Enhancing ESnet’s Unicast-Only OSCARS with a Manycast Overlay Service

Jeremy M. Plante
Vinod M. Vokkarane

Dept. of Electrical and Computer Engineering
UMass Lowell
Jeremy_Plante@student.uml.edu
Use-Cases

- **Large Hadron Collider – CMS/ATLAS**
  - 2000 physicists, 170 institutions, 36 countries.
  - 300 trillion collisions per year.
  - 25 petabytes per year.
  - One copy of data maintained at CERN.
  - Second copy collectively contained by multiple “Tier 1” institutions.
  - Tier 1s send to tier 2s for science analysis.

- **Climate Modeling – Earth System Grid Federation**
  - Replicated data storage.
  - Distributed remote access.

- **Nuclear Fusion – ITER**
  - Distributed workflows
  - Coordinating and co-scheduling of compute nodes/storage/instruments.
  - Fast recalibration and experimentation.
  - Compute nodes high interdependency between tasks and tight deadline restrictions.
Use-Cases

- What do all these applications have in common?
  - Need to provide fast communication from a single sender to multiple receivers (hopefully in parallel).

- Need to send ALL data to ALL storage sites?
  - Costly
  - Complex
  - Inefficient
  - Unnecessary
Communication Paradigms

UNICAST

UNICAST

MULTICAST

ANYCAST
Generalization

- Manycast request:
  - N nodes in network.
  - K candidate destinations.
  - K’ required destinations.

<table>
<thead>
<tr>
<th>K</th>
<th>K’</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Unicast</td>
</tr>
<tr>
<td>1 ≤ K ≤ N</td>
<td>1</td>
<td>Anycast</td>
</tr>
<tr>
<td>1 ≤ K ≤ N</td>
<td>K’ = K</td>
<td>Multicast</td>
</tr>
<tr>
<td>1 ≤ K ≤ N</td>
<td>K’ ≤ K</td>
<td>Manycast</td>
</tr>
<tr>
<td>K = N</td>
<td>K’ = K</td>
<td>Broadcast</td>
</tr>
</tbody>
</table>
Generalization

- Multicast = NP-Complete.
- Manycast = NP-Hard.
How to Select Destinations

- Nearest candidate destinations.
  - Shortest paths.
  - Steiner tree.

- Least-expensive destinations.
  - Processing power.
  - Latency.

- Load-balancing.
  - Choose destinations in least-demand.
  - Choose destinations with fewest prior commitments.

- Overall Cost (non-monetary)
  - Power-efficiency.
  - Energy-efficiency.
  - GHG emissions.

- Arbitrarily
  - All destinations assumed equal
  - Reach at least $K'$ out of $K$ destinations.
On-demand Secure Circuits and Advance Reservation System

- Developed by DOE’s Scientific Networking Group (ESnet).
- ESnet has built and maintains the world’s fastest scientific communication network (100 Gbps).
- Provides guaranteed performance on dedicated virtual circuits (VCs) for transmitting data.
- Most popular circuit-provisioning software amongst networking/research communities.
- 50% of ESnet’s monthly 14 petabytes of transmitted data is carried on OSCARS VCs.
What Does OSCARS Look Like?

OSCARS Inter-Domain Controller (IDC)

- Notification Broker
  - Manage Subscriptions
  - Forward Notifications

- AuthN
  - Authentication

- Web Browser User Interface

- IDC API
  - Manages External WS Communications

- Lookup
  - Lookup service

- Coordinator
  - Workflow Coordinator

- Topology Bridge
  - Topology Information Management

- PCE
  - Constrained Path Computations

- Path Setup
  - Network Element Interface

- AuthZ*
  - Authorization
  - Costing
  - Distinct Data and Control Plane Functions

- Resource Manager
  - Manage Reservations
  - Auditing

Users

User Apps

Other IDCs

Local Network Resources
OSCARS/Client Interaction

Client → OSCARS → Client

RequestType:
- CreateReservation

ResponseType:
- CreateReservation Response
  - Global Reservation ID (GRI)
  - Status
  - Start Time
  - End Time
  - Bandwidth
  - Circuit Source
  - Circuit Destination
  - VLAN
OSCARS

- Currently only supports point-to-point communication: Manycast/Multicast inherently not possible at the optical layer.

- Not only is OSCARS incapable of point-to-multipoint communication, but up until very recently, ESnet was limited to unicast by its hardware infrastructure.

