Traffic Optimization for ExaScale Science Applications

draft-xiang-alto-exascale-network-optimization-00

Q. Xiang [1] H. May Wang [1] H. Newman[2]

G. Bernstein [3] A. Mughal [2] J. Balcas[2]

[1] Tongji/Yale University [2] California Institute of Technology [3] Grotto Networking

July 21@IETF 96

LHC: Large Hadron Collider

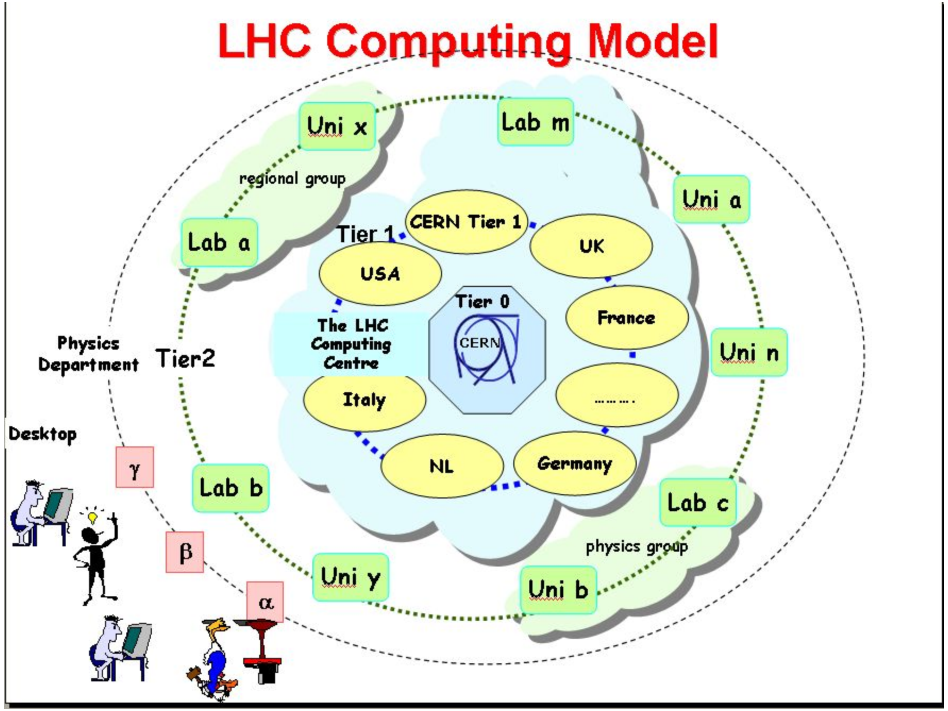


Figure source: cern.ch

Problem Settings

- Exascale Data Transfer
- Requirement for CMS / ALTAS experiment in LHC project
- Features (Multiple Datasets, Domains, Software Infrastructures)

