ALTO Use Case: Resource Orchestration for Multi-Domain, Geo-Distributed Data Analytics

draft-xiang-alto-multidomain-analytics-02

Qiao Xiang[1,2], Franck Le[3], Y. Richard Yang[1,2],
Harvey Newman[4], Haizhou Du[1], J. Jensen Zhang[1, 2]

[1] Tongji University, [2] Yale University,
[3] IBM Watson Research Center,
[4] California Institute of Technology

December 11, 2018, IETF 103 ALTO Interim Meeting

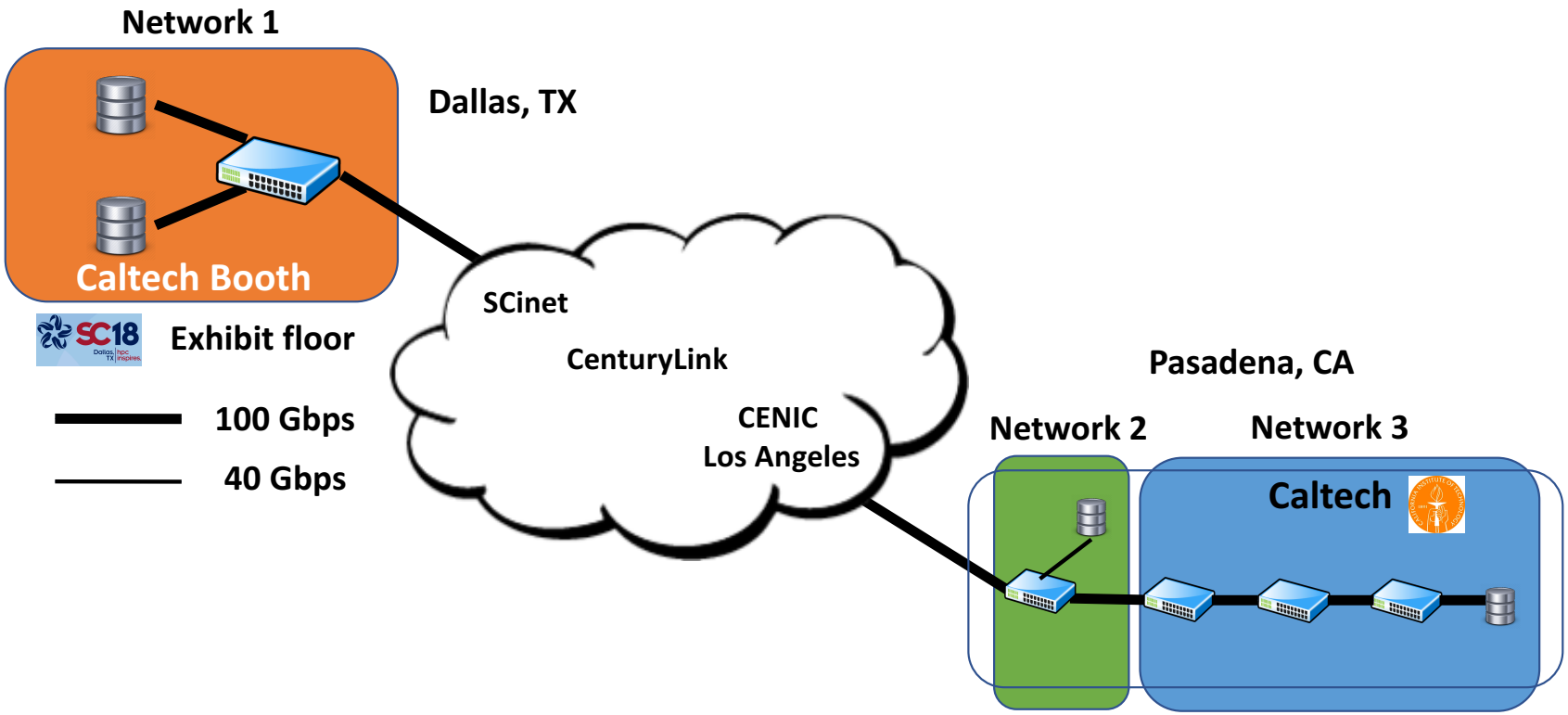
Takeaway from IETF 102

- Two technical updates for the resource abstraction discovery phase (Phase 3).
  - Update the design of the privacy-preserving multi-domain resource abstraction aggregation protocol .
    - The new design does not require a chaining aggregation process between different ASes.
  - Introduce a super-set projection technique to improve the scalability.

