

I E T F®

ALTO Integration and Implementation Supporting CERN Data Management (FTS/Rucio Integration)

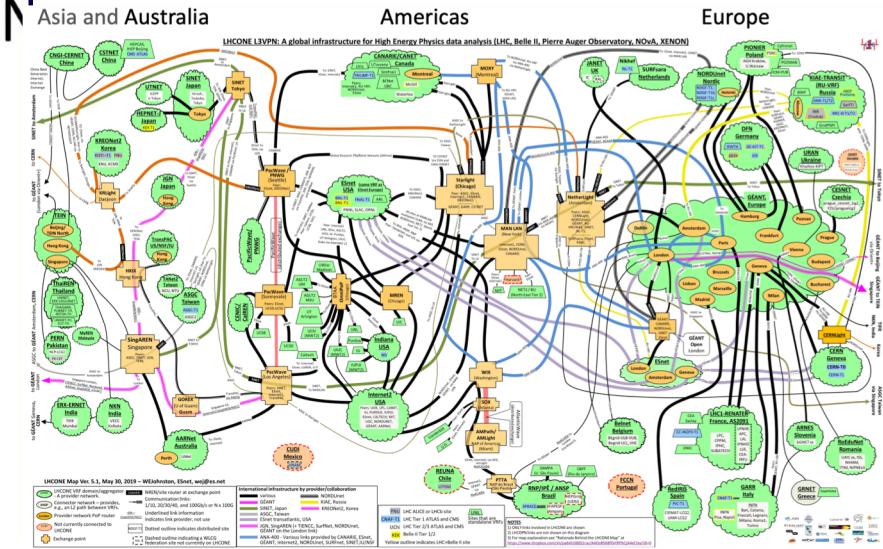
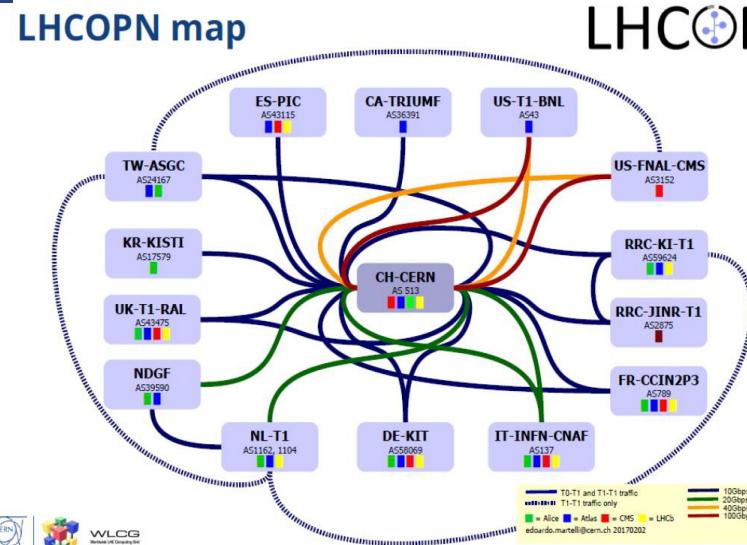
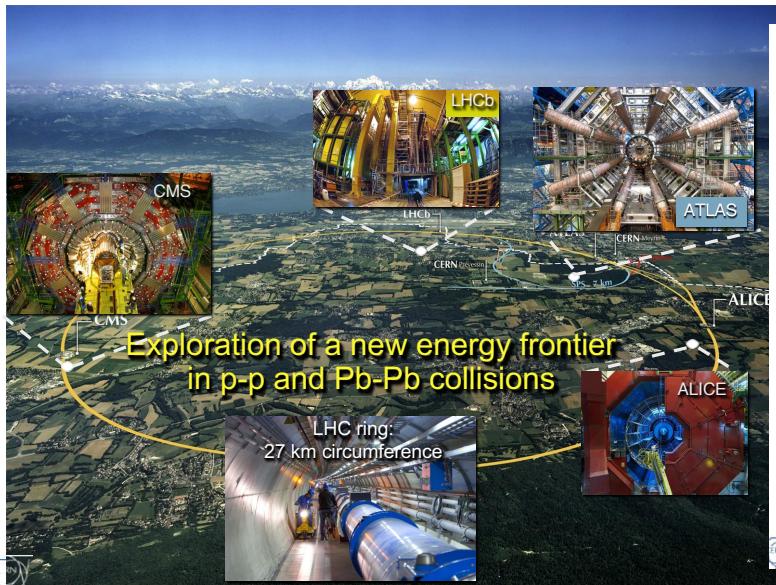
Presenters: Jordi Ros Giralt, Mario Lassnig, Mihai Patrascoiu, Y. Richard Yang
On behalf of team (Jensen Zhang, Kai Gao, Lauren Delwiche, Mahdi , Ryan Yang)