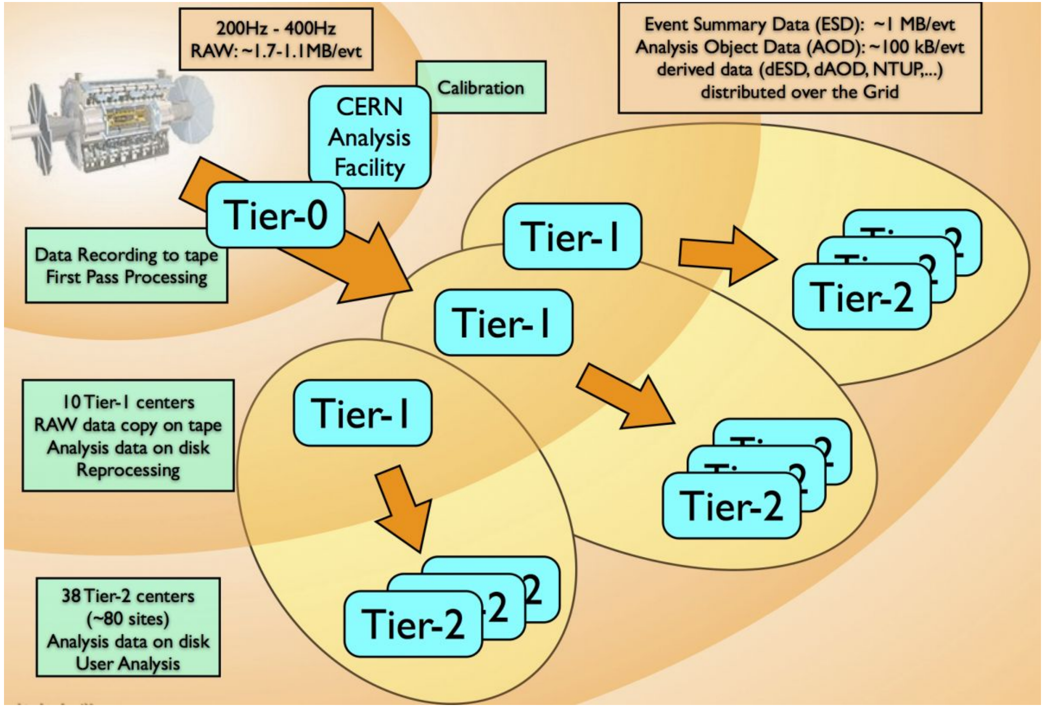


Figure source: cern.ch

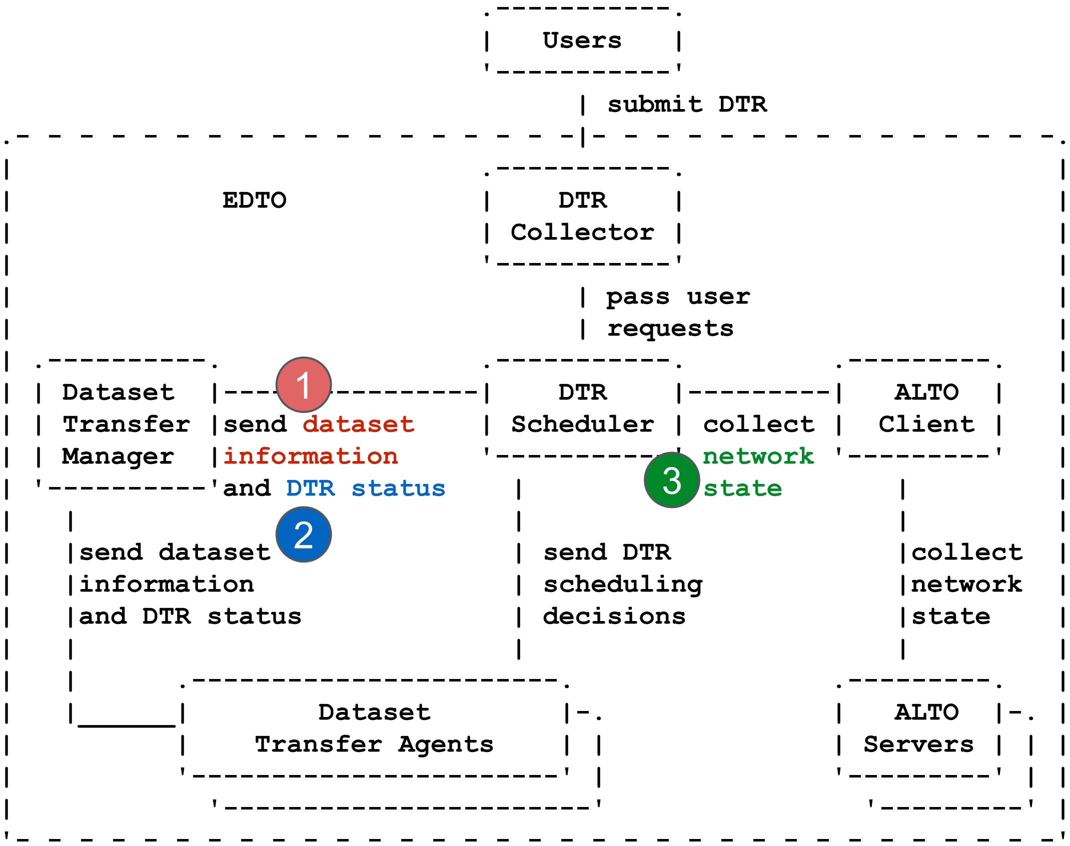
Tier2 Use Case

Identify specific hosts (IP addresses) in a subnet participating in a dataset transfer; direct (only) those flows

Tier1 Use Case

Direct flows to and from specific subnets of data transfer nodes

Exascale Dataset Transfer Orchestrator (EDTO)



```
                              .-----------.
                              |   Users   |
                              '-----------'
                                    | submit DTR
.- - - - - - - - - - - - - - - - - -|- - - - - - - - - - - - - - -.
|                             .-----------.                       |
|             EDTO            |    DTR    |                       |
|                             | Collector |                       |
|                             '-----------'                       |
|                                   | pass user                   |
|                                   | requests                    |
| .----------.                .-----------.         .---------.   |
| | Dataset  |----------------|    DTR    |---------|   ALTO  |   |
| | Transfer |send dataset    | Scheduler | collect |  Client |   |
| | Manager  |information     '-----------' network '---------'   |
| '----------'and DTR status    |           state       |         |
|   |                           |                       |         |
|   |send dataset               | send DTR              |collect  |
|   |information                | scheduling            |network  |
|   |and DTR status             | decisions             |state    |
|   |                           |                       |         |
|   |      .-----------------------.                .---------.   |
|   |______|        Dataset        |-.              |   ALTO  |-. |
|          |    Transfer Agents    | |              | Servers | | |
|          '-----------------------' |              '---------' | |
|            '-----------------------'                '---------' |
'- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -'
```

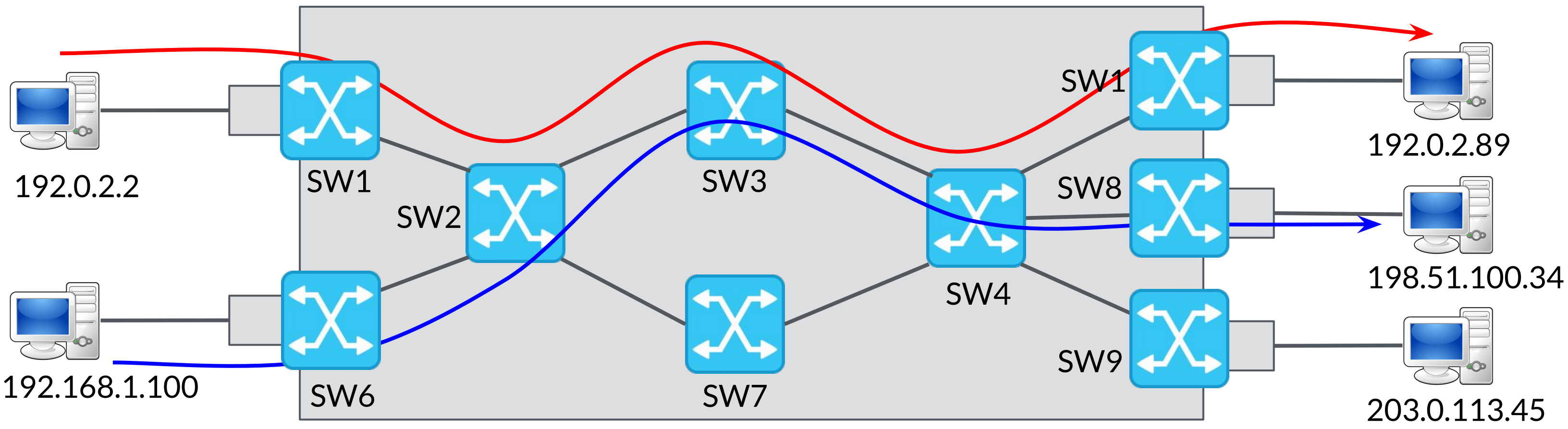
EDTO Architecture

The DTR scheduler requires three classes of information as input:

- Dataset information
- DTR status information
- Network state information (provided by ALTO)

Question: what network information can DTR scheduler get from ALTO?

