of the network and optimize the routing.

In this work, we present the IB performance monitoring tool (IBPM). We implemented IBPM as an extensible framework based on open-source components. Our goal was to give network administrators a tool to monitor and analyze their IB networks.

The paper is organized as follows. In the next section, we briefly present the core idea of rate measurement in IB networks. The features of IBPM are presented in Section 3. Section 4 describes our architecture and implementation and Section 5 concludes this work.

2 Rate Measurement in IB Networks

The IB network utility software package [4] provides tools to extract raw network information from the network. The tools are outlined in the following. We use these tools, analyze their output, and derive statistics about the performance of the network.

Topology Extraction. The tool `ibnetdiscover` performs subnet discovery in the IB network. It produces a human readable file displaying all nodes and links of the topology. We further process the output to produce a graphical representation of the network which is easier to comprehend than the textual representation, especially if the network is large.

Remote Counter Readout. The IB standard [1] defines performance counters for each port of an IB device. Tools such as `perfquery` can be used to remotely read out IB counters. These counters measure, e.g., the amount of transfered data in the absence of congestion, transfered data in the presence of congestion, errors, or changes of the physical state of the link. We use the term *performance counter* for the rest of the paper.

Counter Limitations. Performance counters are unsigned 32 bit wide saturating counters. Each counter step represents a doubleword of transfered data. Thus, the maximum amount of transfered data which can be measured by an IB performance counter is $(2^{32} - 1) \cdot 32$ bit $= 16$ Gbyte. After counter saturation, no further traffic measurements can be conducted unless the counter is reset. IB uses 8b/10b line coding, i.e., the effective data rate is 20% lower than the line rate. Given a 40 Gbit/s QDR-link, the saturation time of the corresponding performance counter is calculated as follows.

$$\frac{\text{Max. counter value}}{\text{Max. trans. speed}} = \frac{(2^{32} - 1) \cdot 32 \text{ bit}}{\frac{4}{5} \cdot 40 \cdot 1024^3 \text{ bit/s}} = 3.\bar{9} \text{ s} \approx 4.0 \text{ s}$$

Hence, we may underestimate the counter value if the readout interval is greater than 4 seconds.

3 Features

IBPM offers innovative measurement and analysis methods for IB networks. As a basic feature, our tool provides automatic topology extraction and visualization. Besides the technical view on the network, statistics are available as well. To add value to the plain topology view, additional performance information can be included as an overlay. IBPM features include traffic locality visualization, measurement of congestion, and measurement of link utilization. In the following, we present the features *traffic locality* and *link utilization*.

Traffic Locality. We define the traffic locality as follows. Let $\mathcal{X}$ be a set of connected nodes. We consider a node $x \in \mathcal{X}$. The functions $out(x)$ and $in(x)$ denote traffic rates leaving or entering the set $\mathcal{X}$ in x, and the functions $gen(x)$ and $con(x)$ denote the traffic rates generated or consumed in x. The locality of generated traffic is defined by

$$l_{gen}(\mathcal{X}) = \frac{\sum_{x \in \mathcal{X}} (gen(x) - out(x))}{\sum_{x \in \mathcal{X}} gen(x)} = 1 - \frac{\sum_{x \in \mathcal{X}} out(x)}{\sum_{x \in \mathcal{X}} gen(x)}.$$

and the locality of consumed traffic is

$$l_{con}(\mathcal{X}) = \frac{\sum_{x\in\mathcal{X}} (con(x) - in(x))}{\sum_{x\in\mathcal{X}} con(x)} = 1 - \frac{\sum_{x\in\mathcal{X}} in(x)}{\sum_{x\in\mathcal{X}} con(x)}.$$

The locality of all traffic with respect to node set $\mathcal{X}$ is

$$l(\mathcal{X}) = 1 - \frac{\sum_{x\in\mathcal{X}} out(x) + \sum_{x\in\mathcal{X}} in(x)}{\sum_{x\in\mathcal{X}} gen(x) + \sum_{x\in\mathcal{X}} con(x)}.$$

Thus, the locality nicely shows the percentage of local traffic with respect to the overall traffic. The concept of traffic locality is useful to assess whether nodes within a subtree of an IB network mainly communicate with each other or whether they heavily depend on communication with nodes outside their subtree.

