Considerations of point-to-multipoint QoS based route optimization using PCEMP

Dipnarayan Guha[1], Seng Kyoun Jo[1], Doan Huy Cuong[1], and Jun Kyun Choi[2]

[1] Researcher, BcN ITRC, Broadband Network Laboratory,
Information and Communications University,
119 Munji-Dong, Yuseong-Gu, Daejeon 305-714, Republic of Korea
{dip, skjo, cuongdh}@icu.ac.kr
[2] Associate Professor, BcN ITRC, Broadband Network Laboratory,
Information and Communications University,
119 Munji-Dong, Yuseong-Gu, Daejeon 305-714, Republic of Korea
jkchoi@icu.ac.kr

Abstract. This paper describes the basic concepts of point-to-multipoint (p2mp) path computation on the basis of the Path Computation Element Metric Protocol (PCEMP). PCEMP, being soft-memory based, has the capability of dynamic configuration of its finite state machines (FSMs) in the participating PCEMP peers, and thus can support a wide variety of traffic engineering techniques that are needed to guarantee bandwidth demand and scalable fast protection and restoration in PCE based p2mp frameworks ensuring end-to-end QoS support. The authors have proposed this concept in the newly constituted PCE WG (Path Computation Element Work Group) in the RTG sub-area of the IETF. In this research-in-progress paper, we show how PCEMP as it is defined, and the optimal number of PCE Domain Areas (PCEDAs) that might be allocated to a PCE node for the best performance in end-to-end QoS management based on a tight optimal Cramer Rao bound for the state machine executions.

1 Introduction

One of the key work items involving the functional specification of MPLS and GMPLS Traffic Engineering LSP path computation techniques in the proposed PCE WG [1] charter is the case for TE LSP path computation for inter-domain areas applying to both point-to-point (p2p) and point-to-multipoint (p2mp) TE LSPs. Most of the existing MPLS TE allows for strict QoS guarantee, resource optimization and fast failure recovery, but the scope is mostly limited to p2p applications [2]. In the context of path computation, one of the important application areas is the reliable support of bandwidth-on-demand applications, where the QoS provisioning needs to be dynamic and robust. A scenario where a PCE node acts a server which are connected to several clients, which may or may not be PCE peers, needs a clear requirement addressal so far as p2mp TE tunneling is considered. In this paper, we consider that such p2mp TE LSP path computation is QoS triggered, and we show how PCEMP finite state machines (FSMs) might help in achieving a scalable architecture involving PCEDAs

where p2mp path computation metrics are independent of the number of clients to which the PCE server is attached to. The Path Computation Element Metric Protocol (PCEMP) [3] acts as a generic computational model for path based metrics in large multi-domain and/or multi-layer networks. This paper also shows that feature of the PCEMP protocol in degenerating the setup and teardown of p2mp TE LSP computation to the PCEMP protocol processing itself, thus enabling support of an arbitrary number of clients as well provisioning of guaranteed robust path protection and restoration and dynamic QoS provisioning for bandwidth-on-demand services [4].

2 p2mp QoS based path computation fundamentals

For the scenario involving robust and dynamic provisioning of bandwidth-on-demand services, the p2mp applications request p2mp forwarding paths in case of different topology deployments. The robustness must be thought in the context of path re-optimization, so a quick change in the topology must be accommodated with every PCEDA level optimization. The p2mp path will have several observed metrics as constraints, such as cost of path establishment and teardown, delay bounds of the p2mp path, both delay bounded and cost optimized constraints in tandem for path computation, etc. One of the features as brought out in the PCE WG charter is the co-existence of different path computation algorithms on the PCE node, so that depending upon the data that is processed, a particular algorithm is invoked. It is also evident that for p2mp applications, a CPU intensive path computation is necessary, primarily because most of the bandwidth-on-demand applications tend to be resource-intensive applications like streaming multimedia, real-time videoconferencing, etc. The ideal thing would be to let the data that is under processing in the PCE node determine the path computation algorithm directly, which would mean that the constraints imposed by the QoS provisioning requirements would directly determine the path computation algorithm and path re-optimization, which in turns drives the resulting topology architecture. Thus, it is easy to see why PCEMP is a possible solution for p2mp TE LSP computation, as it drives a protocol driven architecture for topology changes in path re-optimization based on QoS constraints. The traffic engineering techniques involved with p2mp TE LSP computation involve mainly with the case of p2mp path computation over multiple domains. There are three main issues involved with this feature: 1. load sharing among paths, 2. ability to modify the p2mp paths in different PCEDAs even when the PCEDA entities lie in different multiple domains, 3. p2mp path computation for corresponding clients in muliple domains must be able to support scalability, i.e. the number of clients entering/leaving the p2mp tree at a given time.