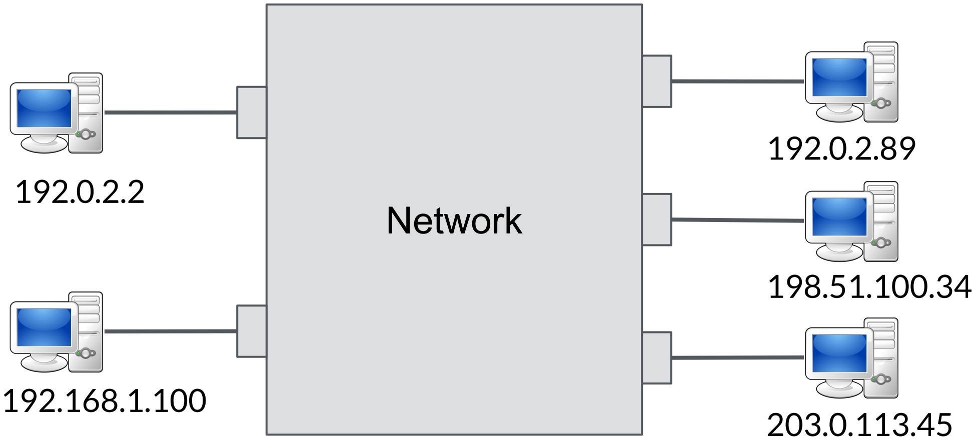
- base ALTO protocol (RFC7285): Cost Map, Endpoint Cost Service, etc.



ALTO: One-Node Topology Abstraction

- ECS/Cost Map services provide one-node abstraction

```
“endpoint-cost-map”: {
  “ipv4:192.0.2.2”: {
    “ipv4:192.0.2.89”: 100
  },
  “ipv4:192.168.1.100”: {
    “ipv4:198.51.100.34”: 100
  }
}
```



DTR Scheduler

Each flow’s rate is set to 100Mbps.

Network Congestion!

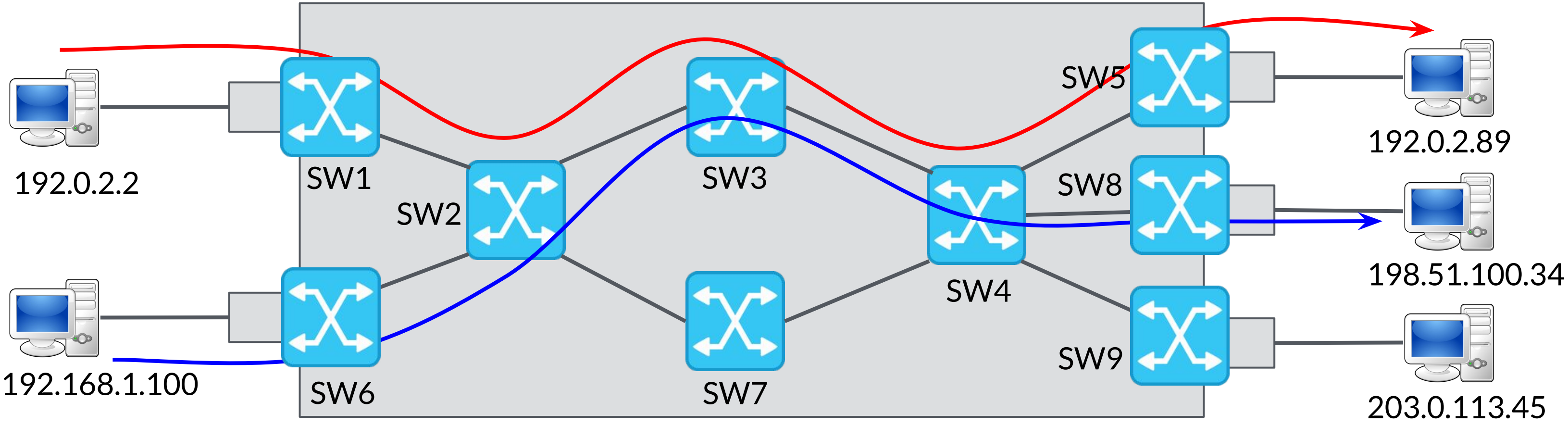
ALTO: One-Node Topology Abstraction

- The efficiency of DTR scheduler depends on the abstraction level of topology
- Returning the raw, complete network state?
  - High overhead
  - Violation of network providers' privacy

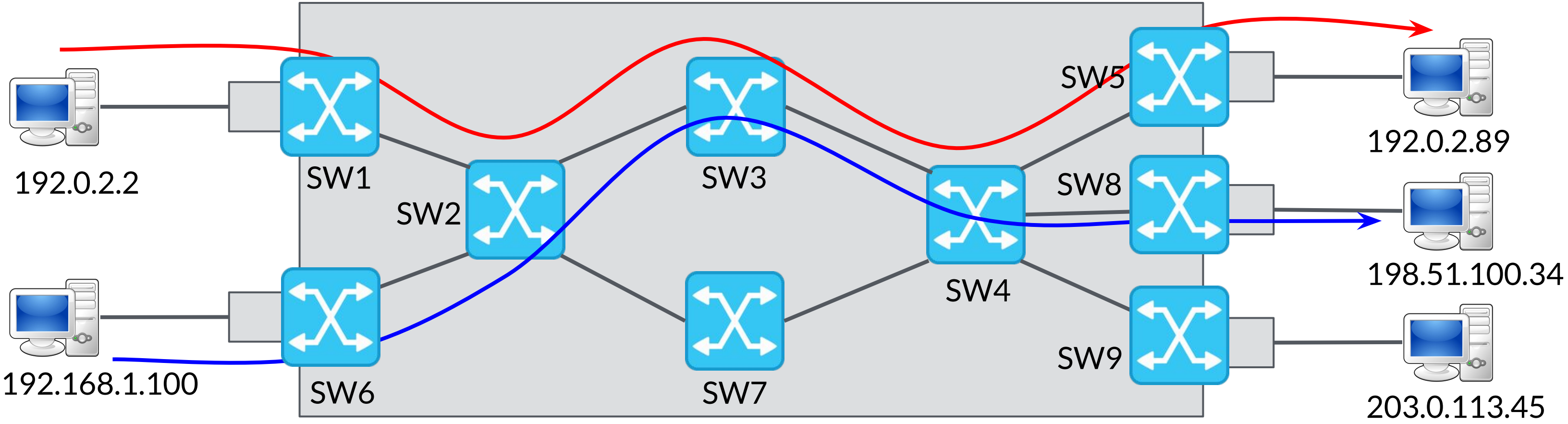
- **Question**: how can DTR scheduler get **sufficient** topology information from ALTO server?