March 27, 2023

IETF 116

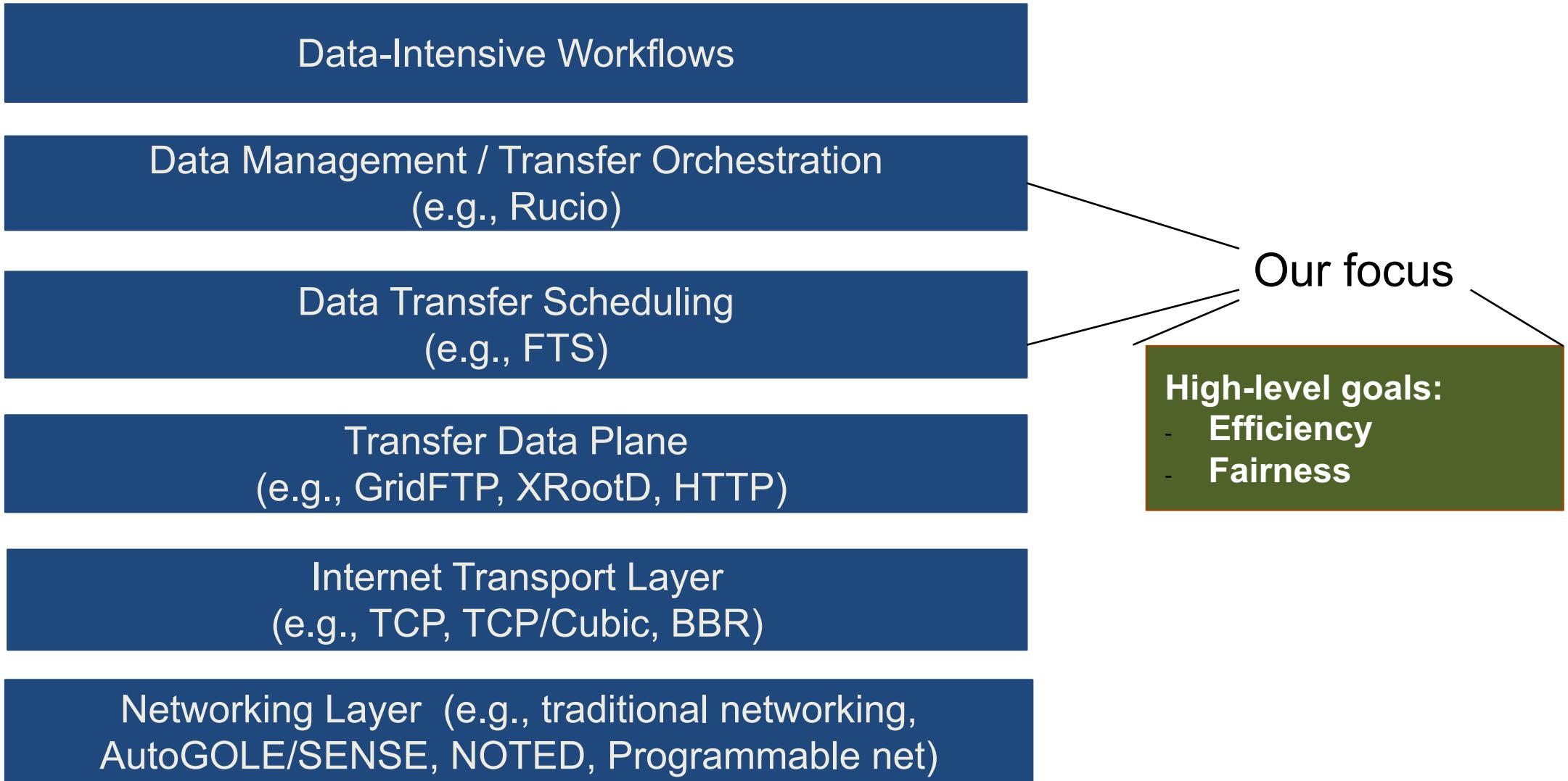
Overview: LHCONE

LHCONE



- Part of CERN infrastructure (experimental facilities, LHCOPN, LHCONE)
 - Topology: 600 distributed storage systems, distributed globally (170 data centers, in 127 sites, across 40 countries)
 - Workload: support data movement for four LHC experiments, and also Belle II, Pierre Auger Observatory, NOvA, XENON, and JUNO
 - Traffic: 2022, the aggregated outgoing traffic just from CERN to its ten largest connected data centres: 457 Petabytes of data.

Overview: Related Software Stack



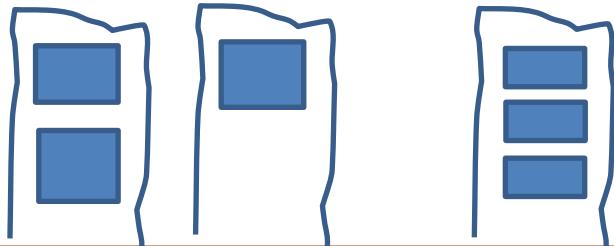
Outline

- Overview
- Transport scheduling integration (ALTO/FTS)

Transfer Scheduling (FTS): Objective and Design

Related objectives:

- Efficiency control: Avoid overloading transfer resources (both network and storage), fully utilize all capacity
- Fairness/allocation: resource allocation beyond congestion



Existing mechanisms:

- Keeps transfer queue for each src/dst pair (pipe)
- Adjusts # concurrent TCP connections per pipe;
- Dispatches transfer if allowed by concurrency level

New but not fully integrated mechanisms:

- Each file transfer (src->dst) is marked as on behalf of an activity of an experiment
 - Experiments identified as virtual organizations (aka tenants)

Data-Intensive Workflows

Data Management / Transfer Orchestration
(e.g., Rucio)

Transfer Scheduling
(e.g., FTS)

Transfer Data Plane
(e.g., GridFTP, XRootD, HTTP)

Internet Transport Layer
(e.g., TCP, TCP/Cubic, BBR)

Networking Layer (e.g., traditional networking,
AutoGOLE/SENSE, NOTED, Programmable net)

Protocol 1 FTS Model Analyzed (Called for High Success Rate)

```
1: Define  $RL(x) = \text{round}(\log_B(x))$ 
2: procedure OPTIMIZEGOODSUCCESSRATE(state)
3:   if cur.ema < prev.ema then
4:     if RL(cur.ema) < RL(prev.ema) then
5:       decision = prevValue - decreaseStepSize
6:     else
7:       decision = prevValue
8:     end if
9:   else if cur.ema > prev.ema then
10:    decision = prevValue + increaseStepSize
11:   else                                 $\triangleright$  emas are equal
12:    decision = prevValue + increaseStepSize
13:   end if
14: end procedure
```

A Semi Zero-Order Gradient Alg Optimizing for Each Pipe

Keep track of the exponential moving average (EMA) of throughput.

$$E_i(t+1) = \alpha T_i(t+1) + (1 - \alpha) E_i(t)$$

Update the number of connections based on EMA.

$$n_i(t+1) = \begin{cases} n_i(t) - 1 & RL_B(E_i(t+1)) < RL_B(E_i(t)); \text{ Line 4} \\ n_i(t) + 1 & E_i(t+1) \geq E_i(t); \text{ Lines 9,11} \\ n_i(t) & \text{else} \end{cases}$$

resource flexibility/fairness
control gap

efficiency gap

THEOREM 4.2 (CONSERVATION THEOREM). Let $K = \max \frac{M_i}{m_i}$. Then as long as $B > (1 - \alpha + \alpha K^2)$, the quantity

$$V_t(t) = n_i(t) - \text{round}(\log_B(E_i(t)))$$

only ever stays constant or increases.

Theorem: In a **Throughput-Deterioration Model**, semi zero order will achieve throughput that is $\leq 1/\sqrt{B}$ of the **optimal** (under default settings).

ALTO/FTS Objective: Application-Defined Networking

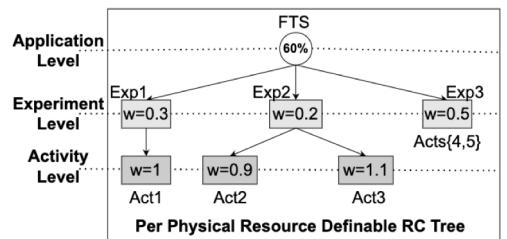
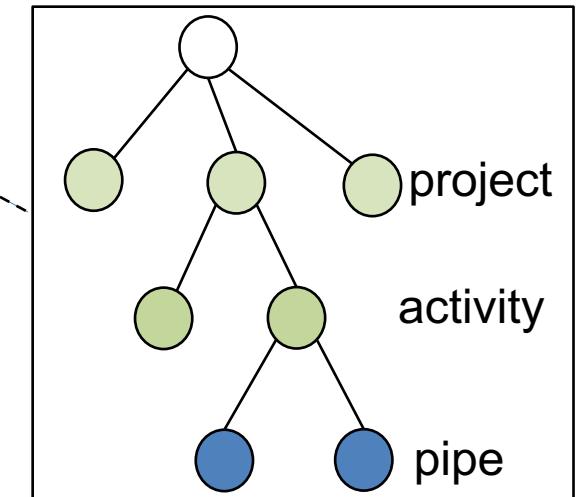
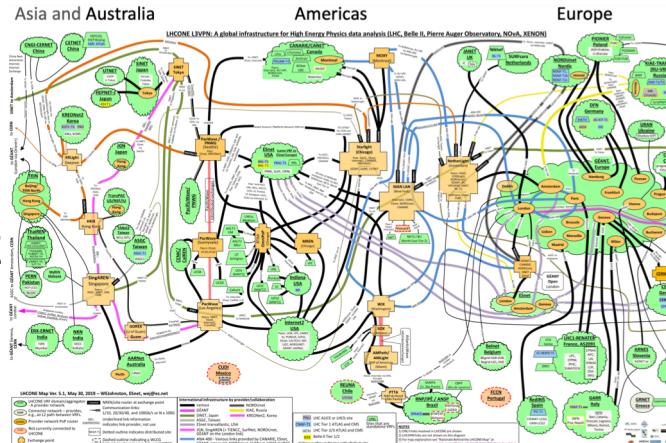
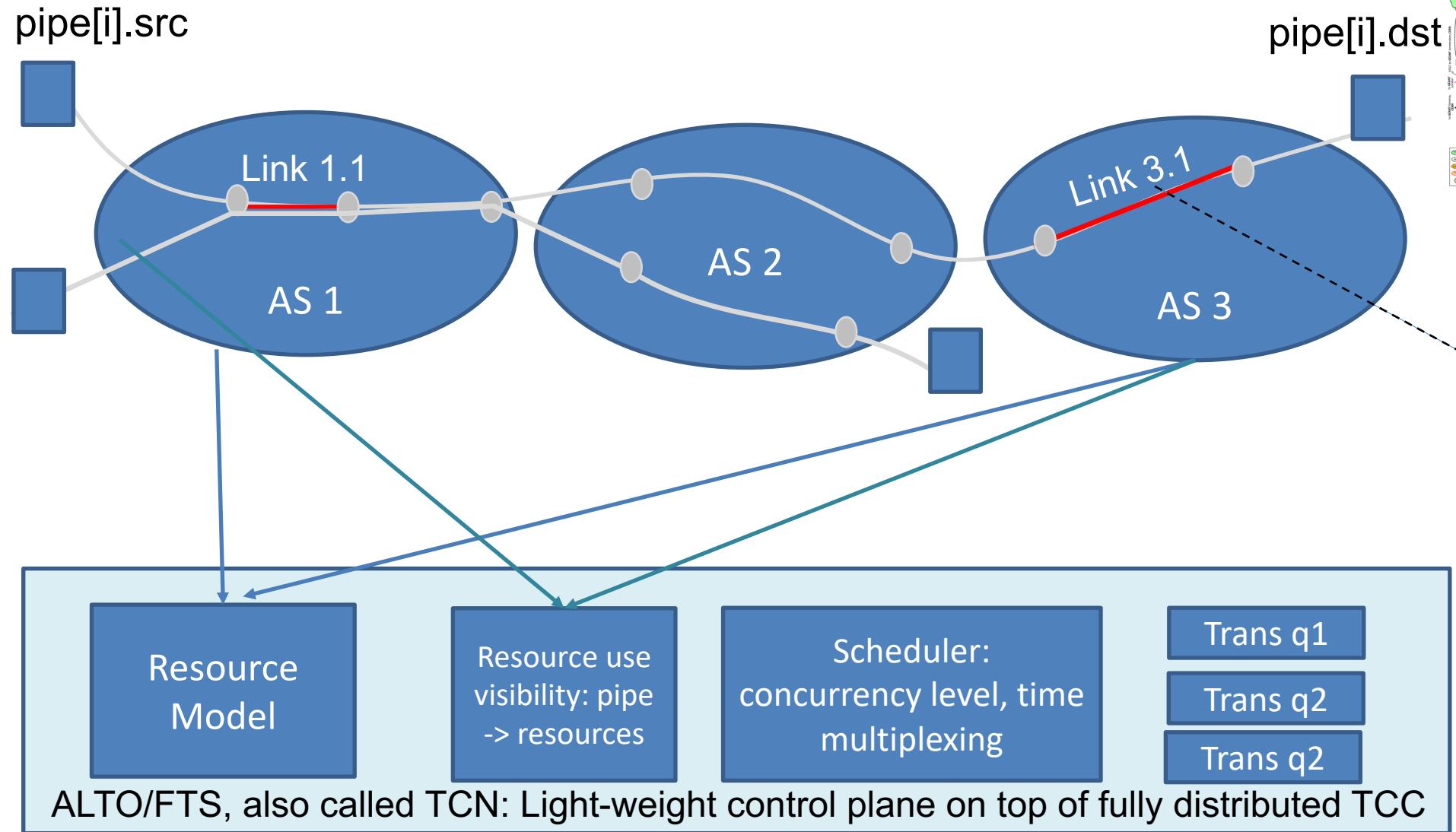
Efficient, Diverse, High Level Resource Models