Update for IETF 103

- Demonstration of -02 design at SuperComputing'18
- Design update:
  - Separation of resource orchestrator and ALTO client for better privacy preservation of bandwidth feasible region
  - A learning-based orchestrator that automatically interacts with the ALTO client and learns the optimal resource reservations without knowing the bandwidth feasible region

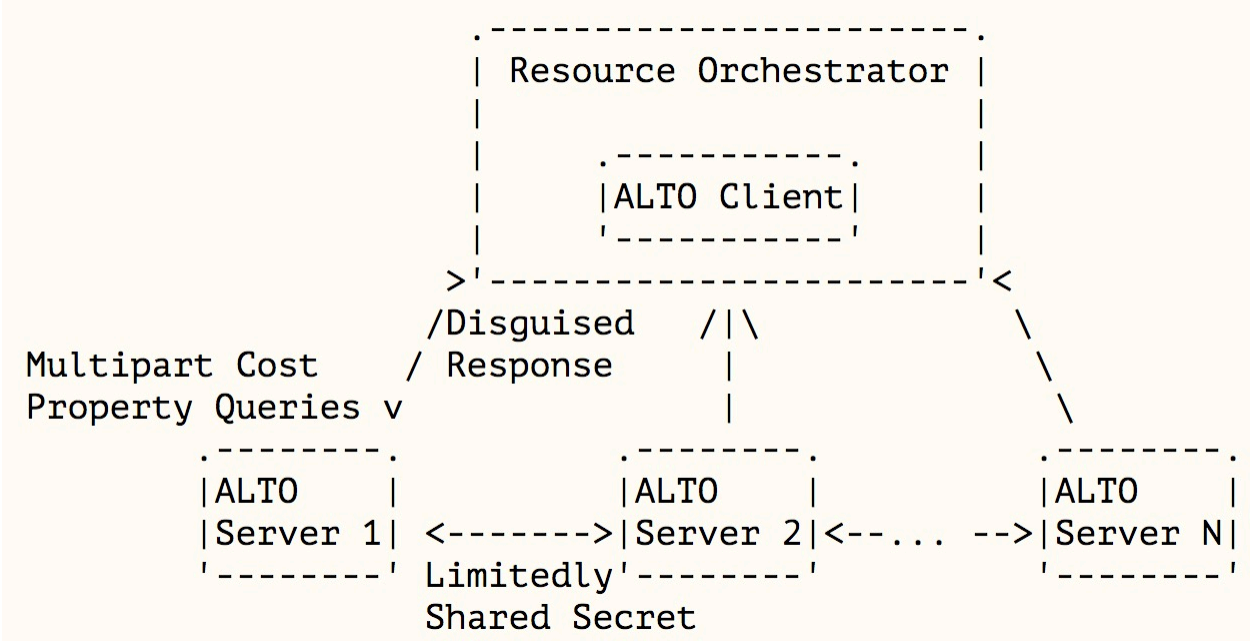
Demonstration at SuperComputing'18



- Full demonstration of Unicorn (now named as Mercator) to orchestrate the transmission of a set of scientific workflows from Dallas to Pasadena at 100 Gbps
- Demo video recordings: https://youtu.be/kUK78gHlQDI

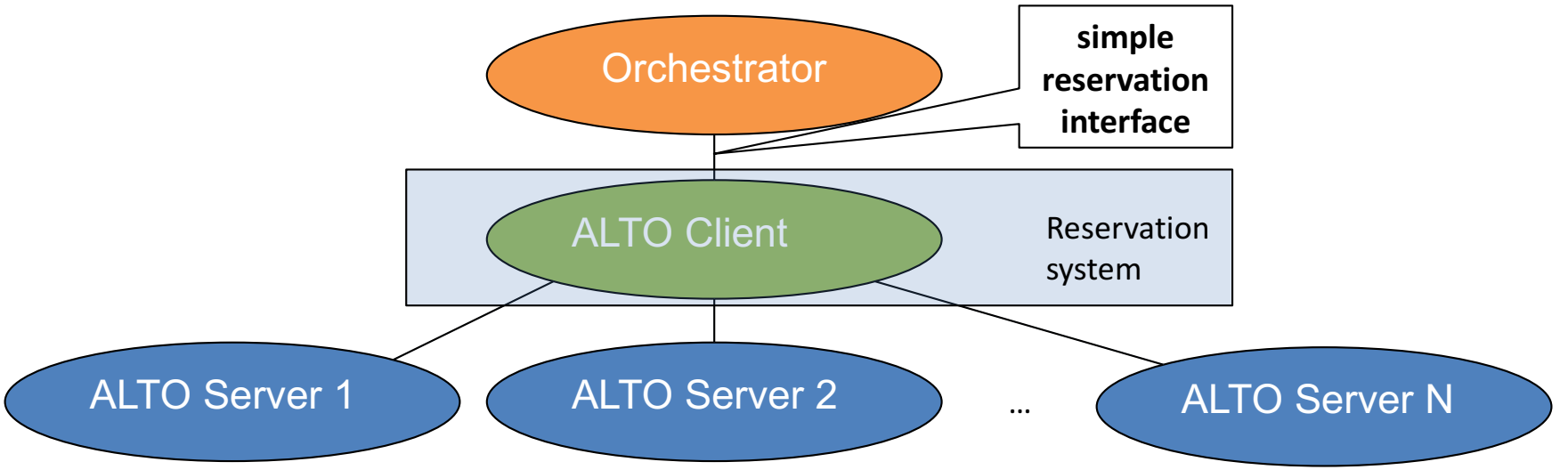
Design Update

- **Previous design in -02**: the resource orchestrator directly receives the resource information collected from the ALTO client.
- **Issue**: ALTO path vector returns the bandwidth feasible region to the application. Such information is still private to networks. The impact of revealing such information to application is still unclear.