Solution 1: Path Vector Extension Service

- ALTO Server exposes topology information of the computed path
- Client retrieves path vector based on endpoint cost map
- Client converts endpoint cost map to graph based format



Example: Path Vector Response



```
"cost-map" : {
  "ipv4:192.0.2.2": {
          "ipv4:192.0.2.89":    ["ne12",   “ne23”, “ne34”, “ne45”],
  },
  "ipv4:192.168.1.100": {
          "ipv4:198.51.100.34":    ["ne62", “ne23”, “ne34”, "ne48"]
  }
}
```

```
"net-map" : {
        "ne12" : {"bw" : 100},
        "ne23" : {"bw" : 100},
        "ne34" : {"bw" : 100},
        "ne45" : {"bw" : 100},
        "ne62" : {"bw" : 100},
        "ne48" : {"bw" : 100}
}
```

Set rate of 50Mbps for each flow for fairness.

Solution 1: Path Vector Extension Service

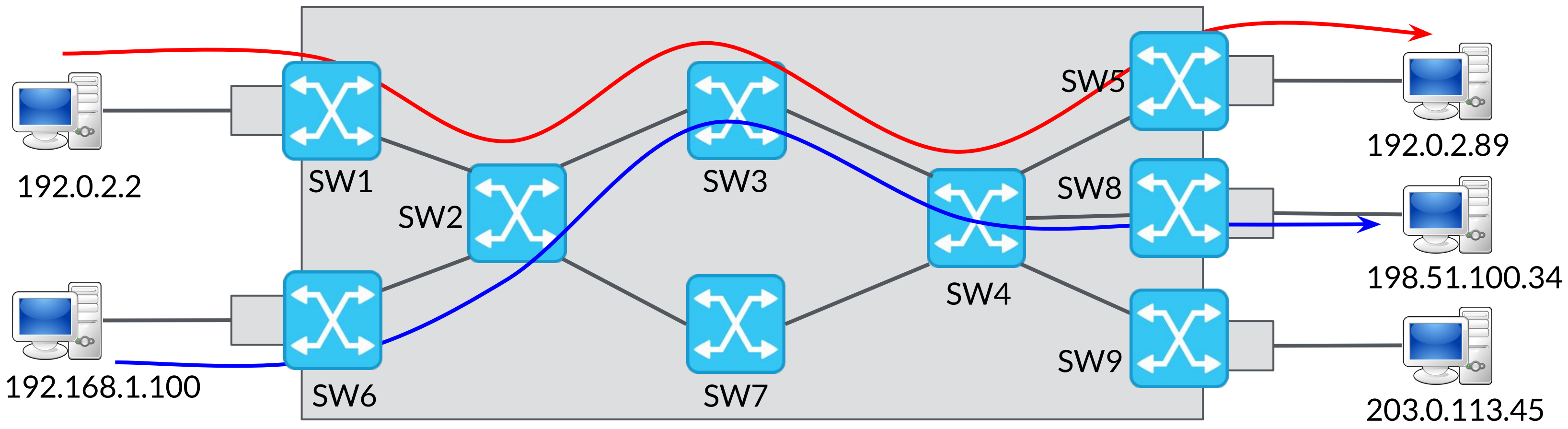
- Advantage
  - DTR scheduler receives sufficient topology information
- Limitations
  - Redundant topology information
  - Amplify the problem scale for DTR scheduler -> slow schedule computation
  - Expose privacy of network provider

- **Question**: how can DTR scheduler get **sufficient** yet **minimal** topology information from ALTO server?

Solution 2: Routing State Abstraction Extension Service

- Path vector: provide topology information of every routing path
- RSA: provide lossless compression of topology information of routing paths by using equivalence conditions
    - **Equivalent**: applications can make the same decision based on RSA response and PV response
    - **Minimal**: only expose topology information that is absolutely necessary to applications

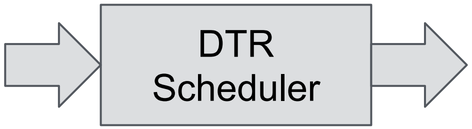
Example: RSA Response



```
"cost-map" : {
  "ipv4:192.0.2.2": {
          "ipv4:192.0.2.89":    [“ane24”],
  },
  "ipv4:192.168.1.100": {
          "ipv4:198.51.100.34":    [“ane24”]
  }
}
```