Point2Point TCP
Congestion
Control (TCC)

- **Universally available**
- **Fast, efficient, robust** building block
- But **single** resource allocation (fairness) model

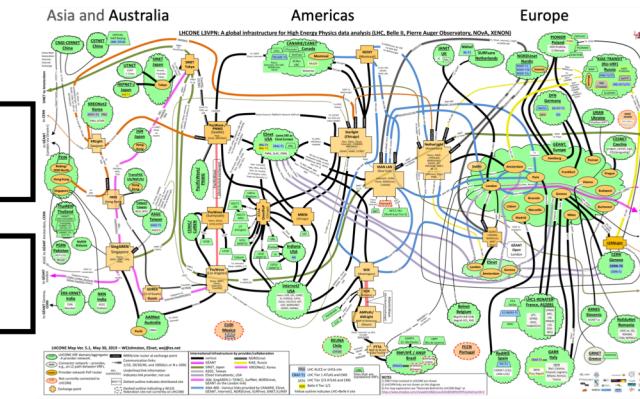
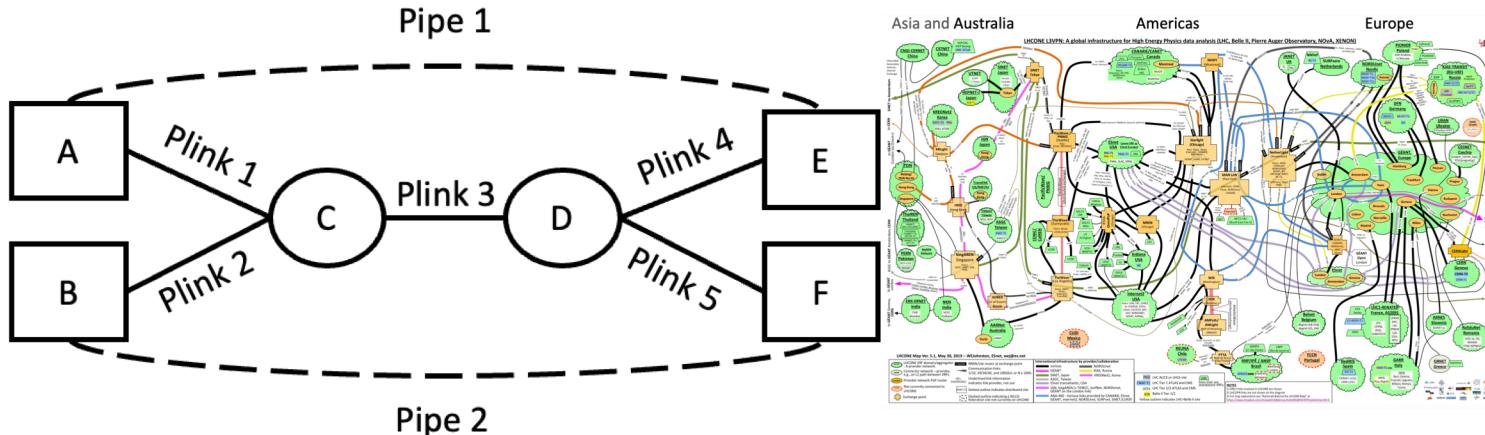
Diverse Network Settings

ALTO/FTS Architecture



Simplified Example Illustrating ALTO/FTS Visibility, Control

Resource Model:
Project 1 pipes: Pipe 1, Pipe 2
R1: <Project 1, Plink 1> <= 5G
R2: <Project 1, Plink 2> <= 10G
R3: <Project 1, Plink 3> <= 10G



App Provided State:

Pipe1.traffic = 3G, Pipe2.traffic = 9G

ALTO Provided State:

Pipe 1: {Plink 1, Plink 3, Plink 4}.
Pipe 2: {Plink 2, Plink 3, Plink 5}.



