Visualizing the Invisible: InfiniBand and Rack-Level Monitoring for HPC Environments

Ashwini Wabale Aishwarya Bhagwat Sucheta Pawar Ashish Ranjan Sanjay Wandhekar
{ashwiniw, aishwaryab, suchetap, ashishr, sanjayw} @ cdac.in
Centre for Development of Advanced Computing, Pune, India

Centre for Development of Advanced Computing (C-DAC) is the premier research and development organization under the Department of Electronics & IT, Ministry of Communications & Information Technology, Government of India established in 1988 to spearhead India's supercomputing capabilities. As part of the National Supercomputing Mission (NSM), C-DAC emphasizes building expertise in HPC system management, including advanced interconnect technologies such as InfiniBand. This enables efficient management of large-scale HPC clusters and ensures reliable, high-performance network communication.

Abstract

High-performance computing (HPC) systems consist of thousands of compute nodes connected through InfiniBand networks that offer high bandwidth and low latency. As these systems scale up, real-time monitoring, diagnosing faulty nodes, and optimizing performance become more challenging. Existing HPC monitoring tools are isolated, measuring a subset of performance metrics and failing to relate logical behavior to the hidden physical infrastructure. These limitations decreased overall system awareness and slowed decision-making.

This work presents an integrated framework for real-time monitoring and visualization based on InfiniBand topology awareness and 3D rack-level visuals. The framework captures live port health, bandwidth utilization, and node metrics (CPU, memory, temperature) into an interactive web interface. It visualizes switch-to-node connectivity and highlights degraded links, bottlenecks, and anomalies across the cluster.

The framework has improved situational awareness and operational reliability. Administrators can identify link degradations and node-level issues in seconds, reducing downtime. Enhanced correlation between logical and physical layers enables faster fault localization, better performance tuning, and economic HPC management practices.

Need

- Unavailability of a visual representation of connecting the network and the hardware rack.
- Large InfiniBand fabric have complex, multi-tier topologies also placement of thousands of nodes in a multiple rack which is hard to visualize.
- Identifying faulty switches, nodes, links, ports with manually parsing logs is a tedious task for an administrator when the size is enormous which in result delays the troubleshooting
- Early detection of network and rack issues and initiate maintenance to ensure the stability of the system.

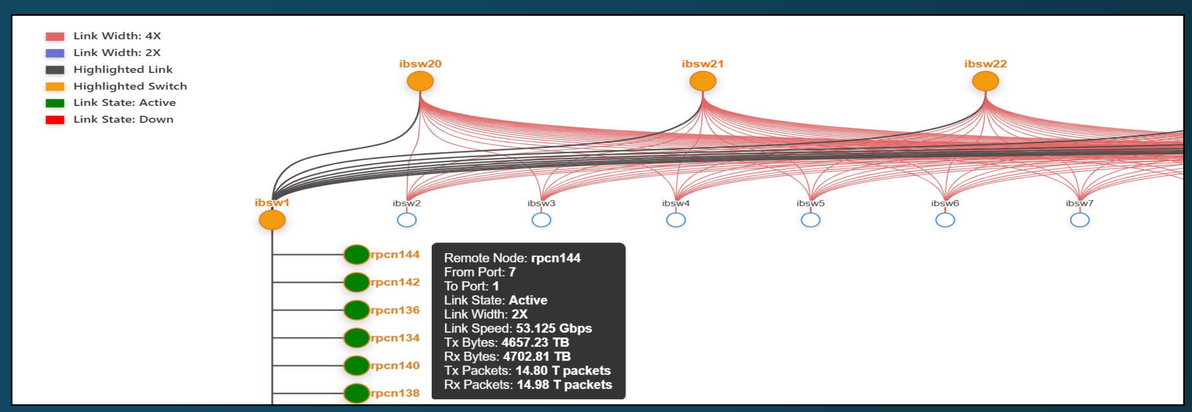

Fig [A]: Topology Port Link health and Traffic usage

Fig [B] : 3D Real-time Rack View

Features

- Visualization of logical and physical network topology including routing of the entire network including nodes, switches, ports and packet routes
- Subnet data management for the administrator
- Identification of congestion hotspots and tracking latency across nodes and switches
- Tracks system metrics to spot problems early and send alerts
- Color-coded network map and live metrics representation of the node
- Auto-generated network topology map helps admins understand fabric layout: Clickable links show stats of switches, nodes

Technology Stack

Backend (REST APIs)

Frontend

Databases

Facts and Figure

- **Load:** 0.5 % CPU only on the management node
- **Data Capture Frequency:** Default is 30 seconds and can be tuned as per requirements
- Loosely coupled Architecture for integration with any monitoring platform

Observations

- The framework provides real-time visibility of nodes per rack with a 3D view and interactive, color-coded visualization of the network topology for better understanding of network bottlenecks and anomalies
- Quick Troubleshooting with less downtime: Easy identification of faulty switch, node, link, port, error and congestion due to integration of structure and visible data metrics with 3D rack view allowing informed decisions about link management, and prioritizing maintenance work, as well as optimizing the overall network.
- Better job and application performance analysis: Integration with SLURM to see which nodes or links are more active during job run to find out communication imbalance
- Capacity planning and scaling decision: This helps administrators to rebalance the routes and add bandwidth when needed

Future Work

- Implement AI-assisted diagnostics to predict link failures, congestion, and performance issues using historical data collection
- Time series visualization to track error growth per h/w component to enable proactive replacement/repairing.
- Enhance visualization with 3D topology views, traffic heatmaps, and customizable dashboards for administrators and researchers.
- Integrate Rack level job scheduling/utilization and network monitoring with workload managers like Slurm or PBS to optimize resource allocation based on job-specific performance.