Optical Packet Routing in Distributed Grid Computing Architectures

Odile Liboiron-Ladouceur and Keren Bergman
Columbia University, Department of Electrical Engineering
500 West 120th Street, New York 10027; ol2007@columbia.edu

Abstract *Requirements for timing alignments among optical packet switched clusters in grid computing are developed and expressions for required packet guard times and fibre transmission lengths are obtained. Timing tolerances are verified in an experimental demonstration.*

.

Introduction

Over the last decade, grid computing emerged as a viable method for efficiently harnessing the computing power of geographically distributed clusters of processors. Projects such as DRAGON, USN, CHEETAH, as well as others recently reported in [1], have demonstrated the advantages of leveraging distributed resources into a single focused application using an optical control plane. Large-scale distributed grid applications can approach multi-terabits per second of throughput between clusters separated by 100 km with nearly time-of-flight latencies. The current compute nodes in these systems consist primarily of electronic interconnects. In this work we consider optical packet switched (OPS) networks as a potential high-capacity platform for interconnection of the high-performance compute node clusters within the grid. An envisioned simple grid computing architecture that employs three 12x12 OPS networks based on the data vortex [2] to interconnect geographically dispersed high-performance clusters is shown in Fig. 1. Each data vortex cluster is connected to the grid through high throughput transmission fibre optic pathways, and multiple processors and shared memory elements are interconnected within each cluster.

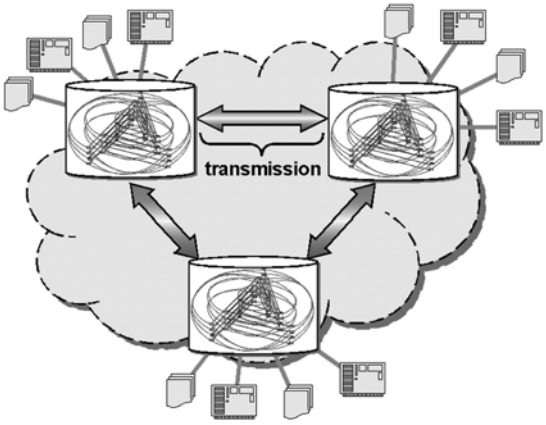

Fig. 1: Three Compute Node Data Vortex Distributed Grid Architecture.

In this work we derive an expression for the minimum guard time in the optical packet protocol required to compensate for the effect of chromatic dispersion on the transmission of multi-wavelength packets. Additionally, an expression for the transmission fibre optic length is obtained for maintaining timing alignments among bufferless OPS clusters. The network tolerance to timing mismatch of incoming packets from other clusters is measured in an experimental demonstration.

Requirements on packet transmission

The data vortex packet format encodes data in both the time and wavelength domains [2]. Due to chromatic dispersion, timing skew between the multiple wavelength channels occurs as packets propagate between clusters through the long distance transmission. Additional guard time is inserted at the tail of the payload channels to compensate for the timing skew on the longest wavelength (Fig. 2). A frame signal indicating when valid data is being transmitted, is encoded on the shortest wavelength. The required dispersion guard time (T_{DG}) is a function of the distance between two data vortex clusters and the type of transmission fibre employed. For a certain length *L* of dispersion-shifted single-mode fibre (DS-SMF), the dispersion guard time required can be calculated from the dispersion slope at the zero dispersion wavelength ($S(\lambda_o)$) taking into consideration the shortest (λ_S) and the longest (λ_L) wavelength of the multiple payload channel packet structure. For two clusters separated by 100 kilometres of dispersion-shifted fiber ($S(\lambda_o)$=0.075 ps/nm^2•km, λ_o=1539.4 nm, λ_S=1530.0 nm, λ_L=1560.0 nm) the required dispersion guard time is 1.26 ns corresponding to only 5% overlap for a 25.6 ns slot time used in data vortex clusters.

$$T_{DG} = \frac{S(\lambda_o)\cdot L}{2}\left\{\left(\lambda_L^2 - \lambda_S^2\right) - 2\lambda_o\left(\lambda_L - \lambda_S\right)\right\} \qquad \textbf{(1)}$$

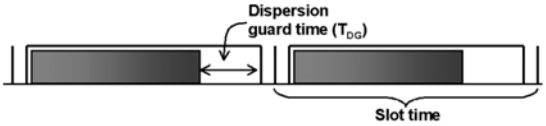

Fig. 2: Packet format with frame signal and added guard time at the tail of the data payload.

To maintain system synchronization between clusters, incoming packets must arrive coincident with the start of the packet slot time intervals of the local data vortex cluster. As in many OPS systems, packet slot times are preserved by the design of the routing path latencies since dynamic buffering is not available [2]. The length of the transmission fibre optic pathway between two clusters should therefore be a multiple *m* of the length corresponding to the packet slot time *T*.