Resource use by project on physical resources:

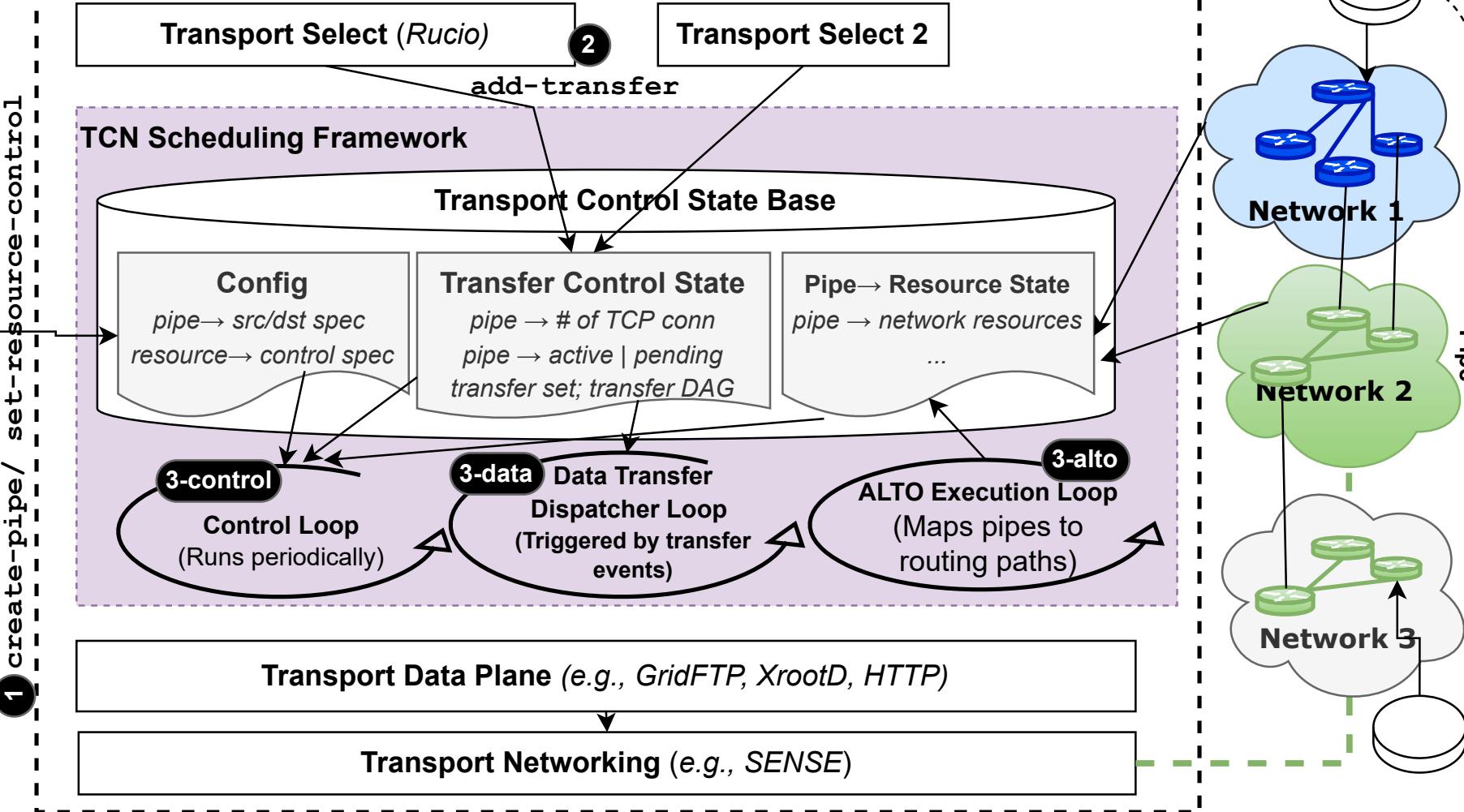
Plink1.traffic = 3G, Plink2.traffic = 9G,
Plink3.traffic = Pipe1.traffic + Pipe2.traffic = 12G

Penalty for resource control constraints:

$P(R1) = 0$ (Plink1 = 3G <= 5G), $P(R2) = 0$ (Plink2 = 9G <= 10G)
 $P(R3) = 2$ (Plink3 = 12G > 10G)

ALTO/FTS More Details

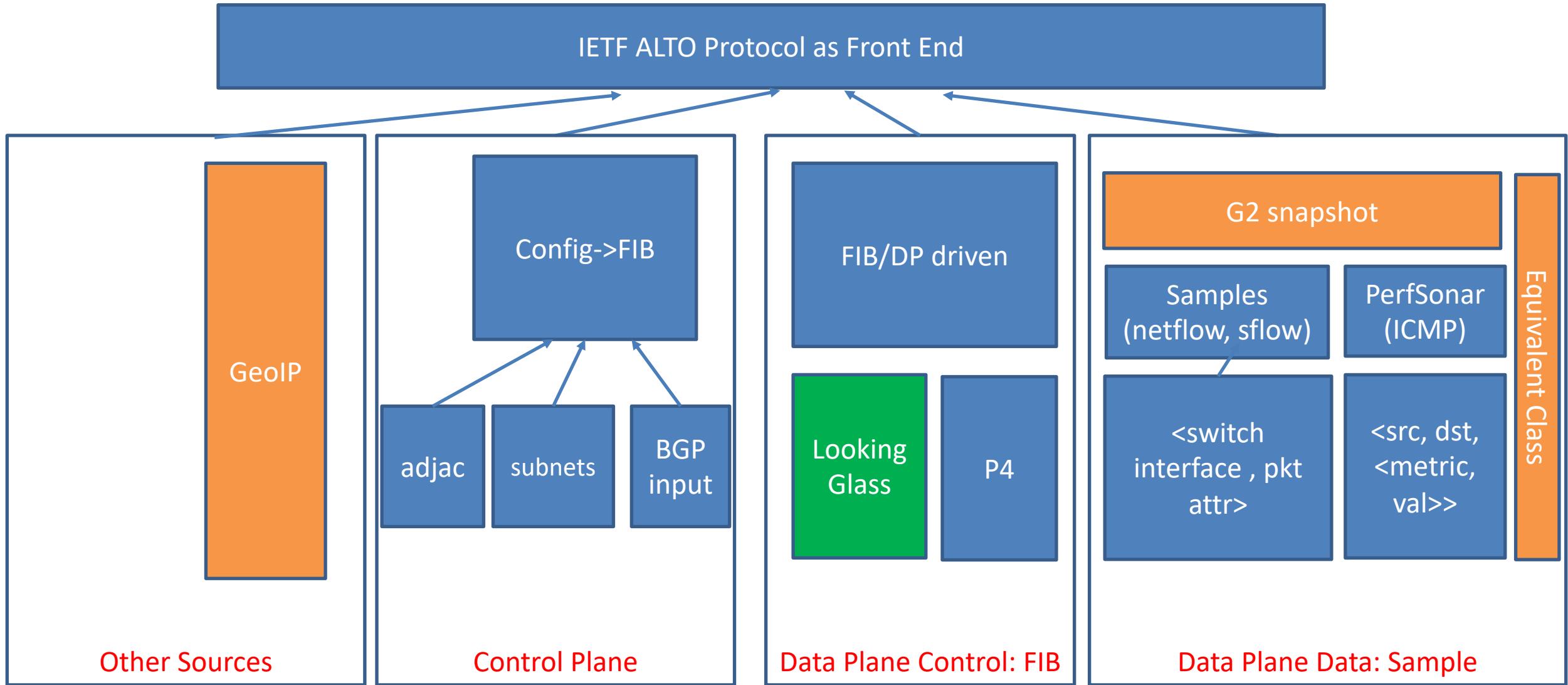
TCN Stack



Integration into FTS 3.12

- Implement [ALTO/TCNOptimizer](#) class for ALTO/TCN control loop
- Add new optimizer mode ([kOptimizerAggregated](#)) to enable ALTO/TCN optimizer
- Extend database schema for pipes ([t_link_config](#)) to support resource control specification ([tcn_abs_limit](#), [tcn_rel_weight](#))
- Build network resource constraints using the ALTO path vector client library

ALTO/FTS Visibility



ALTO/FTS First Hop Visibility

Query Example (ECS with path vector extension)

```
→ cat request-cern.json
{
  "cost-type": {
    "cost-metric": "ane-path",
    "cost-mode": "array"
  },
  "endpoint-flows": [
    {
      "srcs": [ "ipv4:137.138.0.101" ],
      "dsts": [ "ipv4:134.158.84.23", "ipv4:144.16.112.112" ]
    },
    {
      "srcs": [ "ipv4:192.16.166.254" ],
      "dsts": [ "ipv4:140.115.32.101" ]
    }
  ],
  "ane-property-names": [ "next_hop", "as_path" ]
}
```