- Provide front-end logic for grouping individual OSCARS VCs, such that their identities are transparent to the end-user.
  - This is an overlay approach to logical Manycasting!
Overlay Models

Single-hop model

- Establishes $K'$ separate and unique end-to-end VCs from source to destinations.
- ALL lightpaths originate at the source.
- Manycast = Collection of Unicast.

**Manycast Request: (1, {2, 5, 6}, 2)**

Routing in Physical Topology

Routing in Logical Topology
Overlay Models

Multiple-hop model

- Establishes a logical Steiner tree, possibly consisting of multiple logical hops from the source, reaching at least $K'$ destinations.
- VCs may originate/terminate at source OR destinations OR other network nodes, this requires temporary storage for conversion from optical signal to electronic.

**Manycast Request: (1, \{2, 5, 6\}, 2)**
Overlay Models

Conference Proceedings


Journal Publications

## Overlay Models

<table>
<thead>
<tr>
<th>Single-Hop</th>
<th>Multiple-Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔ Fast set-up.</td>
<td>✔ Bandwidth-efficient.</td>
</tr>
<tr>
<td>✔ No delay at “drop-nodes”.</td>
<td>✗ Complex.</td>
</tr>
<tr>
<td>✔ No network-internal storage.</td>
<td>✗ Requires modifications to OSCARS.</td>
</tr>
<tr>
<td>✔ Simple.</td>
<td>✗ Some delay at “drop-nodes”.</td>
</tr>
<tr>
<td>✗ Bandwidth hungry.</td>
<td>✗ Requires network-internal storage.</td>
</tr>
</tbody>
</table>

- **Requirements from ESnet:**
  - No modifications to OSCARS code.
  - No internal storage within the network.
  - Parallel Transfers to multiple destinations.

- **Must use single-hop approach!**
Manycast Client Design

User Application

Manycast Client

API

OSCARS

PCE

Connectivity PCE

Bandwidth PCE

VLAN PCE

Dijkstra PCE

Coordinator

Workflow Coordinator

PCERuntime

Resource Manager

Reserve paths
Manycast Client Flexibility

- Specify a group of Manycast destinations
- Specify minimum threshold/maximum cutoff.
  - By specifying different values for threshold/cutoff, Manycast service flexibility increases:

<table>
<thead>
<tr>
<th>Candidate Destinations</th>
<th>Threshold</th>
<th>Cutoff</th>
<th>Paradigm Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anycast (3/1)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Best-Effort Manycast (3/2)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>Best-Effort Multicast (3/3)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>Manycast (3/2)</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>Bounded Best-Effort[Manycast, Multicast]</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>Multicast (3/3)</td>
</tr>
</tbody>
</table>

- If threshold/cutoff cannot be satisfied, extra sub-requests are cancelled (first-fit) to satisfy constraints of the Manycast request.
Performance Evaluation

Multicast/Manycast OSCARS comparison:

- ESnet topology.
- Advance Reservations (2-hour window)
- Correlation Factor => probability requests overlap in time.
  - 0 Correlation = time-independent set of requests.
- Request set size $R = 100$.
- Average of 10 unique request sets.
Performance Evaluation

- Manycast flexibility lowers blocking, despite same number of reached destinations.
- Particularly true at higher correlation factors.
- The relative blocking reduction due to destination flexibility is less dramatic as more candidate destinations are added.
Conclusions

- Collaboration and distributed workflows are becoming omnipresent.
- Desired parallelism must be taken into account in the network to prevent bottleneck.
  - Manycast communication.
- Many of these applications already transport data over ESnet, likely using OSCARS virtual circuits.
- Proposed Manycast client makes parallel transfers possible without any modification to OSCARS design.
- Current deployed system is simple, but not ideal.
  - But it’s a tangible step towards a deployable Manycast overlay system!
Enhancing ESnet’s Unicast-Only OSCARS with a Manycast Overlay Service

Manycast client code available
https://www.dropbox.com/sh/0jv518h9ecmz6eq/SrZSy3ug3a

Details on collaboration with ESnet
http://faculty.uml.edu/vinod_vokkarane/common/

OSCARS
http://es.net/services/oscars/

Jeremy M. Plante
Vinod M. Vokkarane

Dept. of Electrical and Computer Engineering
UMass Lowell
Jeremy_Plante@student.uml.edu