Progress Report

DOE Award number: DE-SC0004909
Name of recipient: University of Massachusetts, Dartmouth
Project Title: Coordinated Multi-layer Multi-domain Optical Network (COMMON) for Large-Scale Science Applications
Principal investigator: Vinod Vokkarane
Date of Report: December 1st, 2011
Period covered by the report: September 1st - November 30th, 2011

Introduction

We intend to implement a Coordinated Multi-layer Multi-domain Optical Network (COMMON) Framework for Large-scale Science Applications. In the COMMON project, specific problems to be addressed include 1) anycast/multicast/manycast request provisioning , 2) multi-layer multi-domain quality of service (QoS), and 3) multi-layer multi-domain path survivability. In what follows, we outline the progress in this quarter for the above categories.

Activities

In this quarter, our research team at University of Massachusetts, Dartmouth held several conference calls with the researchers/software developers of the ESnet team, regarding the anycast design and development issues in OSCARS. Specifically, our team proposed different schemes for the deployment of the anycast communication paradigm and consulted with the software developers at LBNL, as to which would be the more scalable and efficient design to implement, keeping in mind the further services intended to be provided.

Progress/Accomplishments

In this section we describe the progress and accomplishments in each of the tasks (labeled T1, T2) as outlined in the project proposal:

- T1: Anycast/Multicast/Manycast Request Provisioning

  We introduce algorithms for provisioning anycast, multicast, and manycast calls for both, the immediate and advance (IR/AR) reservation systems [1]. With the advent of bandwidth intensive applications, the demand for multicasting/manycast networking capabilities has become an essential component of wavelength division multiplexed (WDM) optical networks. To support these functionalities in an optical network that is Split-Incapable (Multicast Incapable (MI)), i.e., the optical cross connects are incapable of switching an incoming optical signal to more than one output interface, one must implement a logical overlay to the underlying optical layer.

  We begin first by outlining our research papers which have been accepted (were under review in the last quarter) and then describe our current research focus and outline some tasks which we propose to perform for the next quarter.

  For the static case of provisioning the multicast requests, we had earlier presented integer linear programs (ILPs) to solve these problems with a goal of minimizing the total number of wavelengths required to service the request set. We further developed two lower bounds on the minimum number of wavelengths required to provision the request set. Note that the lower bound is not the *actual* minimum number of wavelengths required, but just a *theoretical* bound. On comparing the lower bounds with the ILPs it was observed that the ILP was within (7-10%) of this bound. We also evaluated the performance of our heuristic algorithms for a dynamic traffic scenario and considered real-world large scale networks. A combination of this work [2, 3] has been submitted to a journal [4]. Work related to the manycast communication paradigm [5, 6], along with the appropriate bounds is in preparation to be submitted to a journal.

  We have also investigated the problem of provisioning holding-time-aware (HTA) dynamic circuits in all-optical wavelength division multiplexed (WDM) networks. We employ a technique

called lightpath switching (LPS) wherein the data transmission may begin on one lightpath and switch to a different lightpath at a later time. Lightpath switches are transparent to the user and are managed by the network. We show that LPS can significantly reduce blocking compared to traditional RWA. We then address the problem of routing dynamic anycast HTA dynamic circuits. We propose two heuristics to solve the anycast RWA problem: anycast with continuous segment (ACS) and anycast with lightpath switching (ALPS). This work received an invitation from the *Elsevier Journal on Optical Switching and Networking* and was accepted for publication in the 2011 issue [7].

In this quarter, we extended our previous work in this area, by evaluating the performance for a set of static advance reservations (unicast). It is our plan to extend this work to also encapsulate the multicast and manycast communication paradigms and eventually evaluate the traffic grooming problem associated with provisioning of IR/AR requests. In the context of HTA AR requests, we plan to extend our framework to *non-continuous* wherein the bandwidth allocated to a particular request (for a certain time period) may potentially be zero. The approach of *non-continuous* transmissions is expected to utilize the underlying network resources efficiently.