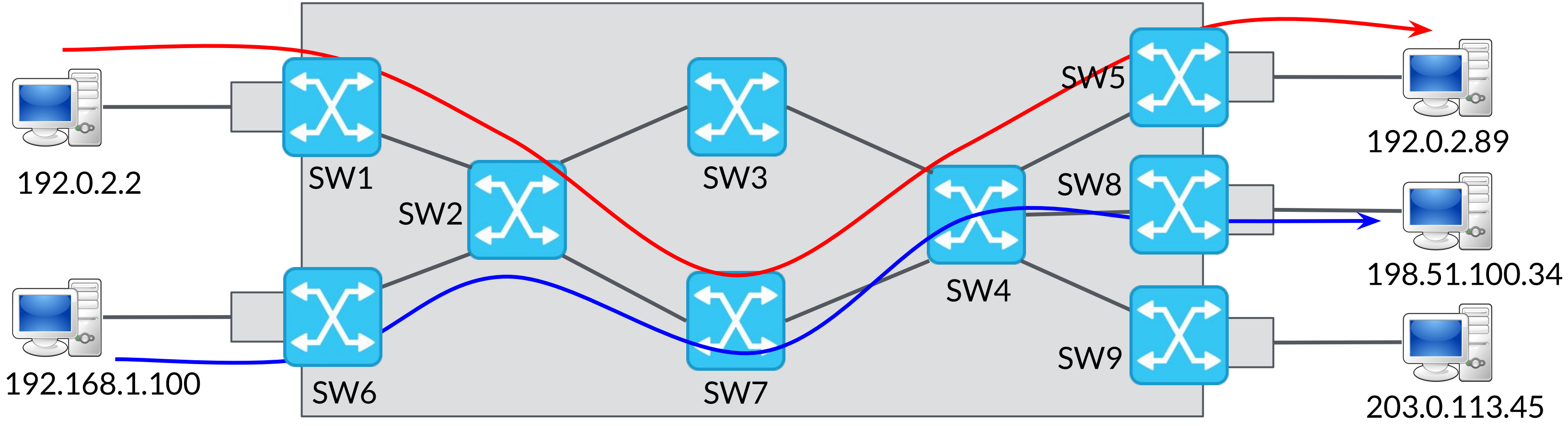
```
"nep-map" : {
        "ane24" : {"bw" : 100},

}
```



Set rate of 50Mbps for each flow.

Example: RSA Response

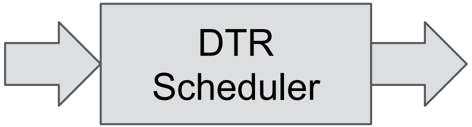


```
"cost-map" : {
  "ipv4:192.0.2.2": {
          "ipv4:192.0.2.89":    [“ane24”],
  },
  "ipv4:192.168.1.100": {
          "ipv4:198.51.100.34":    [“ane24”]
  }
}
```

```
"nep-map" : {
        "ane24" : {"bw" : 100},

}
```

DTR scheduler is oblivious of actual underlying topology



Set rate of 50Mbps for each flow.

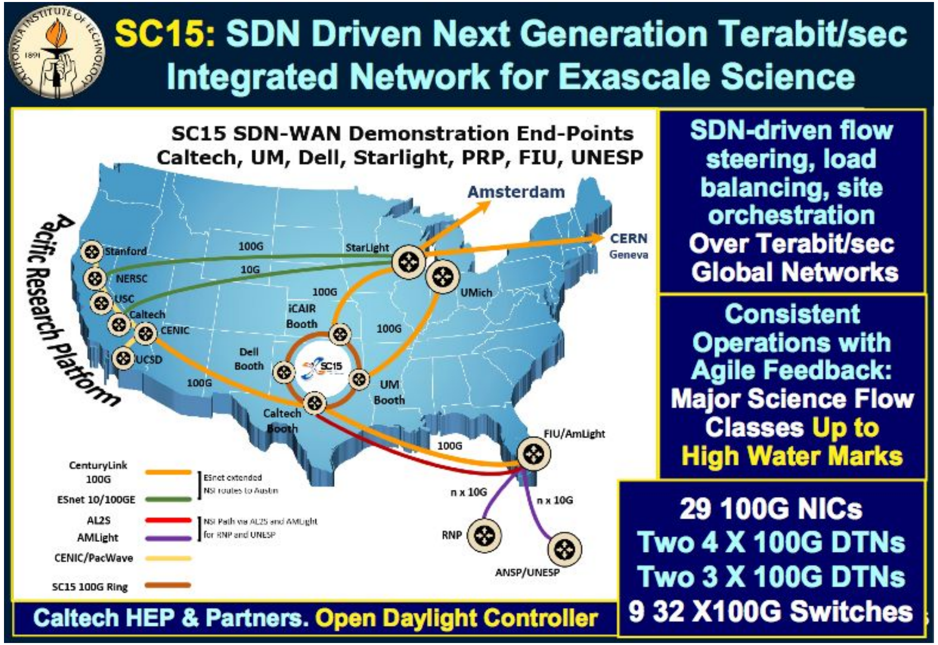
Progress and Next Steps

Current Status:

- An SDN (OpenDaylight) based implementation on a single-domain testbed
- Prototype of RSA to be deployed soon

Future Work:

- Multi-domain deployment on LHC networks
- Topology information aggregation from multiple ALTO servers



SC15: SDN Driven Next Generation Terabit/sec Integrated Network for Exascale Science

SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

Consistent Operations with Agile Feedback: Major Science Flow Classes Up to High Water Marks