3 Analysis of QoS based p2mp path computation using PCEMP

This protocol driven inter-domain network environment architecture requires performance analysis techniques that can accurately model QoS driven protocols. In our

model we adapt a technique that is a modification of a trajectory splitting simulation technique based on direct probability redistribution (DPR) [6,7]. We have developed techniques and demonstrate their utility by applying DPR to a network model that includes inter-domain autonomous systems, and a detailed PCEMP protocol. DPR was found to work best in settings where equalizing the number of samples in each of the subsets (subsets of the state space of the system) for probability tree redistribution occurred. In our model, derived from the PCEMP State Machines design, probability is redistributed from the subsets of states with high probability to the subsets with small probability. One consequence of this redistribution of probability is a decrease in the accuracy of estimates corresponding to high probability subsets. We now see how QoS based TE LSP path computation can be modeled using this concept.

3.1 Simulation of QoS based p2mp TE LSP path computation using PCEMP

In this section we will explain the basic principles of DPR -based splitting, a more complete presentation can be found in [6], [7]. In DPR, we partition our state space S into m mutually exclusive, non-empty subsets $S_1, S_2, \cdots, S_m$. Each observation of the simulation is mapped into the appropriate subset according to function $\Gamma(X_i)$ which is given by,

$$\Gamma(X_i) = j,\ j = 1, 2, \cdots, m \tag{1}$$

where $X_1, X_2, \cdots, (X_i \in 1, 2, \cdots, m)$ are the observations of a discrete-time Markovian system. DPR modifies that transition probability matrix of the system such that each state has a new steady-state probability which can be written as,

$$\pi_i^* = \theta \pi_i \mu_{\Gamma(i)},\ \ i = 1, 2, \cdots, n \tag{2}$$

where n is the number of states, and π_i and π_i^* are the steady state probabilities before and after the DPR process for state i. θ is the normalization constant which is given by,

$$\theta = \frac{1}{\sum_{i=1}^{n} \pi_i \mu_{\Gamma(i)}} \tag{3}$$

The oversampling factor $\mu_{\Gamma(i)}$ is the main parameter that controls the modification of the steady-state probabilities. We assume that $\mu_1 = 1$ and $\mu_1 \leq \mu_2 \ldots \leq \mu_m$, which can be achieved for arbitrary ordering of the subsets by changing the ordering and the indexes of the oversampling factors (μ values).

3.2 Functional Parameters for p2mp TE LSP computation in PCEMP structures

The input data sequence is arranged into an ordered set called the Input Data Type (IDT) which is a subset of the input vector S and a function of the network transform to be computed T. A State Subset is a member of the cardinal product of S and T. It is shown to be isomorphic with the logical decoder outputs [3, 9]. The IDTs invoke the hardware for computing across the partitioned kernel in the PCE nodes.

Input: IDT T_j, State Subsets S_l and S_m, Integers l and m, Label L_b, Semi-Ring R. For p2mp support, there will be multiple state subsets, and we will pairwise consider all such states. In case where the total number of states is odd, one state will be paired with the identity state.

Output: Flow metric/measure p(A,B), which maps to the PCE descriptor ID. For p2mp cases, the PCE descriptor IDs are setwise collected to form the pp2mp ID.

3.3 QoS based path computation using PCEMP

Concept: Iterative applications of the PCEMP DS. Two or more IDT encoders separated by an interleaver (respectively CC and SPC). This structure suggests a decoding strategy based on the iterative passing of soft-computing information between two decoding algorithms. This is equivalent to dynamically partitioning the path computing engine core into sub-units that are linked together real-time based on the input data and the protocol handler. This is what makes PCEMP a protocol driving architecture, and is one of the key features of realizing a NP-hard path computation for p2mp TE LSPs.