- T2: Anycast Development in OSCARS

  We have developed a new multi-domain path computation element (PCE) implementation for the OSCARS 0.6 framework that takes advantage of the anycast paradigm. The new anycast PCE modules will allow researchers to execute future destination-agnostic applications over ESnet, thus broadening the number of available services, and improving the network resource utilization globally. Furthermore, we extend the OSCARS framework to not only perform intra-domain path anycast computation, but also extend such computation across different network domains managed by different instances of OSCARS for making inter-domain anycast path computation possible. Overall, our development demonstrates the feasibility of our proposed implementation and evaluates the improvement of the anycast routing over unicast.

  Design

  The PCE is responsible for computing a single path given the existing network topology, and a connection request. In OSCARS 0.6, this service is provided as a framework which allows third-party PCE implementations to be developed and deployed alongside the rest of OSCARS' modules. The four main modules involved in the path computation request flow are the user interface (or IDC API), coordinator, topology bridge, and PCE modules (shaded modules in Fig. 1).

  Like the rest of the OSCARS framework, PCEs are modules and each one is represented within the OSCARS Coordinator by a PCE Proxy that handles the communication between the Coordinator and the PCE. Requests to PCEs are assumed to be asynchronous. The PCE framework provides:

  - **Modularity:** each PCE is executed as an independent process.
  - **Distribution:** PCEs can be deployed on different (virtual or physical) hosts other than the OSCARS IDC host.
  - **Security:** PCEs follow the OSCARS 0.6 security model in regard to authentication, authorization, and accounting.

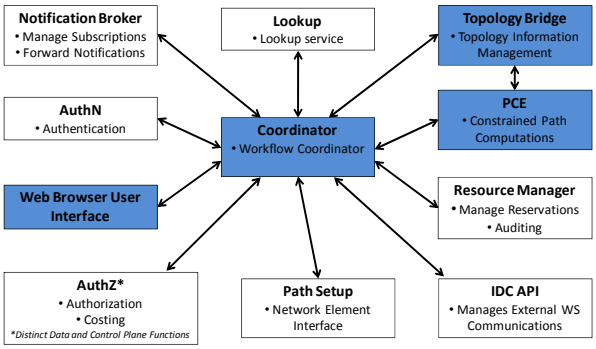

Figure 1: OSCARS modular framework.

– **Language neutrality:** while the default binding is JAVA, the APIs are based on web-services, thus allowing for independent developers to use any language as long as they comply with the API specification.

OSCARS 0.6 allows several PCEs to be deployed, each one of them responsible for computing a specific subset of local paths in a given domain. The execution process is defined as a flexible PCE workflow module, whereby purpose-specific component PCEs are connected in a workflow graph to incrementally prune network resources that do not meet the constraints of the user or network operator. As such, the output from one module can then be fed as input to the next. Specifically, our proposed anycast PCE processes a network topology (domains + nodes + ports + links) as input and outputs a single path from the source to a selected destination.

Implementation

Following the unicast model, our proposed anycast PCE is composed of four core modules which take as an anycast request and a network topology as input, and output an updated, pruned topology (refer to Fig. 2):

– **AnycastConnectivityPCE**: This PCE module is responsible for computing the network topology corresponding to the network connectivity graph between the source node and all the candidate destination nodes of the anycast group. The output of this module is an updated topology with node-pairs not physically connected by a physical fiber pruned out. This module is responsible for dynamically interpreting the network domain so that all other PCEs do not improperly assume additional connectivity.

– **AnycastBandwidthPCE**: This PCE removes the links, ports, and nodes that do not guarantee the bandwidth capacity of the user's anycast request. Fibers which are over-subscribed at the starting time of the request will be pruned from the topology. The behavior of this PCE is largely responsible for the existence of resource-driven connection blocking. Our testing results, show how the probability that requests will be blocked is reduced as an effect of utilizing anycast communication.