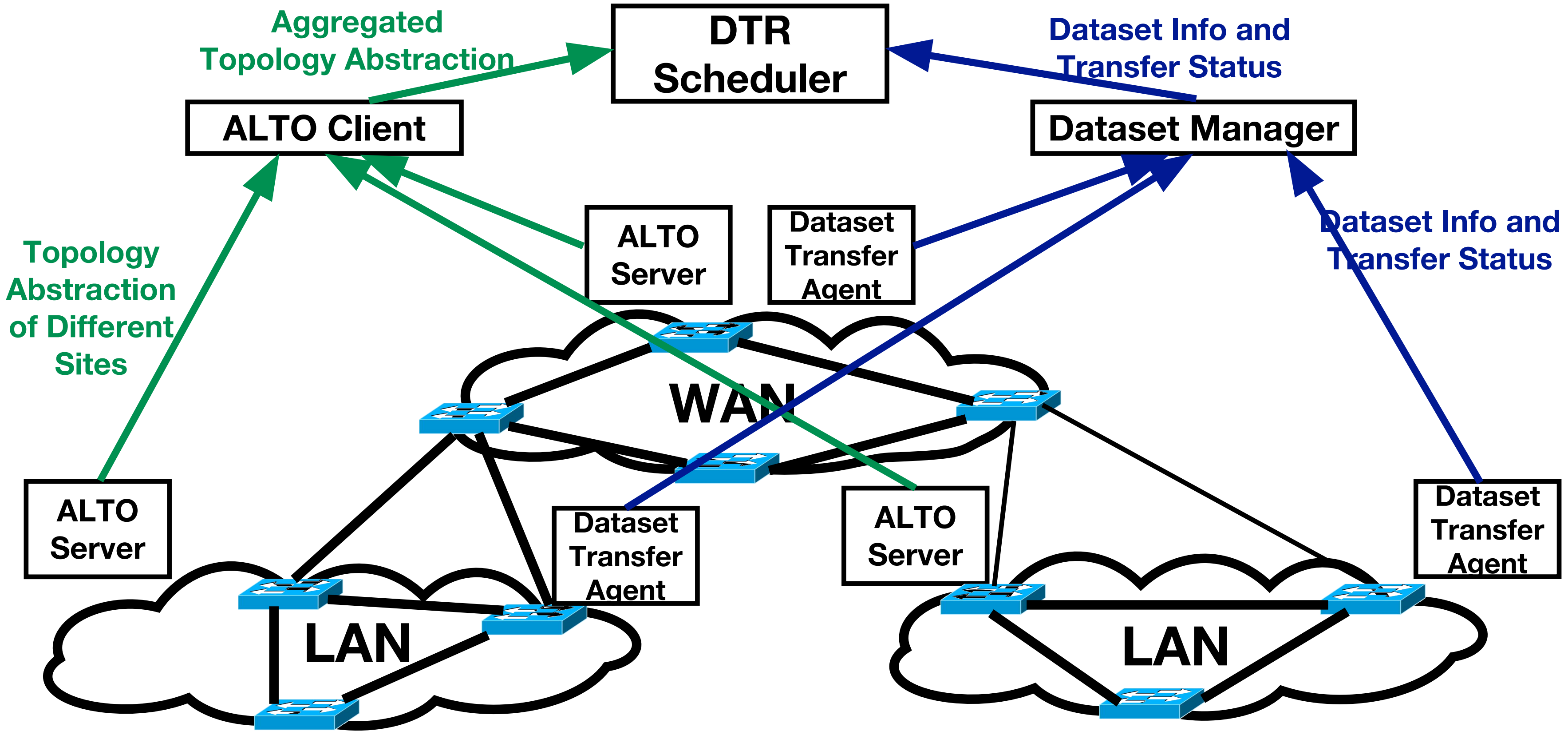
29 100G NICs
Two 4 X 100G DTNs
Two 3 X 100G DTNs
9 32 X100G Switches

Caltech HEP & Partners. Open Daylight Controller

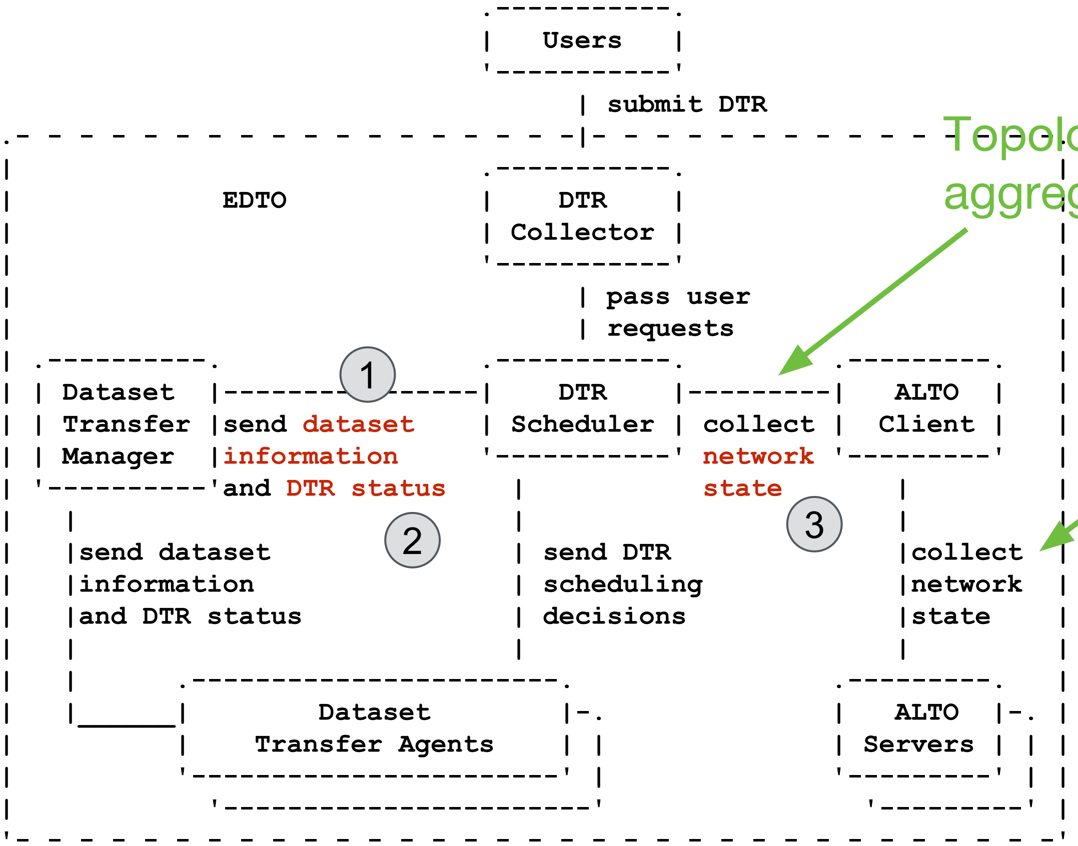
Figure source: http://monalisa.caltech.edu:8080/Slides/Public/SCICReports2016Final/ICFASCICPresentation2016_Final022216.pdf

Backup Slides

System Overview



Solution (Proposal)