Query/Response

```
→ curl -s -H 'Content-Type: application/alto-endpointcostparams+json' --data-ascii @
request-cern.json https://science.jensen-zhang.site/pathvector/cern-pv | ./pprint
--d41d8cd98f00b204e9800998ecf8427e
Content-Type: application/alto-endpointcost+json
Content-ID: <ecs@science.jensen-zhang.site>

{'endpoint-cost-map': {'137.138.0.101': {'134.158.84.23': ['autolink_1',
'autopath_2'],
'144.16.112.112': ['autolink_1',
'autopath_3']},
'192.16.166.254': {'140.115.32.101': ['autolink_1',
'autopath_1']}},
'meta': {'cost-type': {'cost-metric': 'ane-path', 'cost-mode': 'array'},
'vetag': {'resource-id': 'cern-pv.ecs',
'tag': 'e615bf984f7249949f8903c5cf56f02d'}}}
--d41d8cd98f00b204e9800998ecf8427e
Content-Type: application/alto-propmap+json
Content-ID: <propmap@science.jensen-zhang.site>

{'meta': {'dependent-vtags': [{resource-id': 'cern-pv.ecs',
'tag': 'e615bf984f7249949f8903c5cf56f02d'}]},
'property-map': {'.ane:autolink_1': {'next_hop': '192.65.184.145'},
'.ane:autopath_1': {'as_path': '20965 24167 7539 1659'},
'.ane:autopath_2': {'as_path': '20965 2091 789'},
'.ane:autopath_3': {'as_path': '20965 9885 55824'}}}
--d41d8cd98f00b204e9800998ecf8427e--
```

Routing Plane Retrieval (Looking Glass of CERN and GEANT)

```
etc > {} lg-agent.json > ...
1  {
2    "namespace": "default",
3    "agent_class": "alto.agent.cernlg.LookingGlassAgent",
4    "uri": "http://lhcone-lg.cern.ch/lg.cgi",
5    "default_router": "ex2j.cern.ch:juniper",
6    "refresh_interval": 300
7 }
```

Implementation

Jensen/Kai/Lauren

Outline

- Overview
- Transport scheduling integration (ALTO/FTS)
- Transport orchestration integration (ALTO/Rucio)

Transfer Orchestration (Rucio) Objective and Design

Related objective:

- Given replication rules (automatic workflow) or potential sources (manual workflow), pick the right source/destination

Existing mechanism:

- Sorting algorithm based on distance
 - Existing distance focuses on past tput

Data-Intensive Workflows

Data Management / Transfer Orchestration
(e.g., Rucio)

Transfer Scheduling
(e.g., FTS)

Transfer Data Plane
(e.g., GridFTP, XRootD, HTTP)

Internet Transport Layer
(e.g., TCP, TCP/Cubic, BBR)

Networking Layer (e.g., traditional networking,
AutoGOLE/SENSE, NOTED, Programmable net)

ALTO/Rucio Current State: Uniform Interface and Backend to Sort Replicas

BY=as_hopcount,delay_ow WHERE continent="EU"

Step 1: Configuration

Configure ALTO client at Rucio server to fetch visibility using ALTO

Step 2: Express Sorting Demands

ALTO sorting expression enables Rucio download command to sort replicas based on a combination of distances and properties

```
[client]
# ALTO server
default_ird = https://science.jensen-zhang.site/directory/default
metrics = {
    "as_hopcount": {
        "resource_type": "path-vector",
        "resource_id": "cern-pv",
        "prop_name": "as_path",
        "prop_transformer": "tolist | len",
        "aggr_transformer": "sum"
    },
    "delay_ow": {
        "resource_type": "cost-map",
        "resource_id": "delay-ow",
        "dependent_network_map": "default-networkmap"
    }
}
```

```
ainernet> rc rucio list-file-replicas --sort='alto;stmt="BY as_hopcount,delay_ow"' --metalink test
l version="1.0" encoding="UTF-8" >
alink xmlns="urn:ietf:params:xml:ns:metalink">
le name="file1">
identity>test:file1</identity>
ash type="adler32">69fe2b13</hash>
ash type="md5">12969016e761864f30f97dd5fb259e30</hash>
ize>1048576</size>
lfn name="/atlas/rucio/test:file1"></lfn>
rl location="XRD1" domain="wan" priority="1" client_extract="false">root://xrd1:1094//rucio/test/
rl location="XRD3" domain="wan" priority="2" client_extract="false">root://xrd3:1096//rucio/test/
rl location="XRD4" domain="wan" priority="3" client_extract="false">root://xrd4:1097//rucio/test/
ile>
talink>
```

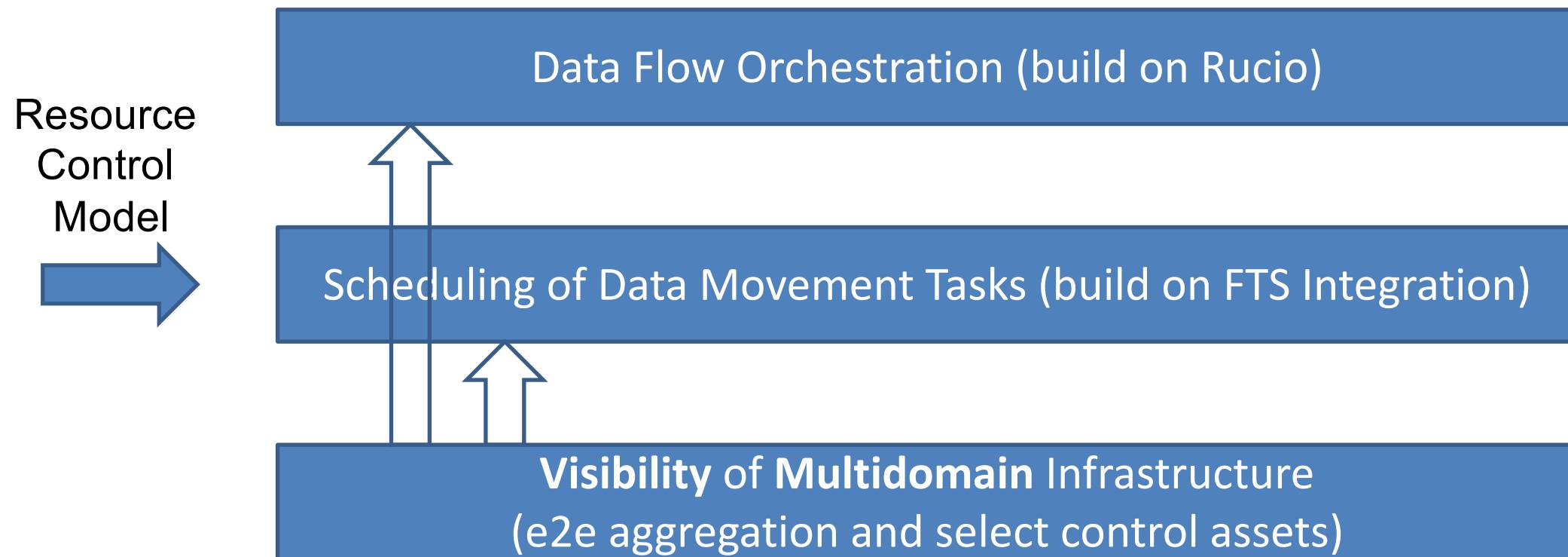
ALTO/Rucio Using ALTO as Uniform Interface: GeoIP and GeoDistance

- Providing geoip property using the standard ALTO endpoint property service **[RFC 9240]**
- Providing geo distance between endpoints using the standard ALTO Endpoint Cost Service (ECS) **[RFC 7285]**

```
→ curl -s -k -u cern:lhcone -H 'Content-Type: application/alto
-propmapparams+json' -d '{"entities": ["ipv4:198.17.101.70"]}'
https://localhost:8443/entityprop/geoip | jq .
{
  "property-map": {
    "ipv4:198.17.101.70": {
      "geolocation": {
        "lat": 32.8515,
        "lng": -117.2798
      }
    }
  }
}
```

```
etc > {} geoip-delegate-agent.json > ...
1  {
2    "namespace": "default",
3    "agent_class": "alto.agent.delegate.DelegateAgent",
4    "data_source_name": "geoip",
5    "data_source_config": {
6      "data_source_cls": "alto.agent.geoip.GeoipAgent",
7      "db_path": "/opt/geoip2/GeoLite2-City.mmdb"
8    },
9    "refresh_interval": 300
10 }
```