– **AnycastVlanPCE**: Each port on a node has a designated number of VLAN tags which represents the maximum number of virtual circuits which may be accommodated at that node. The *AnycastVlanPCE* module prunes out the links, ports and nodes that do not

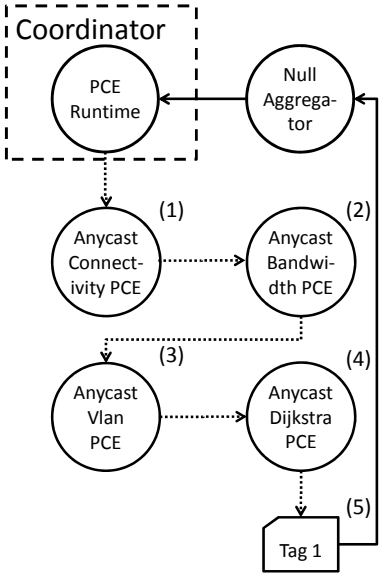

Figure 2: Anycast PCE stack flow-chart.

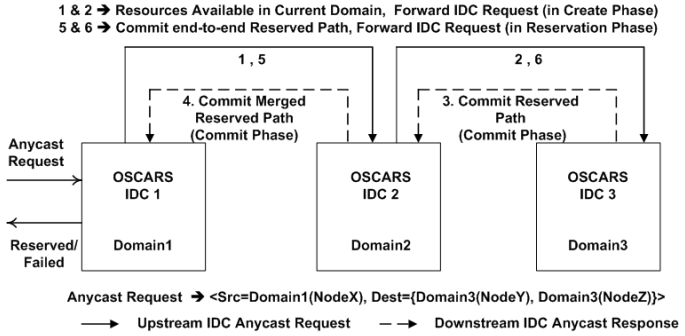

Figure 3: Multi-domain anycast in OSCARS.

have enough VLAN tags to support the virtual circuit., thereby guaranteeing secure connection establishment for all successfully provisioned requests.

– **AnycastDijkstraPCE**: This PCE module computes the potential end-to-end paths to each destination in the anycast set and then selects the final destination based upon some criteria. In this work, we select destinations to satisfy an anycast request such that the candidate along the shortest path is preferred. Alternative metrics can easily be incorporated with our existing *AnycastDijkstraPCE* design to select destinations based on a path's available bandwidth, and/or other metrics.

The worst-case runtime complexity of our anycast PCE implementation is increased over its unicast counterpart by a factor of $|D_s|$, the number of destinations in the anycast set.

OSCARS is responsible for providing the understanding of inter-module relationships and the ordering of the module executions. A *PCERuntime* agent controls this ordering through a customizable XML configuration file that prescribes rules for arranging the PCE module

executions. PCE modules need not be aware of the relative execution ordering. The *NullAggregator* module aggregates a set of paths based on the result from several PCEs. In our case, the *NullAggregator* captures the result, *Tag 1* (refer (5) in Fig. 2), from the last PCE to execute, *AnycastDijkstraPCE*. The final reply is sent back to the *PCERuntime* module, which governs the request forwarding between PCE modules. The final output from the execution of the anycast PCE workflow is a pruned topology consisting exclusively of the VC along the path from the source to the selected anycast destination.