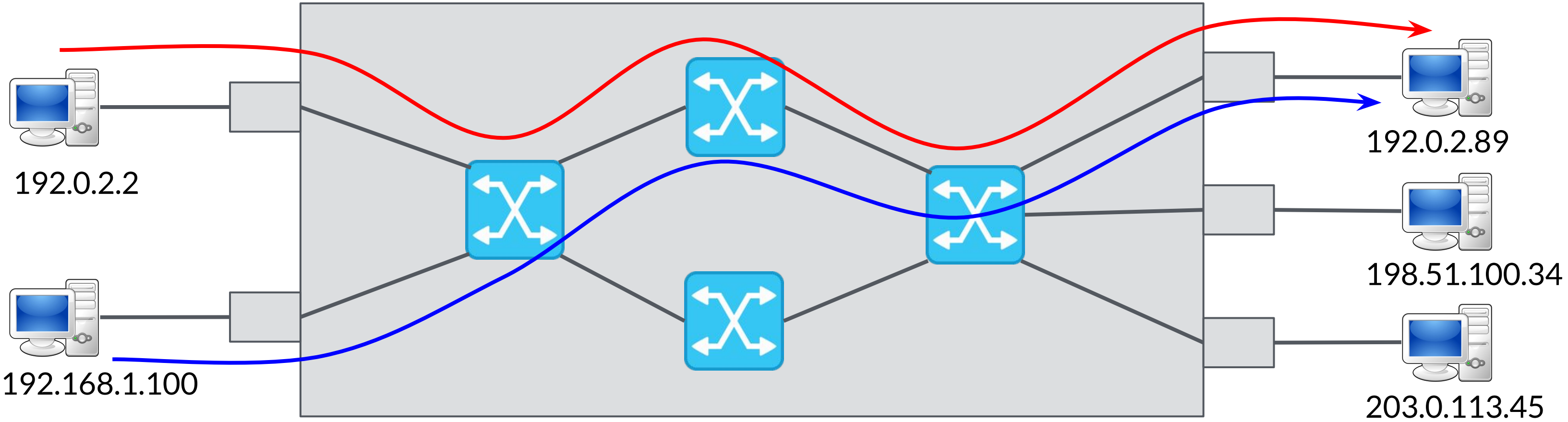


Topology based aggregation

key service: Topology extension

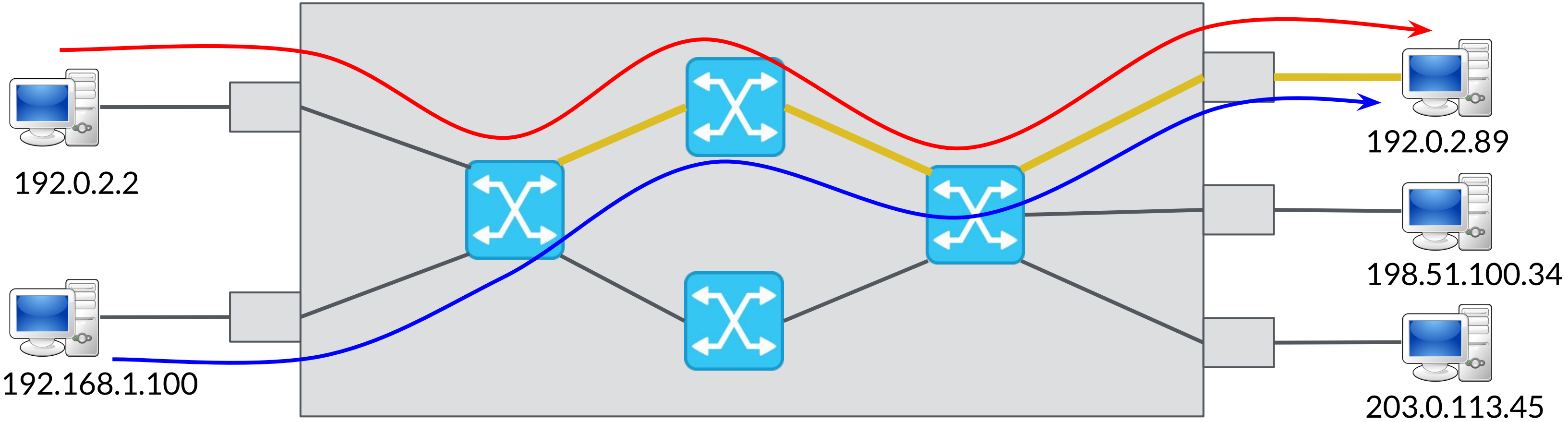
Path Vector based Solution

- Client retrieves path vector based endpoint cost map
- Client converts endpoint cost map to graph based format



RSA based Solution

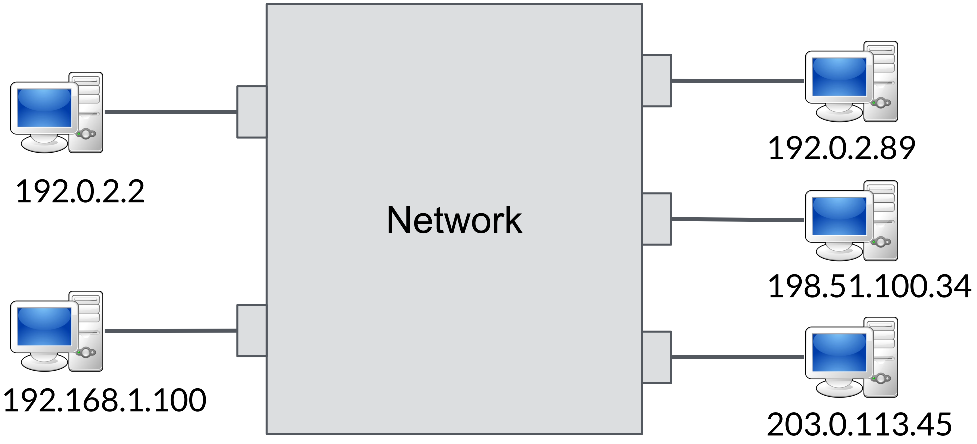
- Path Vector may provide redundant information
- RSA can provide lossless compression of topology information by using equivalence conditions