For a packet slot time of 25.6 ns, the fiber length must be a multiple of 5.23 meters (c_o speed of light, n=1.4682).

$$L = m \cdot T \cdot \left(\frac{c_o}{n} \right) \quad \textbf{(2)}$$

Experimental demonstration

The tolerance to timing mismatch of incoming packets is investigated using the experimental demonstration illustrated in Fig. 3. To generate packets of the correct format for the implemented 12×12 data vortex architecture, four routing addresses and a frame signal are required. The routing header wavelengths are modulated with the address information, which are constant throughout the duration of the packet. They are then coupled to the 16-wavelength packet payload modulated at 10 Gb/s. Packets propagate through five internal nodes in the data vortex cluster. One of the packets exits the cluster and propagates through 1.5 km of DS-SMF fibre before effectively reaching another cluster.

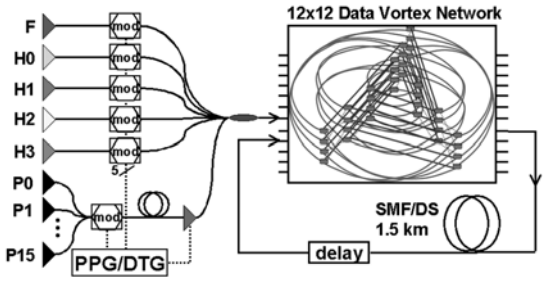

Fig. 3: Implemented 12 × 12 data vortex network in a grid computing experimental demonstration.

In this demonstration, the packet is injected back into the same data vortex cluster through a different input port (port C in Fig. 4). In the absence of packets inside the cluster propagating from node B to node A, the incoming packet is routed to node A, based on its encoded header address. In the presence of concurrent packets, a deflection electrical signal is emitted from node B to inform node C and prevent the incoming packet from colliding with an existing packet at node A. In the data vortex, deflection signals are transmitted from nodes on downstream stages (nodes B and A) to nodes on the adjacent upstream cylinder (nodes C and D) for contention resolution [2]. The electronic deflection signals must be timed correctly for the incoming packet to be efficiently deflected to node D. The network tolerance to the arrival time of an incoming packet from a distant cluster is measured by changing the total length of the transmission fibre optic pathway, thereby delaying the packet arrival time. For a packet to be deflected efficiently, switching node C must receive the deflection signal from node B at the appropriate time, so that the routing decision can be executed while the incoming packet is within node C. If the packets arrive too early or too late, the deflection signal timing requirement is not met and leads to an improperly timed deflection signals. Consequently, the incoming packet is truncated at node D with the truncated part colliding with the packet present in node A (Fig. 4). The tolerance can be measured for the range of delays for which packets do not get truncated. Additionally, the time margins for incoming packets are similar to the ones for packets propagating within the data vortex network [3]. The length accuracy of the transmission fibre optic pathway must be within ±5 cm.

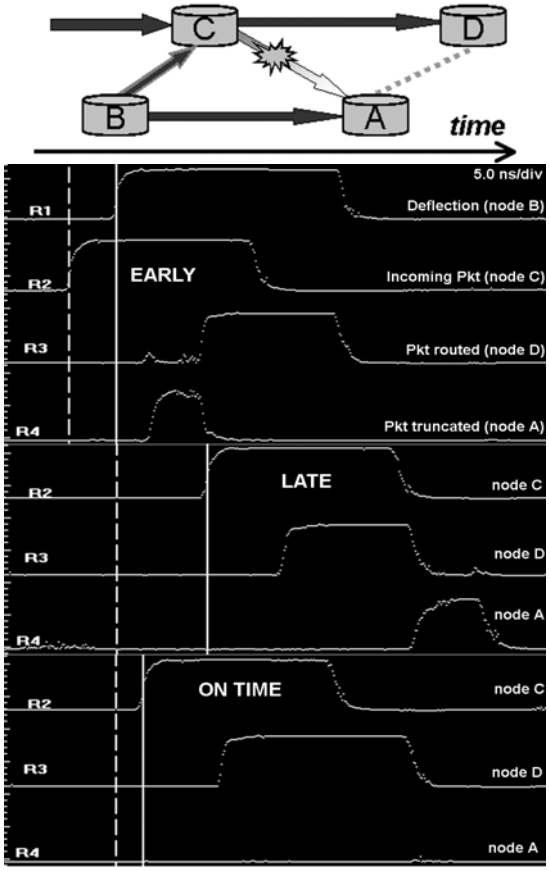

Fig. 4: Illustration of packet deflection (top). Electrical waveforms generated by the nodes (bottom).

Conclusions

We investigated the requirements for maintaining synchronization in a simple distributed grid computing architecture that interconnects among high-capacity optical packet switched clusters. To compensate for the chromatic dispersion of the transmission fibre between clusters, an expression for the required guard time in the packet structure was derived. Additionally, an expression for the transmission fibre length based on the packet injection period was found and its required accuracy was demonstrated experimentally.

References

1 IEEE Comm. Magazine, 44 (2006), pp. 62-131.
2 A. Shacham et al., JLT, 23 (2005), pp. 3066-3075.
3 B.A. Small et al., PTL, 11 (2005), pp. 2478-2480.