Summary: Current ALTO/FTS+Rucio: 3 Main Components



Status and Next Steps

- Implementations
 - ALTO/FTS
 - Visibility: looking glass first-hop links (e.g., CERN border links to peers)
 - Control: Zero-order stochastic gradient algorithm, event-driven programming
 - Scale: 200x40, targeting 600x600 full mesh
 - Resource model: Full linear model
 - ALTO/Rucio
 - Fully integrated, uniform interface
- Deployment
 - Target full production workload in summer 2023, for HL-LHC Data Challenge

Backup Slides

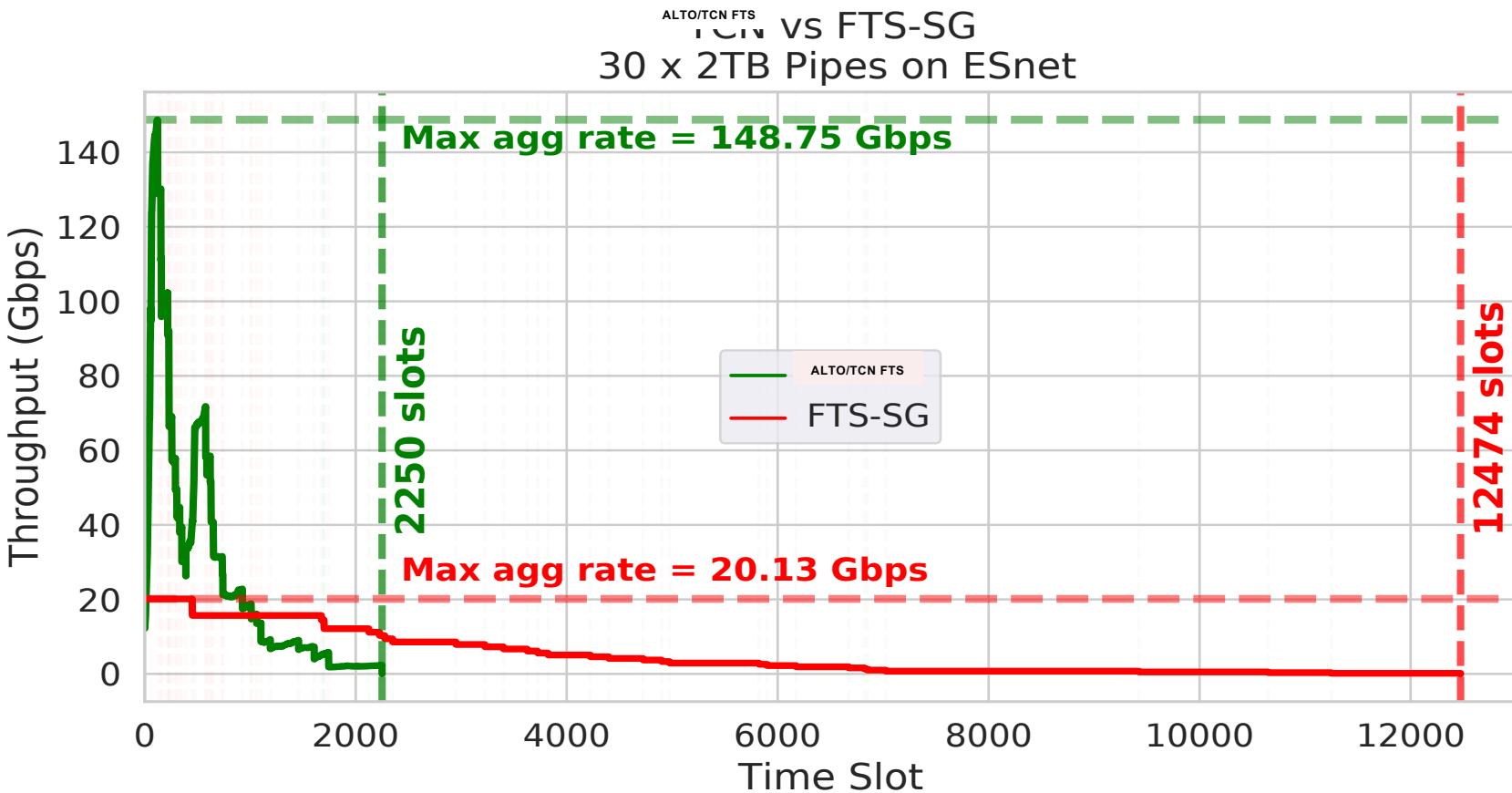
ALTO/FTS Details

- Integral, quadratic distance function
- Zero-order stochastic rounding

$$U(\tau) = \left(\sum_{i=1}^K w_i \tau_i \right) - \eta \cdot d(\tau, t \cdot K)^2$$

1. Basic Gradient	Gradient of control state n_i : $(\frac{da}{dn_i})$	$\frac{da}{dn_i} = \sum_{j=1}^K \frac{da}{dT_j} \cdot \frac{dT_j}{dn_i}$
1.1.	$\frac{dT_j}{dn_i}$ is the gradient of the bottleneck	If $T_j(n) = \min(f_{j,1}(n), \dots, f_{j,b}(n))$ and $k = \operatorname{argmin}_i f_{j,k}(n)$, then $\frac{dT_j}{dn_i} = \frac{df_{j,k}}{n_i}$
1.2.	Decide zero (implicit) or first order (w/ analytical expr)	$\frac{df_{j,k}}{dn_i} = \begin{cases} \text{zero-ord est.} & \text{for blackbox } f_{j,k} \\ \text{first-ord grad.} & \text{otherwise.} \end{cases}$
1.2a.	Zero order estimate	$G(n, z) = \frac{f_{j,k}(n+z) - f_{j,k}(n)}{\ z\ ^2} \cdot z$
1.2b	First order computation	Compute analytical expression: $\frac{df_{j,k}}{dn_i}$
2. Momentum-Based		
Gradient Acceleration		Compute $g = (\frac{da(n)}{dn_1}, \frac{da(n)}{dn_2}, \dots, \frac{da(n)}{dn_K})$; Update $\mathbf{m} = (1 - \alpha)\mathbf{m} + \alpha \cdot (\eta g)$; $n = cur.\mathbf{n} + int(\mathbf{m})$;
3. Discretize	$int(x) = \begin{cases} \lfloor x \rfloor & \text{with probability } 1 - (x - \lfloor x \rfloor) \\ \lfloor x \rfloor + 1 & \text{with probability } x - \lfloor x \rfloor. \end{cases}$	

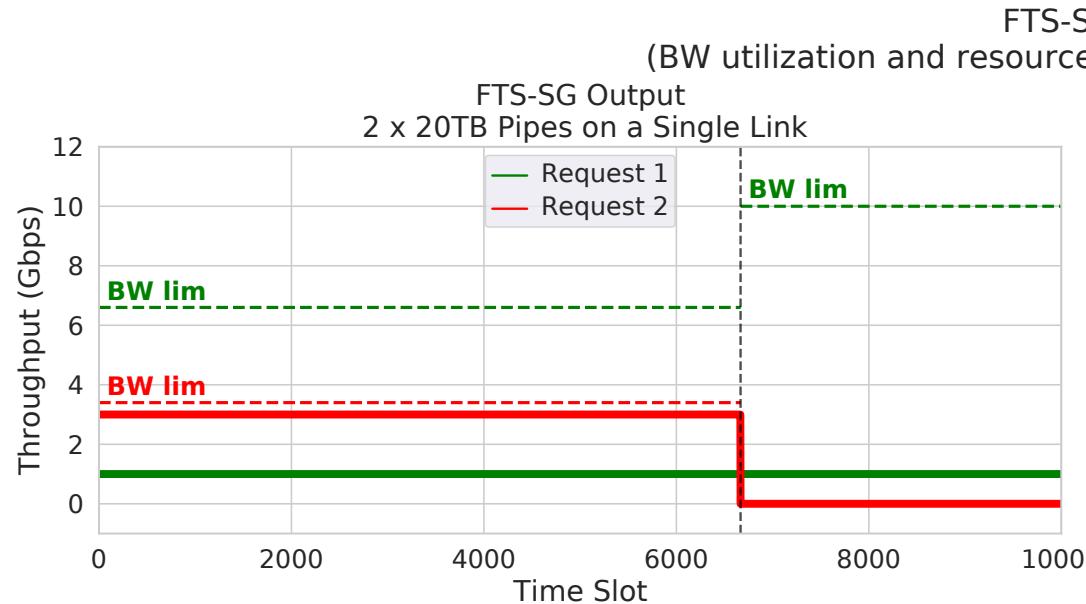
Basic Benchmarking \Rightarrow Real Topology (ESnet)