For a network with a tree structure, the visualization of the traffic locality is simple. Each subtree defines a set of nodes $\mathcal{X}$ for which the traffic locality can be computed and the obtained value can be associated with the root of the subtree which serves to color a node map. For general network topologies, the node sets $\mathcal{X}$ for the computation of the traffic locality need to be explicitly defined.

Link Utilization. Link utilization profiles show the time-dependent utilization of a link. These profiles are a powerful tool to enable network administrators to identify bottleneck links. In case no bottleneck link exists in the network, the profiles can be used to carefully downsize the backbone network while still providing the desired quality of service (QoS) to the customers. In addition, time-averaged utilization values are used to color a network map.

4 Architecture and Implementation

IBPM is implemented in a modular fashion which makes it easy to extend. Figure 1 provides an overview of the program structure. Each module is displayed together with its main features that are currently implemented. The application core is formed by Nagios, a computer and network monitoring software as controller, and Foswiki as corresponding graphical user interface. Open-source packages which are used in the modules include Gnuplot, RRDtool, Graphviz, and many more.

A monitoring run with IBPM normally consists of several steps. First, the topology of the network is automatically extracted, normalized, and stored for further processing. In regular intervals, performance data is collected from all nodes and switches in the network. In the configuration module, the user selects nodes and switches, and specifies measurement periods for statistical analysis. After enough data is collected, the general analysis of both topology and performance data is invoked and the analytical results are computed. Eventually the analytical results are interpreted by choosing one of the proposed comprehensible views, i.e., as network graph overlay or statistic.

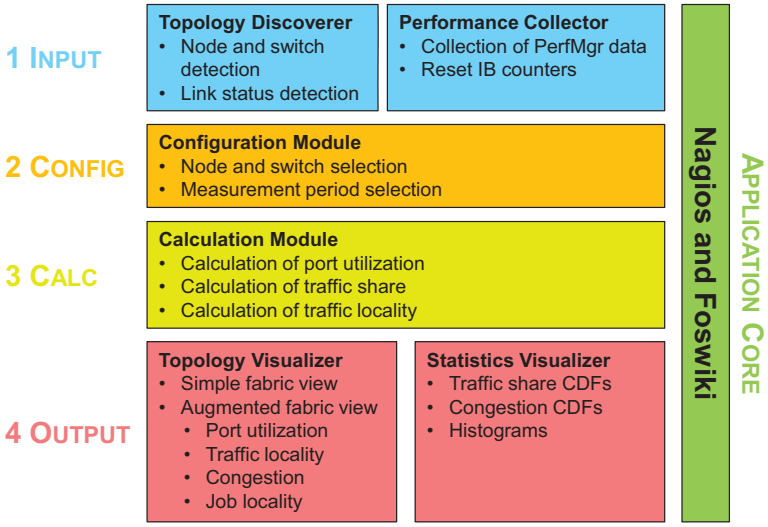

Fig. 1. Program structure of IBPM.

5 Conclusion

In this paper, we presented IBPM, an extensible framework for IB performance monitoring. IBPM offers a clear interface to network administrators to view a health map of the network and perform measurements. In addition, it provides many options for the visualization of the measured performance data. Our approach defines a set of visualization scenarios that are valuable for network administrators and operators of HPC data centers.

Acknowledgments

The authors would like to thank *science + computing ag* for providing them with access to IB test and production networks to test and evaluate IBPM.

References

1. InfiniBand Trade Association: The InifiniBand Architecture Specification (2011) http://www.infinibandta.org/ (last visited November 2011).
2. Mellanox Technologies: Unified Fabric Manager™Software for Data Center Management (2011) http://www.mellanox.com/ (last visited November 2011).
3. Qlogic Corporation: InfiniBand Fabric Suite (2011) http://www.qlogic.com/ (last visited November 2011).
4. OpenFabrics Alliance: OpenSM and InfiniBand Diagnostic Utilities (2011) https://www.openfabrics.org/ (last visited November 2011).