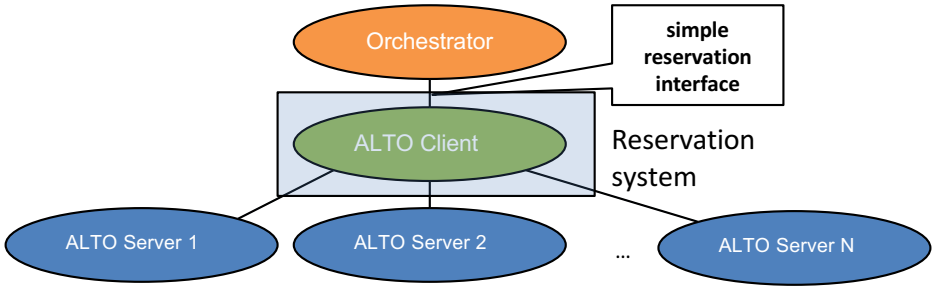
Design Update

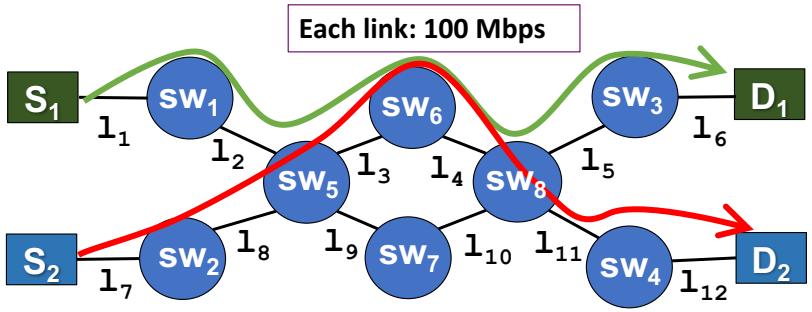
- **New design to be updated in -03**: separation of orchestrator and ALTO client
  - At **Phase 3** (Resource State Abstraction Discovery), The ALTO client does not send the ALTO-PV-encoded resource information (linear inequalities) to the resource orchestrator.
  - In stead, a **simple reservation interface** is provided by the reservation system (e.g., OSCARS) for orchestrator to submit requests for reserving a specific amount of bandwidth, and return either success or failure.



Design Update: Details

- **Goal of orchestrator**: maximize $util(\boldsymbol{x})$
- ALTO client maintains the ALTO PV responses collected from ALTO servers





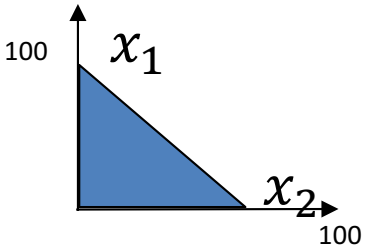
Bandwidth feasible region $\boldsymbol{K}$