Basic Computation: Configure PCEMP DS to compute soft-decisions in terms of measures. An assigned measure is given to each branch of the IDT. This algorithm makes the data intensive path computing much easier and reduces overhead and complexity and is incorporated in the computing core. It also guarantees disjoint path computation that enables fast end-to-end backup and protections. The configuration is totally dependent on the processed data and in a PCE server based bandwidth-on-demand scenario, can be triggered by the QoS service classes. The QoS classes are directly mapped onto the IDT, and thus can realize the p2mp based TE LSP path computation and re-optimization all the time based on the demanded bandwidth ensuring robustness and reliability of services. This follows directly from the PCEMP protocol architecture, details of which can be found at [3].

Section of the pseudo-code for the PCEMP FSM execution algorithm, Guha, Dipnarayan et al, Path Computation Element Metric Protocol, IETF Internet Draft, July 2005

```
for
{
any TE LSP passing through a PCE node P
{
Initialization..
Loop..
```

```
//QoS based p2mp TE LSP path computation support BEGIN //
do
{
repeat for all Si's;
assign boutS for each si in Si for P;
}
//QoS based p2mp TE LSP path computation support END //
..
//QoS based p2mp TE LSP path computation support BEGIN //
do
{
repeat for all Si's;
assign p for all c0's for all Si's;
take the weighted arithmetic mean of the probabilities
along with the branch labels;
assign p1 = probability weighted (mean(c0));
}
// QoS driven p2mp TE LSP path computation support END //
```

$\log p(S_i, x_j) = \log sum(s_i, x_{j-1}) + \log sum(z_j)$, $for \forall b \in B_i$ . .

$p(s_i, x_j) = \min \log(sum(s_i, x_{j-1}))$ . . Stop

3.4 Cramer Rao Bound considerations for the QoS based PCEMP algorithm

With reference to [3], the probability density function of each data block of the PCE input c1, chosen without loss of generality, conditioned on the variable c0, the common symbols between the two encoders, CC and SPC and the other PCE input x1 is given by

$$p_{c_1/c_0,x_1}(c_1(n)/c_0, x_1(n)) = \frac{1}{(2\pi\sigma)^N} e^{-|c_1(n)-C_0 F x_1(n)|^2/(\sigma)^2} \quad (4)$$

where N is any arbitrary integer representing the cardinality of the data blocks, σ being the standard deviation of π_i, the steady state probability that the spanning tree is in state i, as in (2), and F is a unitary matrix whose rank varies as a function of i and N. C_0, F, x_1 are convoluted together. As the cardinality of the data blocks are uncorrelated, the joint pdf of a block $C_{1M} = [c_1(1),\ldots\ldots c_1(M)]$ is given by:

$$p_{C_1(M)/C_0,X_1(M)}(C_1(M)/C_0, X_1(M)) = \frac{1}{(2\pi\sigma)^{NM}} \exp(-\sum_{n=1}^{M} | c_1(n) - C_0 F x_1(n) |^2 / \sigma^2) \quad (5)$$

To simplify the Cramer Rao Bound derivation for determining the PCEMP IDT map and the optimal number of PCEDAs to be allotted to a PCE node, we now introduce a parameter vector θ, as

$$\theta = [c_0^{\ T}, x_1^{\ T}(1), ..., x_1^{\ T}(N), c_0^{\ H}, x_1^{\ H}(1), ..., x_1^{\ H}(N)] \quad (\mathbf{6})$$

and write the log-likelihood function as (6)

$$f(\theta) = \log p_{C_1(M)/C_0, X_1(M)}(C_1(M)/C_0, X_1(M)) = -\log((2\pi\sigma)^{MN}) - \sum_{n=1}^{M} \frac{|c_1(n) - C_0 F x_1(n)|^2}{\sigma^2} \quad (\mathbf{7})$$

This θ is related to the state probabilities as in (3).

We define L as the instantaneous data block partitioning that the CC and SPC enforces on the data block currently under processing, i.e. the cardinality of the data blocks are further partitioned into L for the purpose of parallel processing. It becomes easy to model equation (6) for the bound by introducing the 2(L+1)NM X 2(L+1) NM complex Fisher's Information Matrix, as in [8], thus

$$J(\theta) = E_{C_1(M)/C_0, X_1(M)} = \{(\frac{\partial f(\theta)}{\partial \theta^T})^H \frac{\partial f(\theta)}{\partial \theta^T}\} \quad (\mathbf{8})$$

where $\frac{\partial f(\theta)}{\partial \theta^T}$ is a 1 X 2(L+1)NM row vector. It has been proved in [8] that the CRB of θ can be found by considering the reduced (L+1)NM X (L+1)NM matrix thus:

$$J'(\theta) = E_{C_1(M)/C_0, X_1(M)}\{(\frac{\partial f(\eta)}{\partial \eta^T})^H \frac{\partial f(\eta)}{\partial \eta^T}\} \quad (\mathbf{9})$$

where $\eta = [C_0^{\ T}, X_1^{\ T}(1), ...X_1^{\ T}(M)]$. [8] goes on to derive the CRB for a fixed N, L and M. In our case, we assume that N and L are only fixed for a given observation time interval, that is, when the first route tree alignment corresponding to the first state happens. The model in [8] thus becomes finitely discontinuous at the end of each observation time interval. This doesn't prevent us from calculating the reducing CRB for the PCEDA allocation optimum. What we do is introduce a matrix W' and relax two restrictions which have been assumed in [8], 1. The size of the matrix can be variable. For the purpose of considering a restricted observation time interval and for applying [8]'s model on the piecewise continuous bound in time, we pad the matrix W' with ones so that there is no change in its' overall rank. This essentially means to mask the data block bits out by a series of 1's in the CC-SPC unit so that the overall logic function does not change. 2. The second relaxation is that the columns

of W' need not form an orthonormal basis for the null space of C_0. In our masking bit scheme, it is possible to derive a secondary matrix W'' from W' that forms this basis, which is implemented into the processing hardware that is programmed with transforms, as in [9]. We thus have, from (1), and along the lines of [8], defining L' as derived from the bit masking process,

$$C = \frac{W'}{\sigma^2}\{W'^H \Gamma[I_{L'XL'} \otimes (F^* X_{1M}^* X_{1M}^T F^T)]\Gamma^H W'\}^{-1} W'^H = \frac{1}{\sigma^2}\Gamma[I_{L'XL'} \otimes (F^* X_{1M}^* X_{1M}^T F^T)^{-1}]^{-1}\Gamma^H \qquad (\mathbf{10})$$

It is easy to see that the maximum value of this function (9) is 1. We had started our analysis with c_1, with one of the triads of (c_0, c_1, x_1) arbitrarily without loss of generality. The elements of this triad are independent of each other, and thus, the optimal bound of the data-driven computation function is thrice the value of C that we have computed in (9), which effectively means, that if there is one-to-one map from the PCEMP IDT to the number of PCEDAs allocated to a PCE node, the maximum number of PCEDAs optimally handled for best TE LSP path computation is 3. We shall confirm this result in this paper in our simulations section.

3.5 QoS oriented operation considerations of p2mp in PCE

As we have said before, it is possible to obtain a protocol driven network architecture from a data driven protocol FSM. From the operation point of view, there are two equally likely possibilities for QoS oriented p2mp support in PCEs:

1. The PCE descriptor ID that is obtained from the FSM execution can be carried as a separate optional object in standard OSPF/RSVP-TE extensions, irrespective of whether a routing or a signaling based solution is deployed for TE LSP path computation in p2mp scenarios. Traditionally, if an explicit-route object is used, the PCE descriptor ID can be used in conjunction with it as a sub-object. It is easy to understand that a path change will essentially mean changing the contents of the explicit-route object and/or inserting/deleting a new one for the purpose of p2mp support. The pp2mp ID, which is added on once the PCE descriptor IDs are added with the explicit route object being processed by every next hop, will be thus spanned over the entire PCEDA. At the egress side, the total pp2mp ID is recognized and the LSP contents mapped onto the corresponding client paths.

2. The other option is to have a separate messaging system for PCEMP that only has the PCEMP header and the PCE descriptor ID as the payload. The PCE nodes can maintain a local counter for these IDs, which are generated randomly but become fixed for any set of adjacent path computation. The scope of this deployment is again implementation specific. It might act as an encapsulated packet within standard routing or signaling protocols, or may be run independently before control and management information is exchanged, or may be periodically run to maintain the "soft-state" like conditions. In either case, the p2mp TE LSP path computation is independent of the number of clients (or end points) that are attached to the PCE node, resulting in clear scalability enhancements. It is also evident that the make-before-break condi-

tions in modifying p2mp TE LSPs can be easily done without much overhead and computation intensive operations.