Multi-Domain

The multi-domain workflow for an anycast AR request is shown in Fig. 3. For the sake of simplicity, consider an IDC as a single OSCARS instance. A multi-domain anycast request is first submitted to the local IDC (source IDC is IDC 1 in this case). In this example workflow, the anycast request specifies Node X in Domain 1 which is found locally in the network managed by IDC 1. The request also specifies Node Y and Node Z as part of the anycast destination set, which is remote to IDC 1. Now the Coordinator in IDC 1 initializes the PCE workflow. The *anycastConnectivityPCE* loads all the partially (ingress and egress only) or fully visible (sister network domains share entire topology) topologies as a topology stack to reach from the source to all the anycast destination domains. In this example Domain 2 and Domain 3 are loaded. The *anycastBandwidthPCE* and *anycastVlanPCE* then prune all the local nodes, ports, links in the topology stack which do not fit the user's constraints of bandwidth and VLAN. In case of MPLS, they simply prune the ingress and the egress nodes of the local domain. This pruned topology stack is then fed into the *anycastDijkstraPCE*, which finds the best local path to the egress node for all valid anycast destinations of the local domain and returns this path to the local Coordinator. Now the local Coordinator within IDC 1 determines that the request is inter-domain, flags the anycast request to be in the CREATE phase and forwards this request by loading the profile of the next inter-domain hop in the path which helps to communicate suitably over the inter-network with the next IDC responsible for the inter-domain hop. In Fig. 3, IDC 1 forwards the inter-domain anycast request to IDC 2. Now, IDC 2 performs actions similar to IDC 1 (the OSCARS coordinator and PCE framework are highly re-entrant and efficient by switching logic based on the phase a request is in). If a local path is found feasible, IDC 2 then forwards this request to IDC 3 which manages the destination domain, Domain 3. Now, IDC 3 performs actions similar to IDC 2 in computing the best path to all of the anycast destinations and returns whichever has the shortest number of hops back to the Coordinator. Upon successful receipt of the path, the Coordinator for IDC 3 then locally saves the path in its local database as reserved and changes the anycast request phase to COMMIT and forwards the request back to the sender of the request, IDC 2. IDC 2 sees the phase of the request to be COMMIT, and so it merges the local path with the global path and saves this merged path as the reserved path in the local database. IDC 2 again forwards the updated request to the original sender, IDC 1, which after merging the local and global paths, sets the full end-to-end path in its local database. Subsequently, IDC 1 changes the request status to RESERVED to indicate the end-to-end path is stitched and forwards this end to end reserved path to IDC 2. IDC 2 now overwrites the entire end-to-end anycast path again into its local database and forwards it to IDC 3 which performs similar action of persisting the end-to-end reserved path to the local database. IDC 3 flags the reservation as completed by setting the anycast request status to the RESERVATION-COMPLETE phase, which is then recursively transmitted back to IDC 1

(the original sender). The user is notified that the anycast AR request has been successfully reserved. If the request cannot be provisioned locally in the CREATE phase by any of the domains in the path, then the user is notified that the reservation has failed and the request is blocked.

Results

We stress tested our anycast extension to OSCARS on ESnet and GEANT networks independently for single domain performance of anycast as well as both the domains in case of testing for multi-domain. During the performance analysis, we observed that anycast performs significantly better (about 40% better) in terms of blocking on an average when compared to unicast. This performance betterment is backed with a significant reduction in average hop count (about 55% better when compared to unicast) required to provision the network demand which in turn reduces the number of lightpaths to be setup in the network significantly. In essence by extending OSCARS for anycast we were able to observe significant betterment in performance. We recently submitted a conference paper to IEEE ONDM encompassing the design, development and performance for both single and multi-domain results. We are looking at adding a load balanced anycast as a direct future enhancement to basic anycast other than extending OSCARS with other overlays like multicast, manycast in the next quarter.

- T3: Multi-Layer Multi-Domain Quality of Service (QoS)

  In this task we examine the multi-layer quality of service in optical WDM networks. The aim is to deliver a QoS framework to map input connection requests to a certain number of classes, wherein each of these classes gets a different treatment in the network.

  Work Performed:

  Over the past couple of months, the main tasks performed are follows: (1) development and analysis of partition-based QoS on optical WDM networks with hybrid immediate and advance reservation (IR/AR), and (2) network-wide approximate blocking analysis for hybrid IR/AR.