$$x_1 \leq 100\ Mbps, \forall l_u \in \{l_1, l_2, l_5, l_6\},$$

$$x_2 \leq 100\ Mbps, \forall l_u \in \{l_7, l_8, l_{11}, l_{12}\},$$

$$x_1 + x_2 \leq 100\ Mbps, \forall l_u \in \{l_3, l_4\}.$$

$$x_1, x_2 \geq 0$$

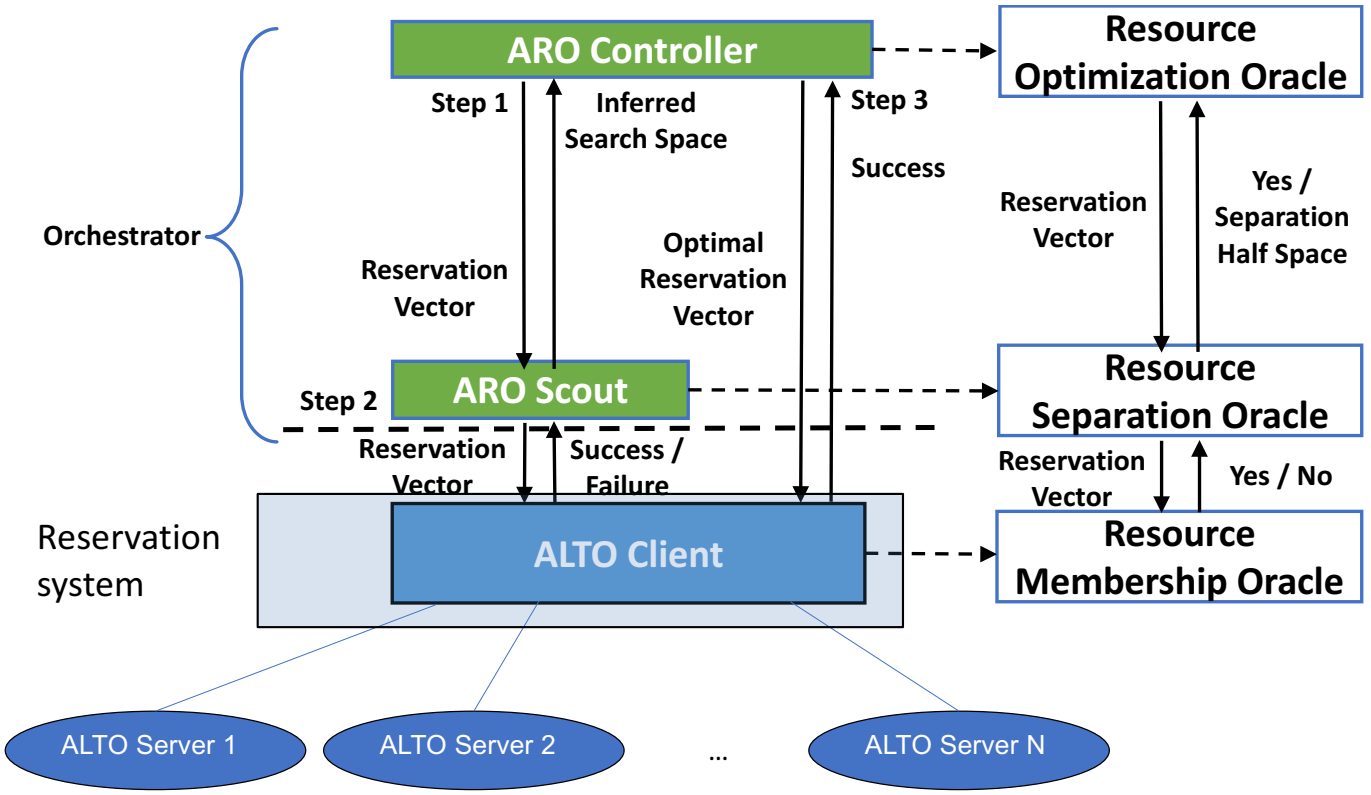


- Model the simple interface provided by ALTO client as a resource membership oracle.

Resource Membership Oracle (ReMEM): Given a reservation vector $\check{\boldsymbol{x}}$, return YES if $\check{\boldsymbol{x}} \in K: \{\boldsymbol{x} | \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$, and return NO otherwise.

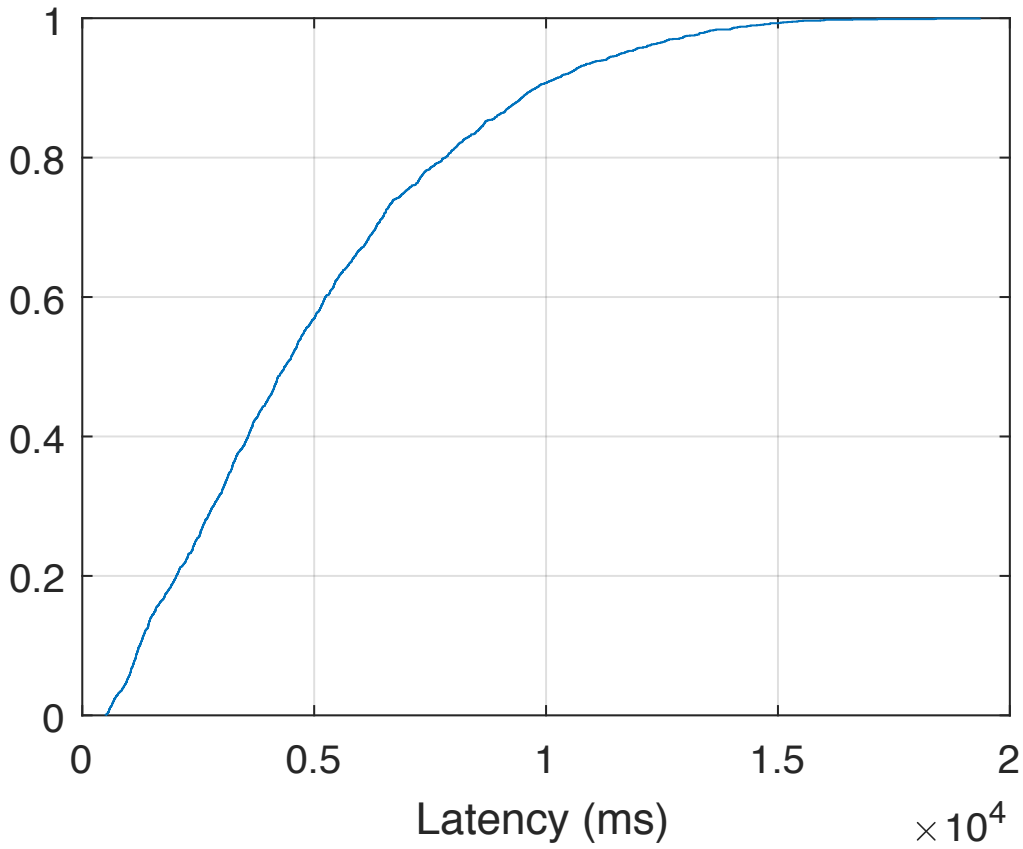
Design Update: Details

- Fast construction of optimization oracle (i.e., optimizing resource reservation) via $O(n^3)$ calls on membership oracle (i.e., calling ALTO client).



Result

- 2215 flow-set requests in a week's CMS trace
- 100% correctness ratio
- For 95% of requests, BoxOpt learns the optimal resource reservation within 12 seconds (assuming the user is in NYC and the network is in LA)



Summary and Next Steps

- Goal: efficient, scalable, privacy-preserving multi-domain resource discovery and orchestration in collaborative science networks
- Previous versions (-01 and -02) focus on efficiency, scalability, and privacy preserving between ALTO servers and ALTO client
- New design further tackles the privacy preserving issue between ALTO client / reservation system and orchestrator
  - This feature will be documented in the next version (-03).

