Achieving Large Bandwidth by Leveraging Parallelism in End-Hosts and Networks

Hirokazu Takahashi, Makoto Takizawa, Shoukei Kobayashi,
Osamu Kamatani, and Osamu Ishida
NTT Network Innovation Laboratories, Yokosuka, Japan
takahashi.hirokazu@lab.ntt.co.jp

Venkatram Vishwanath, Sungwon Nam,
Luc Renambot, and Jason Leigh
Electronic Visualization Laboratory,
Chicago, Illinois, USA

***Abstract*—This paper describes the first experiment on MultiRail and MultiLane technologies using global networks. These technologies leverage end-host and network parallel resources, e.g., processor cores and lambda paths, to achieve large bandwidths.**

I. Introduction

The demand for terabit/sec (Tbps) level large bandwidth applications is increasing. For example, streaming of uncompressed super high-definition (SHD) [1] video requires a bandwidth of approximately 6 Gbps for 30 frames/sec (fps) and 12 Gbps for 60 fps. Multi-channel streaming of such large streams requires a bandwidth of 100 gigabit/sec (Gbps) to 1 Tbps.

The development of an infrastructure technology for such applications called Terabit LAN [2] is expected. The research goal of the Terabit LAN is to provide up to 1 Tbps end-to-end bandwidth over virtual LANs on lambda networks. The Terabit LAN includes both end-host and network technologies to achieve the bandwidth.

Fig. 1 shows trends in end-host microprocessor technology and network fiber communication technology. Both trends show similar paradigm shifts toward parallel resources. Because of the physical limitations of the clock frequency per processor core and bit-rate per lambda, parallel resources, i.e., multi-processor/core and multi-lambda/fiber, were applied to improve the total performance.

In this paper, we describe the first experiment on MultiRail [3] and MultiLane [4] technologies using global networks. These technologies leverage end-host and network parallel resources to ensure a large bandwidth.

II. MultiRail End-Host Technology

Fig. 2 shows the typical structure of the non-uniform memory access (NUMA) architecture, which is becoming mainstream for multi-processor end-host systems. Each processor can access a memory module and a network interface card (NIC), which are connected to the bus of another processor. However, due to the inter-processor communication overhead, the accessing latency is longer than that to access the processor's own bus, and its bandwidth is degraded. To provide a large bandwidth to the applications, process scheduling should be done based on the topology of the end-host architecture, unlike traditional scheduling based on the processor load.

We proposed the MultiRail technology in [3] to establish the scheduling. As shown in Fig. 3, the MultiRail technology prepares resource sets, each of which is called a rail, and allocates each rail to the applications. Since the rails are based on the topology of the end-host architecture, the applications can use the large bandwidth of the rails.

We developed an uncompressed SHD video streaming application software based on the MultiRail technology called

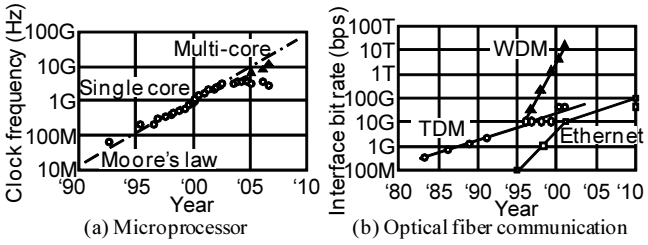

Fig. 1. Technological trends.

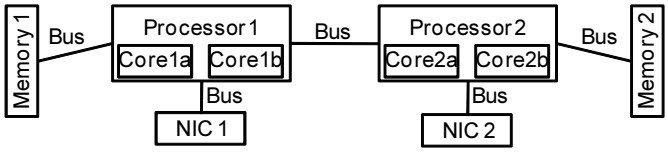

Fig. 2. Example of non-uniform memory access (NUMA) architecture.

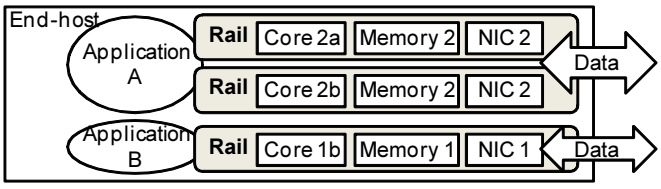

Fig. 3. MultiRail technology.

NetV to confirm the advantage of the technology. The target of NetV is to handle streams of over 10 Gbps, e.g., 60 fps SHD video, and NetV has the capability to handle two 10 Gbps NIC, i.e., two rails. The NetV splits the video image into two parts, an upper half and a lower half, and sends them through each rail.

Fig. 4 shows the achievable sending frame rates of NetV running on the end-hosts shown in Fig. 2 under various rail conditions. As described above, condition A is the optimal condition based on the topology and it achieves the maximum performance of 89 fps.

III. MultiLane Network Technology

We define a lane as a single lambda path that includes one or more optical / electrical switches, and MultiLane as a network technology that leverages parallel lanes to establish large bandwidth communications among end-hosts.