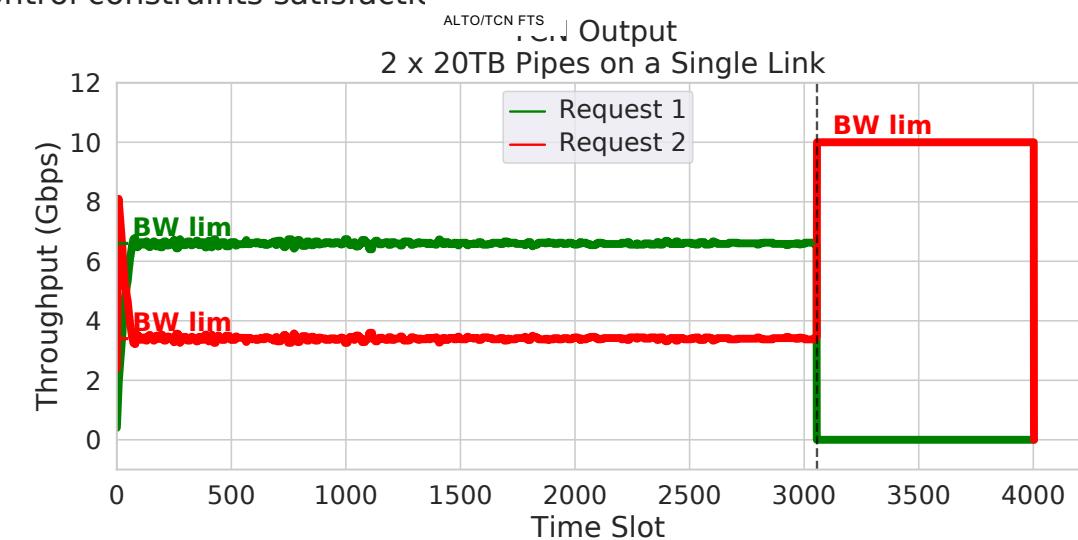
Setting: 30 <src, dst> pipes, one request per pipe, each request 20K transfers, file size = 100MB. the total in the workload is 60TB.
Resource Control goal: all equal weights

1. **7.39x** total BW utilization.
 2. **5.54x** Max RCT improvement. (Short-tailed)
- Global Objective, Zero-order gradients, and Resource Control Constraints.**

Basic Benchmarking: Results



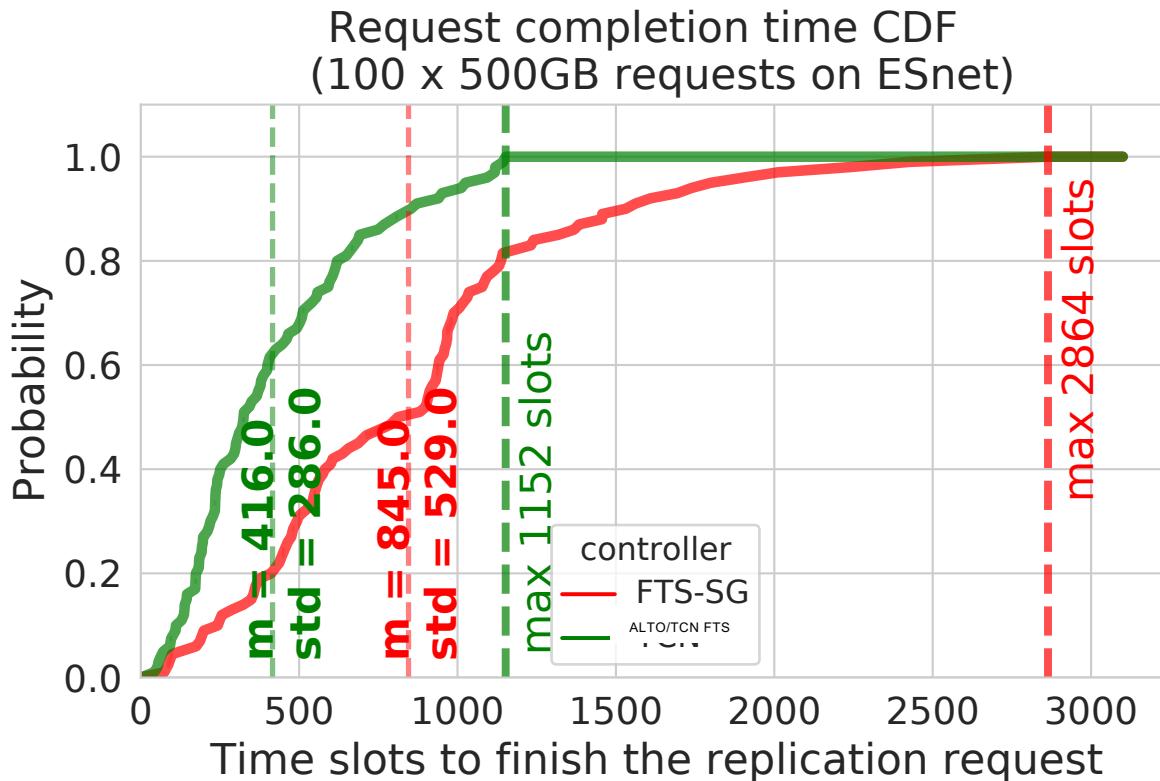
FTS-SG vs ALTO/TCN FTS
(BW utilization and resource control constraints satisfaction)



	FTS Semi-Gradient (FTS-SG)	ALTO/TCN FTS
Total BW Utilization	Not optimal (6GB unused BW).	Fully utilized BW.
Resource Control	BW shares not satisfied.	BW allocated close to 2:1 spec.
RCT	Max RCT = 20,000 slots.	Max RCT = 4,000 slots.

FTS-SG depends on correct configuration (e.g., high enough default). **ALTO/TCN is fully automated.**

Request Performance Distribution



1. 2.03x mean RCT improvement.
2. 2.49x max RCT improvement.