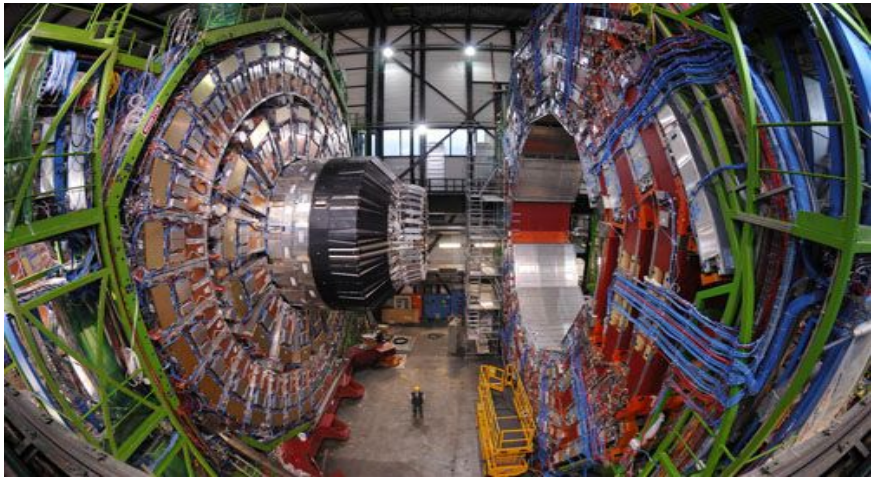
Next step

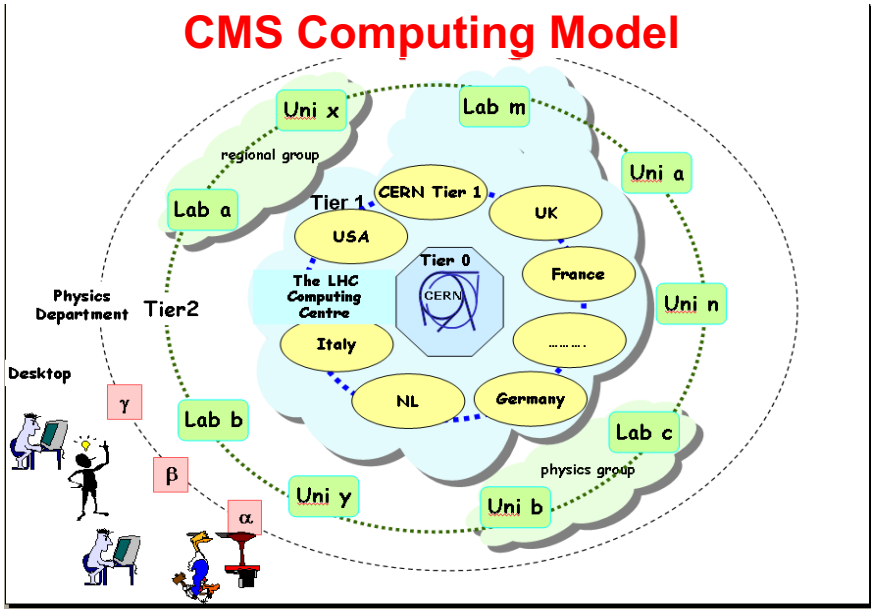
- Full integration with OSCARS, the in-house resource reservation system of CMS

Backup slides

Recap: Multi-Domain, Geo-Distributed Data Analytics

- **Settings**: Different organizations contribute various resources (e.g. , sensing, computation, storage and networking resources) to collaboratively collect, share and analyze extremely large amounts of data.
  - Example: the CMS experiment in Large Hardon Collider.