As one of the MultiLane technologies, we proposed packet-based lane bundling (PLB) [4]. PLB can bundle parallel lanes as a single large bandwidth network path. Fig. 5 shows the details of the PLB. A sender side NIC distributes packets to

	Condition A		Condition B		Condition C		Condition D	
	Rail1	Rail2	Rail1	Rail2	Rail1	Rail2	Rail1	Rail2
Core	1a	2a	2a	1a	1a	1b	1a	1a
Memory	1	2	1	2	1	1	1	1
NIC	1	2	1	2	1	2	1	2

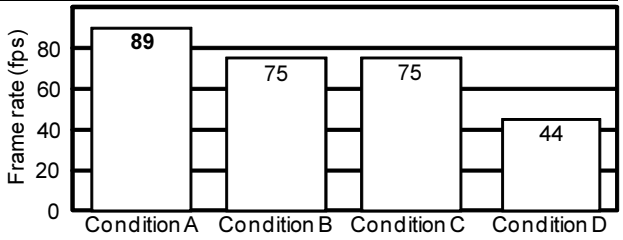

Fig. 4. Achievable frame rates under various rail conditions.

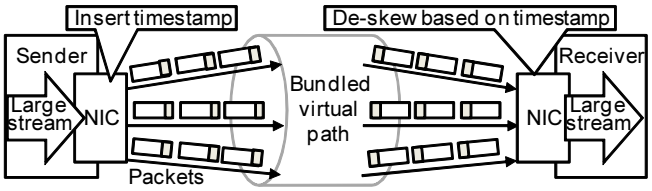

Fig. 5. Packet-based lane bundling (PLB).

parallel lanes, and the receiver side NIC collects the packets. Since each lane has different transmission delays, the packet arrival sequence and timing on the receiver side is different from that on the sender side. This could be a problem in particular for the playback quality of video streaming. To overcome this problem, the PLB de-skews the delay difference. The sender side NIC inserts time stamp information into each packet and the receiver side NIC de-skews the packets based on the information.

We developed a prototype NIC called the Terabit LAN NIC (Fig. 6) to test the MultiLane technologies. The current Terabit LAN NIC is for the advanced telecom computing architecture (ATCA) platform and a 40-Gbps interface (four 10 Gbps interfaces) is equipped at the front and backplane sides. Moreover, the Terabit LAN NIC can achieve over 40 Gbps by combining boards.

We implemented PLB in the Terabit LAN NIC and evaluated the de-skewing performance. In the evaluation, two 10 Gbps Ethernet lanes, one is a back-to-back connection and the other has a 20 μsec delay and 10 μsec (HWHM) jitter from the delay emulator, were bundled, and the delay variations of de-skewed packets were measured. The resultant variations were nearly identical to the results without delay emulation and the differences between them were approximately 10 nsec. This means that PLB and its de-skewing function can bundle parallel lanes as a single large bandwidth stable delay network path.

IV. Joint Experiments Using Global Networks

We conducted a joint experiment of the MultiRail and the MultiLane technologies using global networks. Fig. 7 shows the configuration for the experiment. Each of two end-hosts in Yokosuka, Japan, sent 30 fps uncompressed SHD video streams and the end-hosts in San Diego, CA., displayed the received video stream to each tiled display. One stream was transferred using two bundled lanes through the Terabit

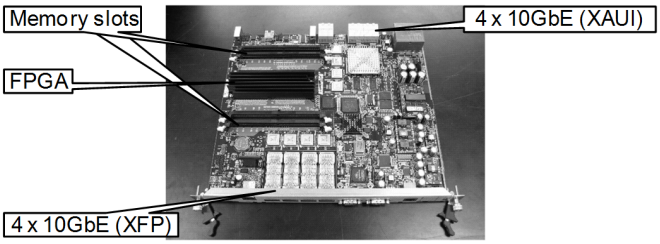

Fig. 6. Prototype of implemented terabit LAN NIC.

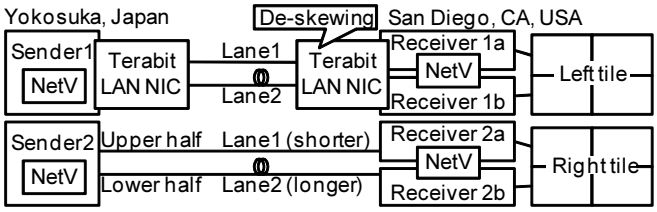

Fig. 7. Configuration for joint experiment.

Fig. 8. Received video streams.

LAN NICs, and the other stream was simply transferred using the two lanes without de-skewing. The two lanes had geographically different paths and the delay difference was approximately 150 msec. Fig. 8 is a photograph of the received video streams displayed on the tiled display. The image to the right without de-skewing is misaligned because of the delay difference. On the other hand, the image on the left is aligned using the PLB de-skewing. The results show that the MultiRail technology handles the large bandwidth on end-hosts and the MultiLane technology enables stable large bandwidth delay transfer on global networks.

V. Future Work

In future work, the connection of rails and lanes should be considered. If the bandwidth of a rail and a lane are equivalent, these can simply be connected one-to-one. However, if there is a significant difference in the bandwidths, one-to-one connection cannot leverage their bandwidth. Connection methods that can optimize the use of the bandwidth should be considered.

References

[1] D. Shirai, et al., *6 Gbit/s uncompressed 4K video stream switching on a 10 Gbit/s network*, ISPACS2007.

[2] O. Ishida, *40GbE, 100GbE, and Terabit LAN Challenges*, 3rd FON, OECC/IOOC2007.

[3] V. Vishwanath, et al., *The Rails Toolkit (RTK) – Enabling End-System Topology-Aware High-End Computing*, e-Science2008.

[4] S. Kobayashi, et al., *Packet-based Lane Bundling for Terabit-LAN*, APCC2008.