Based on the event states of the protocol, the corresponding trajectory partitioning is mapped onto the PCEDAs, enabling a correct performance analysis model for PCEMP based inter-domain architectures. Including this correlation, the estimator for the variance in PCEMP IDT maps, from (1), (2), (3), (6) and (10) is

$$\begin{aligned} \mathrm{var}(\hat{\pi}_i) &= \mathrm{var}\Big(\sum_{k=1}^{N} \Lambda_i(X_k)/N\Big) \\ &= \frac{1}{N^2}\mathrm{var}\Big(\sum_{k=1}^{N} \Lambda_i(X_k)\Big) \\ &= \frac{1}{N^2}\left\{ N\sigma_i^2 + 2\sum_{j=1}^{N-1}(1 - j/N)\rho_j\sigma_i^2 N \right\} \\ &= \frac{\sigma_i^2 R}{N} \end{aligned} \tag{11}$$

3.6 Simulation Results

We have defined a set of PCEMP messages and their types [3] recently at the IETF PCE WG, where we have used 8 different test case scenarios for the PCEMP driven inter-domain network environments where we performed queuing analysis and timing delay analysis. The simulations were carried out by developing external process modules for OPNET Modeler 8.0 Here are some of the results that we got by porting these external process modules to OPNET Modeler 8.0. The equations for modeling are based on (1), (2), (3), (4), (5), (10) and (11). We have developed a path computation and routing protocol software for this purpose based on our analysis and the definition of the protocol [10].

The first result is a classic where we get the results for the Probability Mass Function (PMF) of the NXON-OFF/D/1/K queue for N = 11, pIA = 8.55e-3, pAI = 0.33 and K = 45. The values were considerably in the region of 10e-7 for different normalized queue lengths, and though the trend was similar to [6] and [7], the results used in the PCEMP context seem to be better and provide a better upper bound to the optimal stable queue length. The piecewise continuous PMF over the observation interval is deduced from equations (4) and (5). Equations (6), (7), (8) and (9) help in computing this function and port it onto PCE software.

The second simulation result is carried out using equations (10). As discussed in section 3.4, this shows why the optimal number of PCEDAs is 3, based on the independence of the triad (c_0, c_1 and x_1).

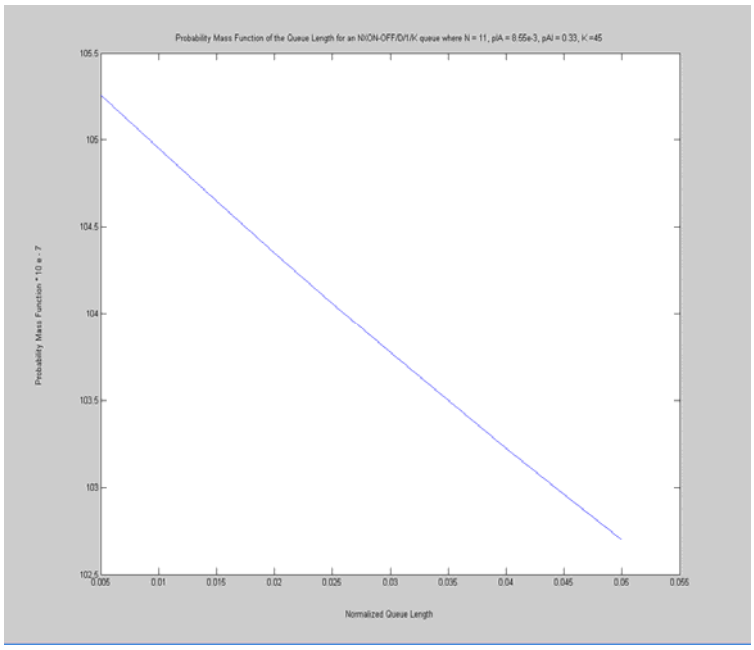

Fig. 1. PMF of the M/D/1/K model of the PCE architecture vs. normalized queue length