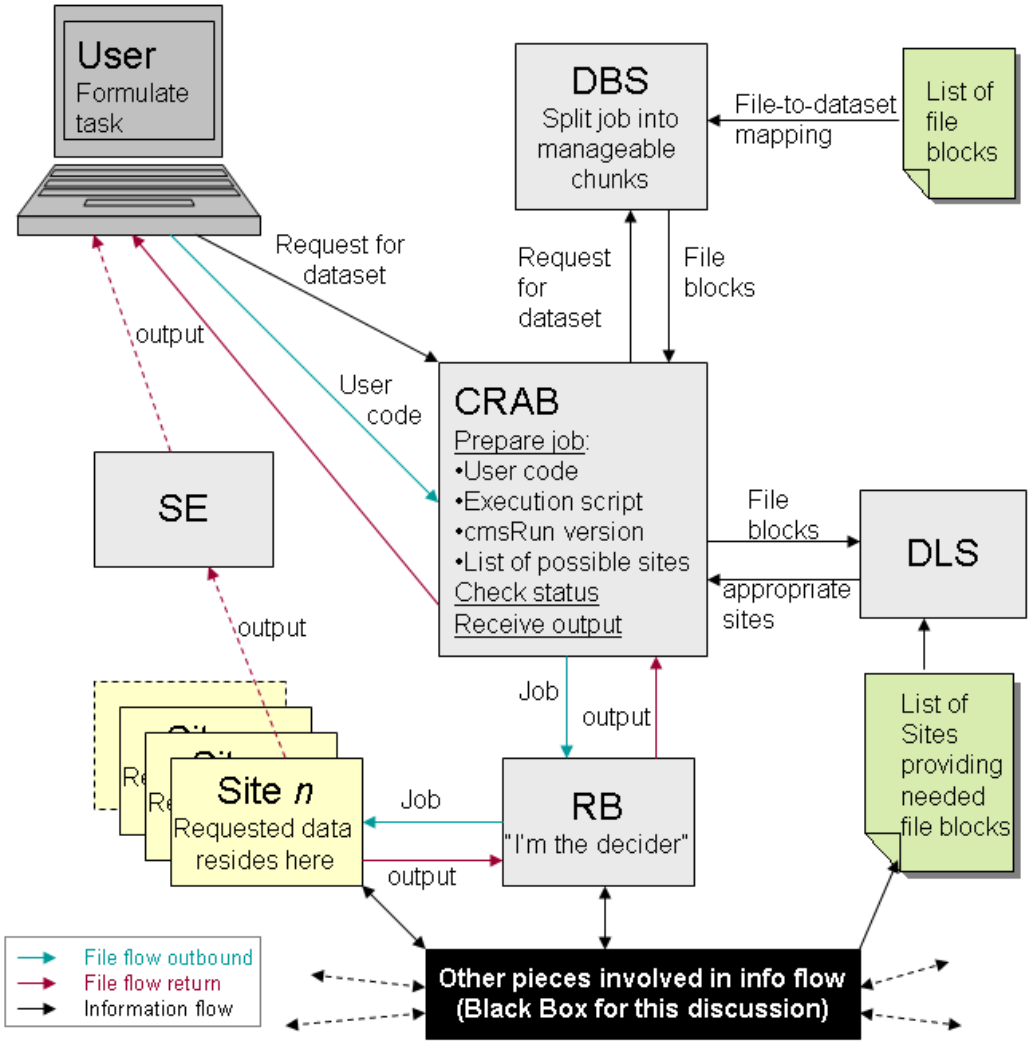




Current CMS Data Analytics Work Flow

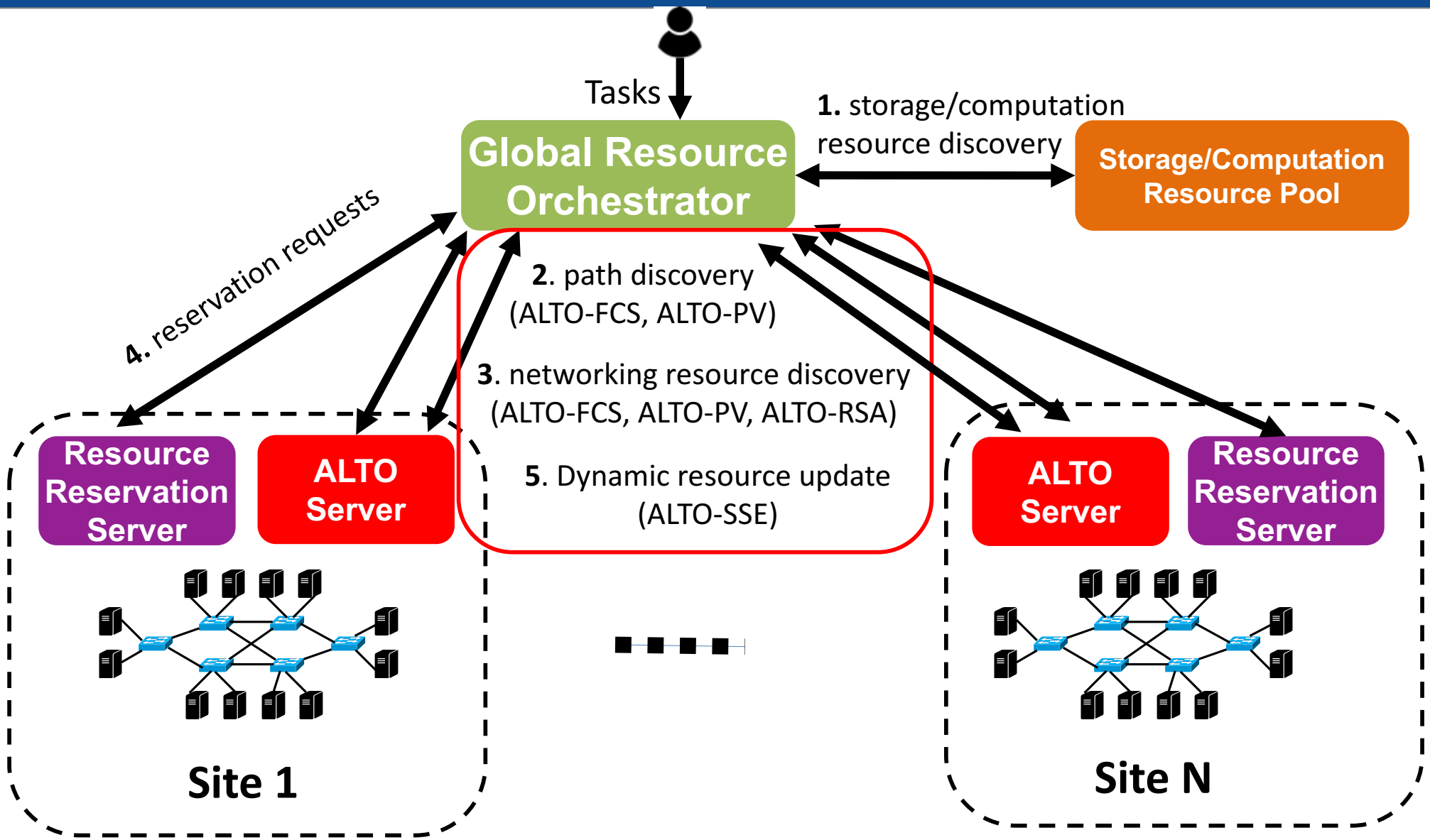
- **Factors determining data analytics task delay**.
  - Task decomposition (parallelization).
  - Data transmission from input dataset location to computation nodes.
  - Data transmission from computation nodes to output dataset sites.