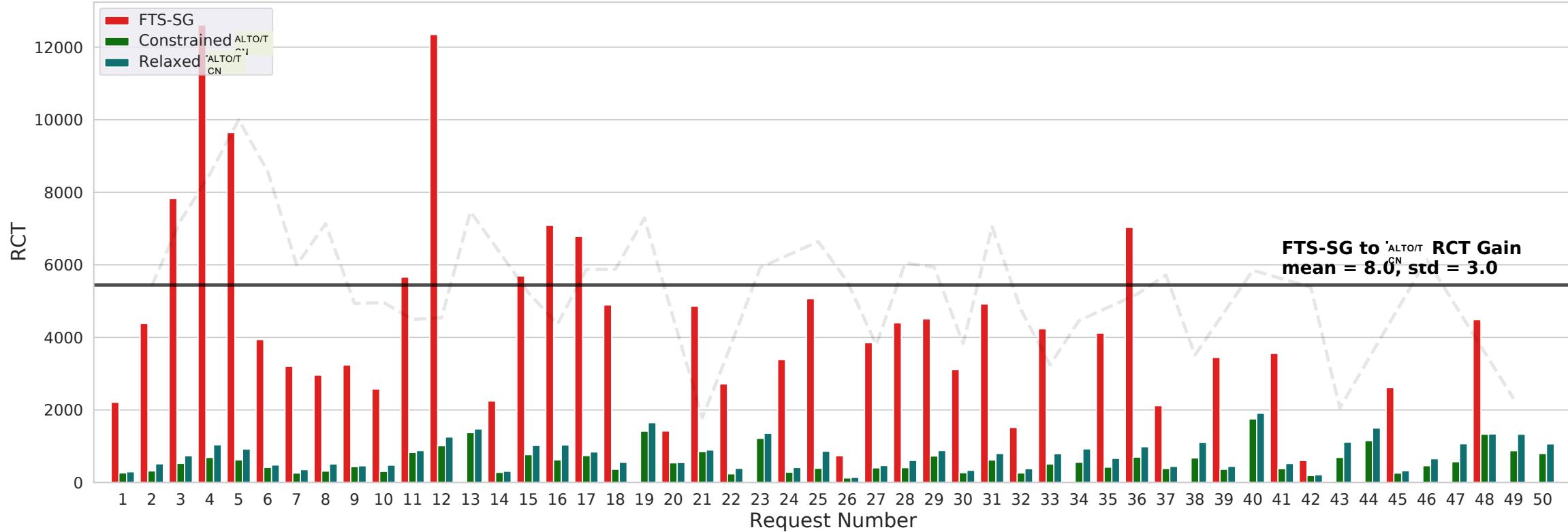
Global Objective, Zero-order gradients, and Resource Control Constraints.

Setting: Similar to previous slide, but with modification to include more requests to show more details: 100 <src, dst> pipes, one request per pipe, each request 5K transfers, file size = 100MB.

Resource Control goal: all equal weights.

From All-Arrival Workload \Rightarrow Dynamic Arrival Workload

Per flow comparison of $\text{ALTO/T}_{\text{CN}}$ and FTS Request Completion Times
50 request with incremental arrival time on ESnet

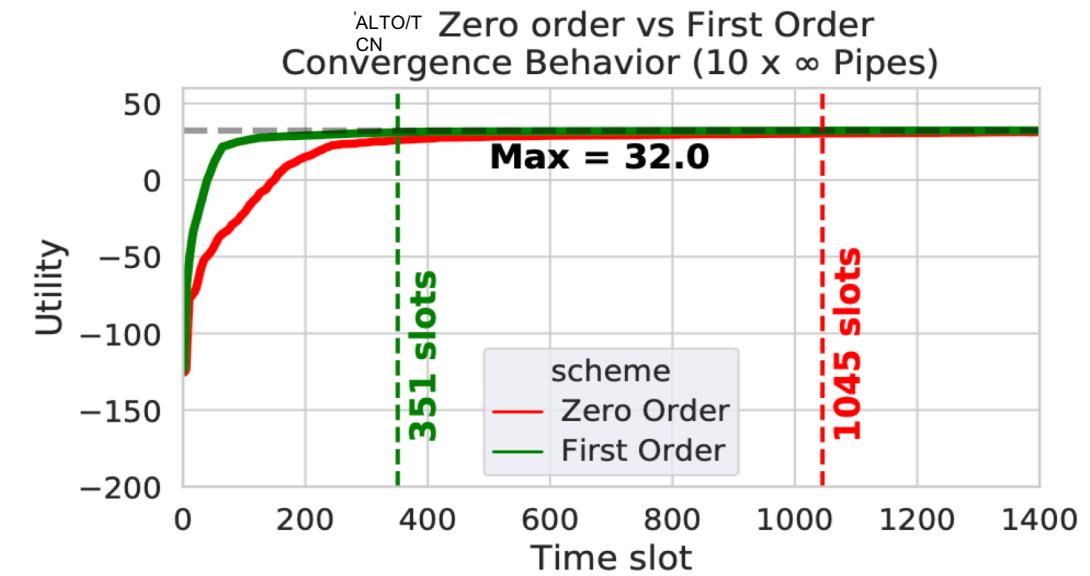
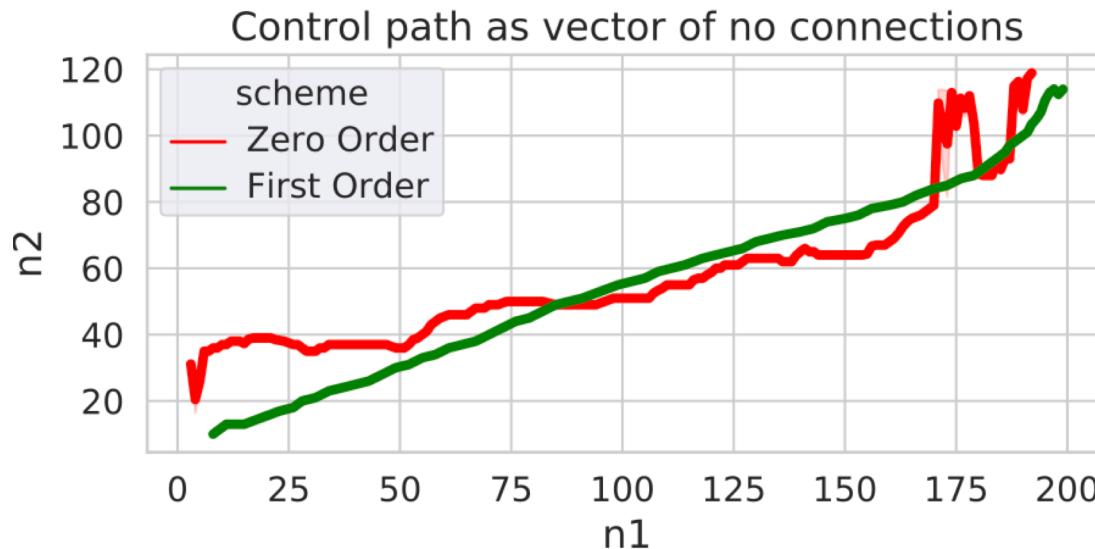


FTS-SG to $\text{ALTO/T}_{\text{CN}}$ RCT Gain
mean = 8.0, std = 3.0

Setting: ESnet (67 nodes), selected 50 active pipes; each pipe has transfer workload arrives according to a arrival distribution (Poisson arrival, with parameter 1200 (every 200 time slot); Each replication request has $N(40k, 20k)$ files, file size is 100MB.

**8.0x improvement in RCT when using ALTO/TCN.
(Global Objective, Full Zero Order)**

ALTO/TCN Zero-order vs First-order



2.97x improvement in convergence speed (# transient slots) with using first-order gradients.

First-order integrated ALTO/TCN vs Full Zero-order ALTO/TCN

Setting: ALTO/TCNology ESnet, selected 10 pipes, infinite backlog.

General Context: Exa-Scale Data Systems



High Luminosity LHC



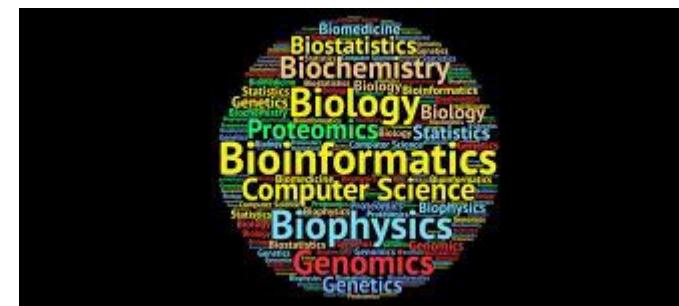
SKA Australia Telescope Facility



Vera Rubin Observatory



KSTAR Korea Superconducting Tokamak Next Gen Advanced Photon Source



Bioinformatics/Genomics

- Collect, distribute, and process data from scientific instruments (aka large IoT)