  As far as the first task is concerned, we followed with the design and implementation on the simulator of the partition-based QoS approach. We evaluated three network scenarios from the mixed IR/AR we investigated in Q2: (a) strict IR/AR partitioning, (b) strict IR/AR partitioning with partial sharing, and (c) flexible IR/AR partitioning with preemption. The second sub-task carried out was the development of a network-wide approximate blocking model for optical WDM networks with hybrid IR/AR. We extended the model for the link blocking analysis [8] to the whole network computation with two different WDM assumptions, under wavelength-continuity constraint and with wavelength conversion. The model is able to calculate the approximate blocking probability given a network offered load and a set of IR/AR classes. One of the contributions of the model with respect to past approaches in the literature is the addition of a flexible method to compute the blocking probability for different IR and AR traffic classes. The analytical model makes use of the Erlang fixed-point approximation to compute the network-wide blocking. As for the second sub-task, we compared the results from the analytical model with those obtained from simulation, and we can found that the blocking probability is well-approximated using the model. We tested different traffic load scenarios with different IR/AR classes and on two different network topologies, NSFNET and ESnet, showing the results that for a diverse number of wavelengths,

the model can compute a good approximation of the blocking probability. This work is also expected to be submitted to a journal [9].

We also extended our anycast algorithm to provide survivability by allowing a link-disjoint backup path to be provisioned to an alternate destination in the anycast set; this technique allows for resilient light paths for destination-agnostic applications. We compared this to the naive approach of provisioning a backup link-disjoint light path only to the primary destination in the anycast set. This approach can consume a large amount of resources because the disjoint backup path is often very long compared to the primary path. In some scenarios no disjoint path may exist to the primary destination regardless of resource availability. Our results indicate that relocation is able to significantly reduce the blocking probability compared to this naive approach because by allowing the backup path to route to an alternate destination, the probability is increased that a shorter path can be found. This reduces the overall load on the network and allows for provisioning of additional requests. Furthermore, relocation offers additional resiliency to destination failure compared to the naive approach.

NOTE: All the above work which has resulted in conference proceedings have acknowledged the DOE-COMMON project.

Cost Status & Unexpected Funds

See attached document.

Next Quarter Deliverables

- Incorporate Traffic Grooming in tasks T1 and T3
- Extend the current OSCARS framework to provision point to multipoint (manycast/multicast) connection establishments.
- Extend our anycast PCE design to allow for survivability via path protection.
- Investigate issues on multi-domain QoS Provisioning.

References

[1] N. Charbonneau, C. Guok, I. Monga, and V. M. Vokkarane, "Advance Reservation Frameworks in Hybrid IP-WDM Networks," May 2011.

[2] A. Gadkar, J. Plante, and V. M. Vokkarane, "Static multicast overlay in WDM unicast networks for large-scale scientific applications," in *Proc. ICCCN*, July. 2011.

[3] A. Gadkar and J. Plante, "Dynamic Multicasting in WDM Optical Unicast Networks for Bandwidth-Intensive Applications," in *Proc. of GLOBECOM 2011*, Houston, TX, USA, Dec. 2011.

[4] A. Gadkar, J. Plante, and V. M. Vokkarane, "Multicast overlay for high-bandwidth applications over optical WDM networks."

[5] J. Plante, A. Gadkar, and V. M. Vokkarane, "Dynamic manycasting in optical split-incapable wdm networks for supporting high-bandwidth applications," in *Proc. IEEE International Conference on Computing, Networking and Communications*, July. 2012.

[6] A. Gadkar, J. Plante, and V. M. Vokkarane, "Manycasting: Energy-Efficient Multicasting in WDM Optical Unicast Networks," in *Proc. of GLOBECOM 2011*, Houston, TX, USA, Dec. 2011.

[7] N. Charbonneau, A. Gadkar, B. Ramaprasad, and V. Vokkarane, "Dynamic circuit provisioning in all-optical WDM networks using lightpath switching," *Journal of Optical Switching and Networking*, 2011.

[8] J. Triay, C. Cervelló-Pastor, and V. M. Vokkarane, "Analytical Model for Hybrid Immediate and Advance Reservation in Optical WDM Networks," in *Proc. of GLOBECOM 2011*, Houston, TX, USA, Dec. 2011.

[9] ——, "Computing approximate blocking probabilities for hybrid immediate and advance reservation in optical WDM networks," *under review, IEEE Journal*, 2011.