Limitation of ALTO One-Node Abstraction

- ECS/Cost Map services provide one-node abstraction
- The efficiency of DTR scheduler depends on the abstraction level of topology
- Resort to ALTO topology extension services

```
“endpoint-cost-map”: {
  “ipv4:192.0.2.2”: {
    “ipv4:192.0.2.89”: 1,
    “ipv4:198.51.100.34”: 2,
    “ipv4:203.0.113.45”: 3
  },
  “ipv4:192.168.1.100”: {
    “ipv4:192.0.2.89”: 2,
    “ipv4:198.51.100.34”: 3,
    “ipv4:203.0.113.45”: 1
  }
}
```



Outline:

1-2 slides: background on LHC and CMS

1-2 slides: EDTO framework

1 slide: what does EDTO need from ALTO (Introduce Problem)

1 slide: ECS/Cost Map based example (limitation of base ALO service)

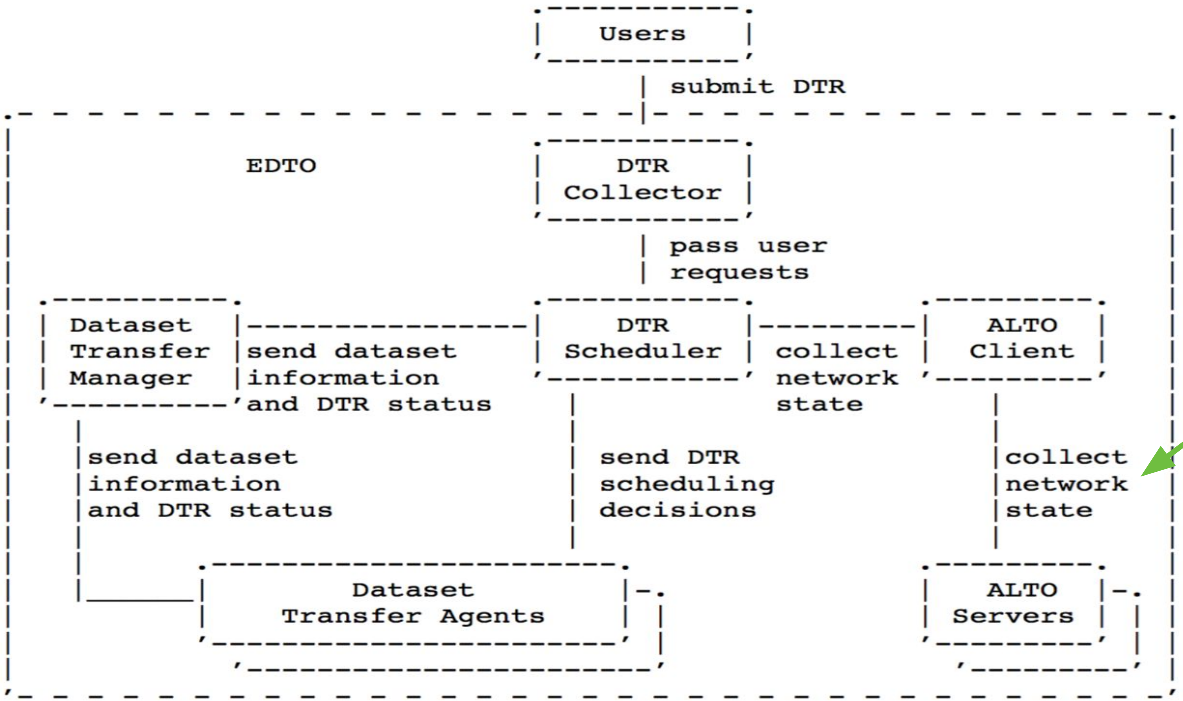
1 slide: PV based solution

1 slide: RSA based solution

1 slide: current status (ODL implementation on a single-domain testbed, prototype of RSA to be deployed soon, etc.)

1 slide: future challenges: multi-domain, server discovery, topology info aggregation from multiple ALTO servers, etc.

System Architecture



```
                              .-----------.
                              |   Users   |
                              '-----------'
                                    | submit DTR
.- - - - - - - - - - - - - - - - - -|- - - - - - - - - - - - - - -.
|                             .-----------.                       |
|             EDTO            |    DTR    |                       |
|                             | Collector |                       |
|                             '-----------'                       |
|                                   | pass user                   |
|                                   | requests                    |
| .----------.                .-----------.         .---------.   |
| | Dataset  |----------------|    DTR    |---------|   ALTO  |   |
| | Transfer |send dataset    | Scheduler | collect |  Client |   |
| | Manager  |information     '-----------' network '---------'   |
| '----------'and DTR status    |           state       |         |
|   |                           |                       |         |
|   |send dataset               | send DTR              |collect  |
|   |information                | scheduling            |network  |
|   |and DTR status             | decisions             |state    |
|   |                           |                       |         |
|   |      .-----------------------.                .---------.   |
|   |______|        Dataset        |-.              |   ALTO  |-. |
|          |    Transfer Agents    | |              | Servers | | |
|          '-----------------------' |              '---------' | |
|            '-----------------------'                '---------' |
'- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -'
```

key service : Topology extension

Solution (Proposal)

- Topology extension
- Server discovery
- Topology based aggregation

```
“endpoint-cost-map”: {
  “ipv4:192.0.2.2”: {
    “ipv4:192.0.2.89”: 100,
    “ipv4:198.51.100.34”: 2,
    “ipv4:203.0.113.45”: 3
  },
  “ipv4:192.168.1.100”: {
    “ipv4:192.0.2.89”: 2,
    “ipv4:198.51.100.34”: 100,
    “ipv4:203.0.113.45”: 1
  }
}
```

Implementing EDTO faces some challenges:

- Limitation of ALTO
  - Lack of sufficient network information for topology properties
- Cross domain issues
  - Consistency of services support
  - Information aggregation

Discussion

- Deployment Issue
- Benefiting From ALTO Extension Topology Services

Limitation of ALTO One-Node Abstraction

- The efficiency of DTE Scheduler depends on the abstraction level of topology