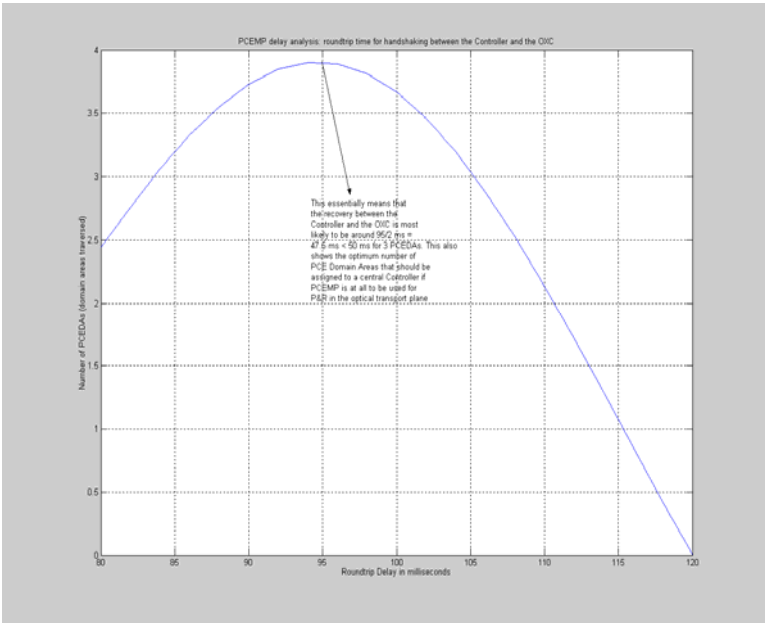

Fig. 2. Optimal number of PCEDAs based on the Cramer Rao bound from QoS based PCEMP

4 Conclusion

PCEMP helps the p2mp participating nodes to advertise their capabilities based on the set of constraints that it can support. Using PCEMP for p2mp TE LSP path computation and global re-optimization also serves the dual purpose of topology reconfiguration. This paper provides a general framework for p2mp support in PCE architectures, and is based on a scenario where the p2mp path computation is triggered by a QoS requirement. The authors are pursuing the standardization of this protocol in the PCE WG for an efficient inter-domain TE LSP path computation, and this is under current discussion within the scope of the newly constituted PCE WG at the IETF.

5 Acknowledgement

This work was supported by the in part by the Korea Science and Engineering Foundation (KOSEF) and the Institute of Information Technology Assessment (IITA) through the Ministry of Information and Communications (MIC), Republic of Korea.

References

1. Farrel, A., Vasseur, J.P., Ash, J.: Path Computation Element (PCE) Architecture, IETF Internet Draft,July2005, http://www.ietf.org/internet-drafts/draft-ietf-pce-architecture-01.txt
2. Yasukawa, S. (Ed.): Signaling Requirements for Point to Multipoint Traffic Engineered MPLS LSPs, IETF Internet Draft, June 2005, http://www.ietf.org/internet-drafts/draft-ietf-mpls-p2mp-sig-requirement-03.txt
3. Choi, J.K., Guha, D.: Path Computation Element Metric Protocol (PCEMP), IETF Internet Draft, July 2005, http://www.ietf.org/internet-drafts/draft-choi-pce-metric-protocol-02.txt
4. Choi, J.K., Guha, D., Jo, S.K.: Considerations of point-to-multipoint route optimization using PCEMP, IETF Internet Draft, July 2005, http://www.ietf.org/internet-drafts/draft-choi-pce-p2mp-framework-01.txt
5. Choi, J.K., Guha, D., Jo, S.K., Cuong, D.H., Yang, O.S.: Fast end-to-end restoration mechanism with SRLG using centralized control, IETF Internet Draft, July 2005, http://www.ietf.org/internet-drafts/draft-choi-pce-e2e-centralized-restoration-srlg-03.txt
6. Akin, Y., Townsend, J.K.: Efficient Simulation of TCP/IP Networks characterized by non-rare events using DPR based splitting. IEEE Computer Society 2001, vol. 3, pp. 1734-1740
7. Nakayama, M.K., Shahabuddin, P.: Quick simulation methods for estimating the unreliability of regenerative models of large, highly reliable systems. ACM source probability in the Engineering and Informational Series archive, vol. 18, issue 3, July 2004, pp. 339-368
8. Barbarossa, S., Scaglione, A., Giannakis, G.: Performance analysis of a deterministic channel estimator for block transmission systems with null guard intervals. IEEE Transactions on Signal Processing, vol. 50, number 3, March 2002, pp. 684-695
9. Guha, D.: An interesting reconfigurable optical signal processor architecture. Proceedings of SPIE, Vol. 5246, pp. 656-659, August 2003
10. Guha, D., Choi, J.K., Jo, S.K., Cuong, D.H., Yang, O.S.: Deployment considerations of Layer 1 VPNs using PCEMP. IEEE INFOCOM 2005 Poster/Demo Session, Miami, FL, U.S.A., March 13-17, 2005, http://dawn.cs.umbc.edu/INFOCOM2005/guha-abs.pdf