- Current CMS workflow.
  - Simple, manual parallelization.
  - Opportunistic, network-unaware computation node assignment.
  - Opportunistic, network-unaware output stage out.

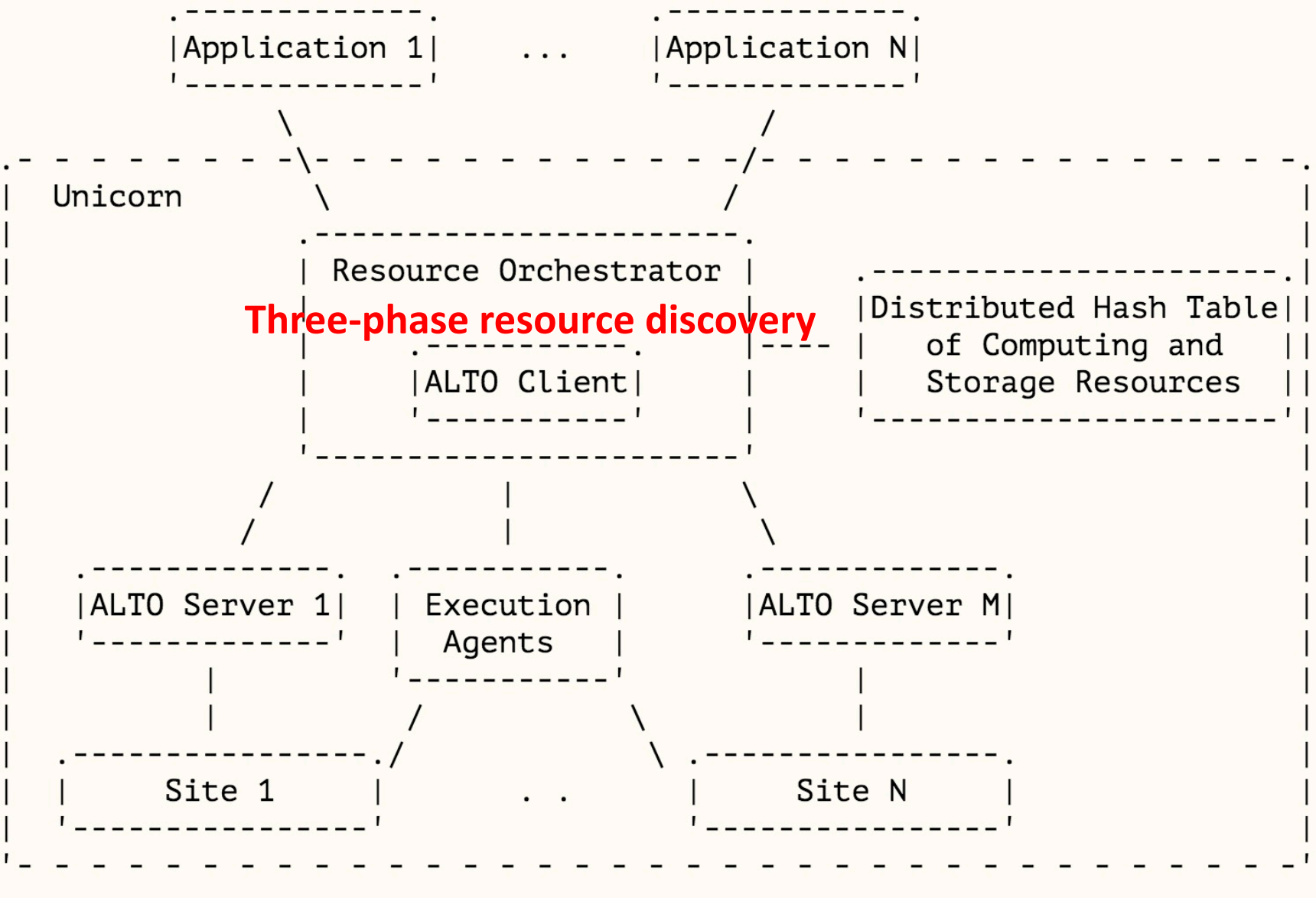


Source: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookAnalysisWorkFlow

Architecture

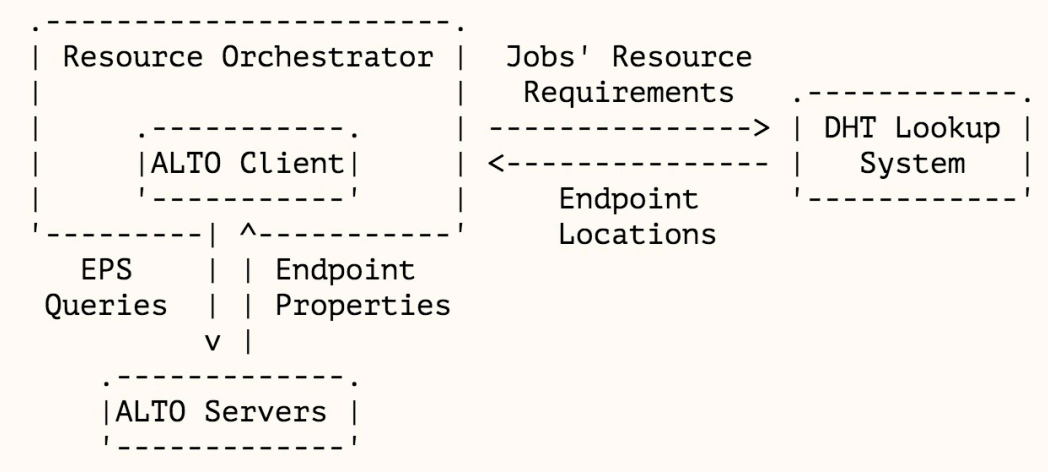


Architecture of Unicorn

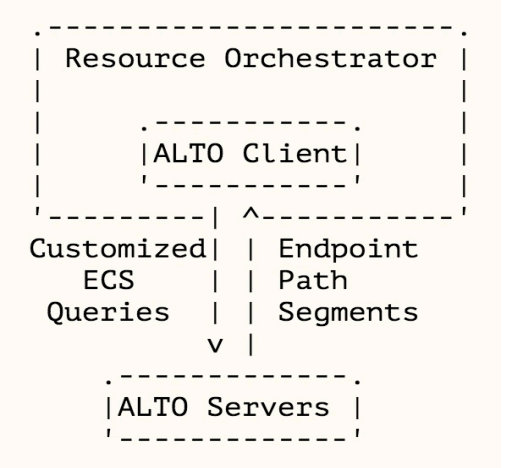


Three-Phase Resource Discovery

- Phase 1: Endpoint Property Discovery
  - Discover the locations and properties of computing and storage resources via ALTO EPS service.

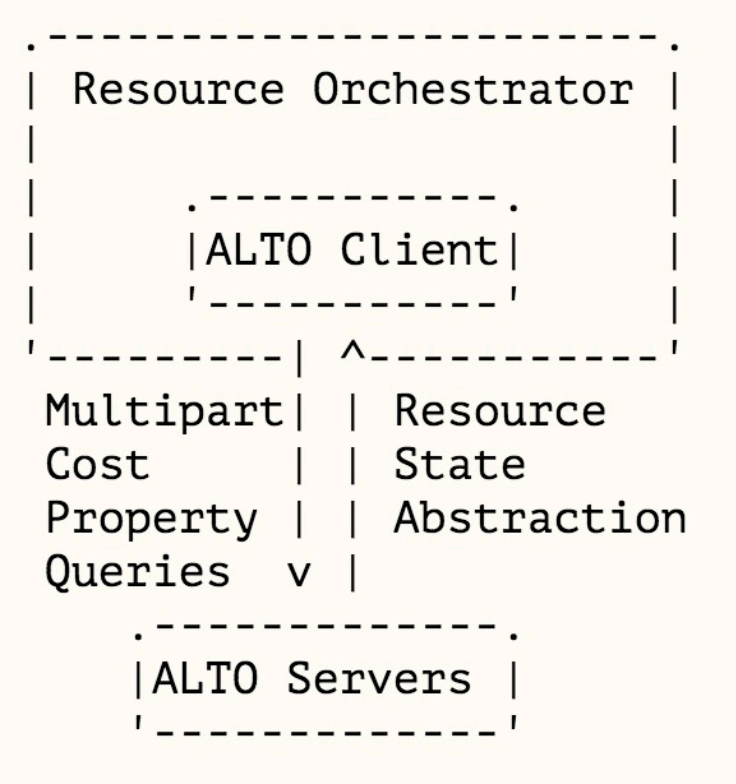


- Phase 2: Endpoint Path Discovery
  - Discover the connectivity between computing and storage resources via network map and ECS service.



Three-Phase Resource Discovery

- Phase 3: Resource State Abstraction Discovery
  - Discover the networking resource sharing between flows via ALTO multipart cost property (MCP) service.
  - **Option 1**: Each ATLO server independently sends the responses to the ALTO client.
    - **Drawback**: expose the private capacity region of each network.



Three-Phase Resource Discovery

- Phase 3: Resource State Abstraction Discovery
  - Discover the networking resource sharing between flows via multipart cost property service.
  - **Option 2**: an ALTO-extension for privacy-preserving interdomain resource information aggregation (see the detailed algorithm in the draft), which returns the **intersected** capacity region of all networks.