Introducing Falcon as a Parsl Data Provider

Mousa AbuGhosh and Engin Arslan
University of Nevada Reno
University of Texas at Arlington

Need for Wide-Area Transfer Optimization

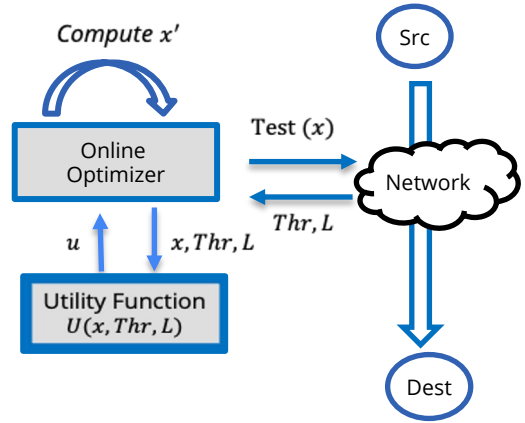
- Science application are becoming increasingly data-intensive
  - Single high throughput genome sequencing data increased from 5MB in 2006 to 5GB in 2018.
- Produced data is often transferred for processing or archival.
- While research networks have sufficient capacity (up to 100Gbps) existing transfer solutions fail to utilize these resources effectively.
- Concurrent transfer of multiple files (i.e., transfer parallelism) improves transfer throughput by improving both I/O and network throughput. How much concurrency is enough?

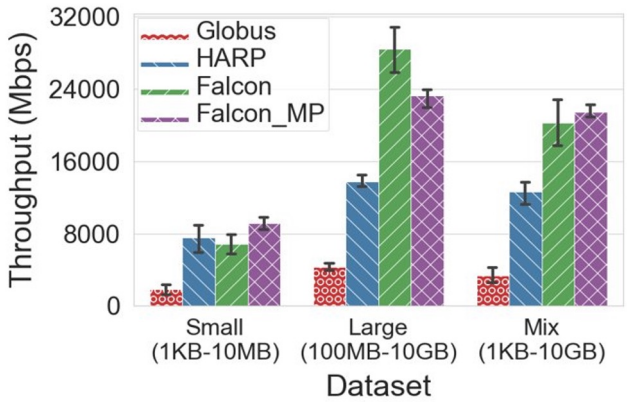
Falcon: Online Wide-Area File Transfer Optimization Algorithm

- Falcon is a transfer protocol is designed and to this day being improved by the joint effort of Md Arifuzzaman, Engin Arslan and recently Mousa AbuGhosh (me)
- Falcons objectives include:
  - Low overhead:
    - High concurrency levels can overwhelm the network, storage and end hosts
  - High performance :
    - Using a fixed and small concurrency level leads to poor system utilization
  - Fair resource sharing:
    - Lack of penalty term in optimization functions lead to aggressive behaviors to maximize the gain

Falcon Design and Performance Comparison

- Online, decentralized and black-box transfer optimization
- Two main components
  - Optimization Algorithm
  - Utility Function





2.2x higher throughput compared to Globus
55% higher throughput than HARP

Falcon as Parsl Data Provide - Version 1

This approach uses the typical approach that other data-providers are currently using:

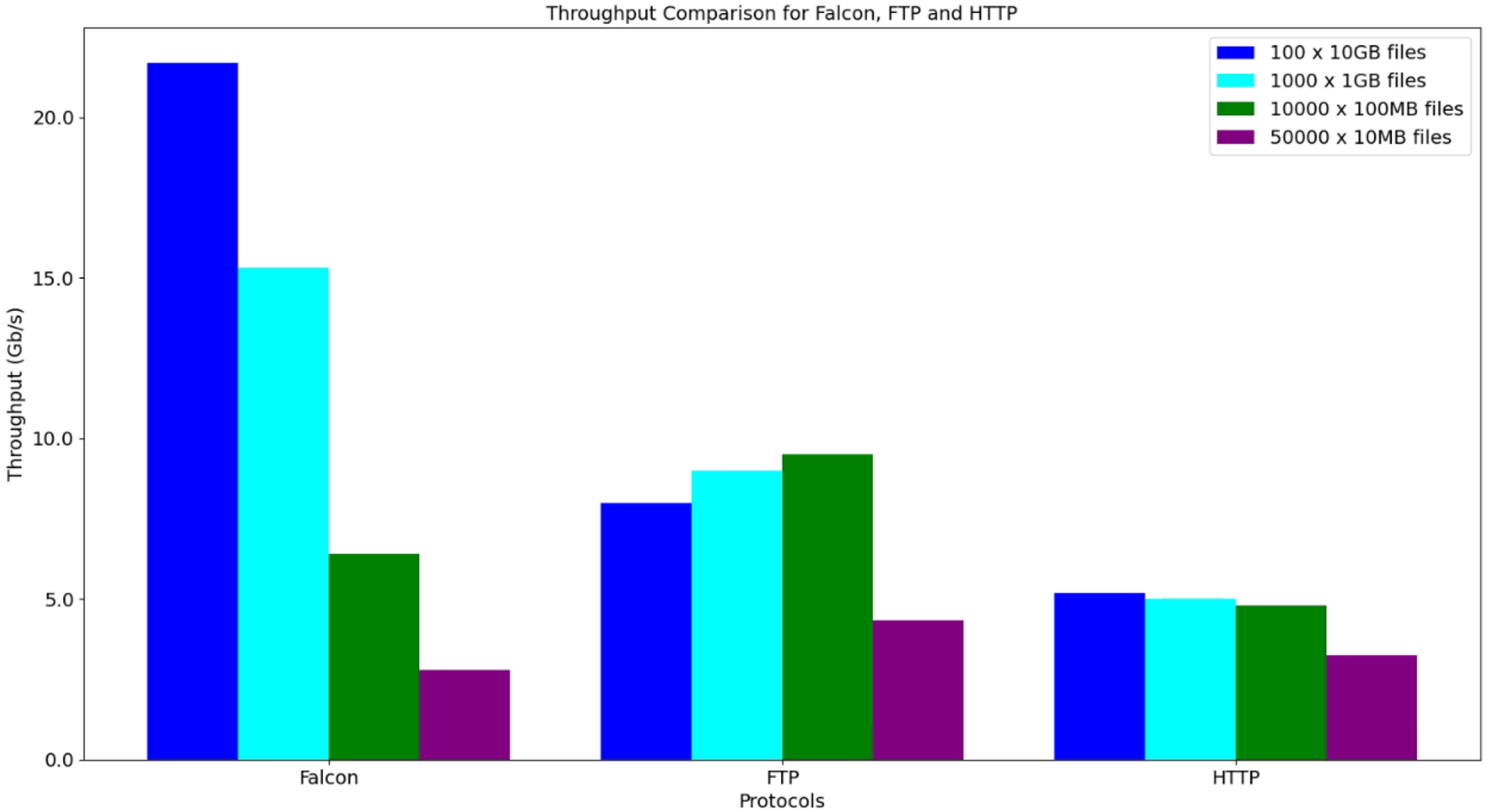
ex:

```
inputs = []
for x in range(0, n):
        inputs.append(File('falcon://{netloc}{working_dir}{filename}'))
```

In this approach we take use of the python package zmq to achieve this transfer and receive file acknowledgment.

In this approach, adjustments needed to be made to Falcon to make it compatible.

Falcon as Parsl Data Provide - Version 1 Throughput



Falcon as Parsl Data Provide - Version 2

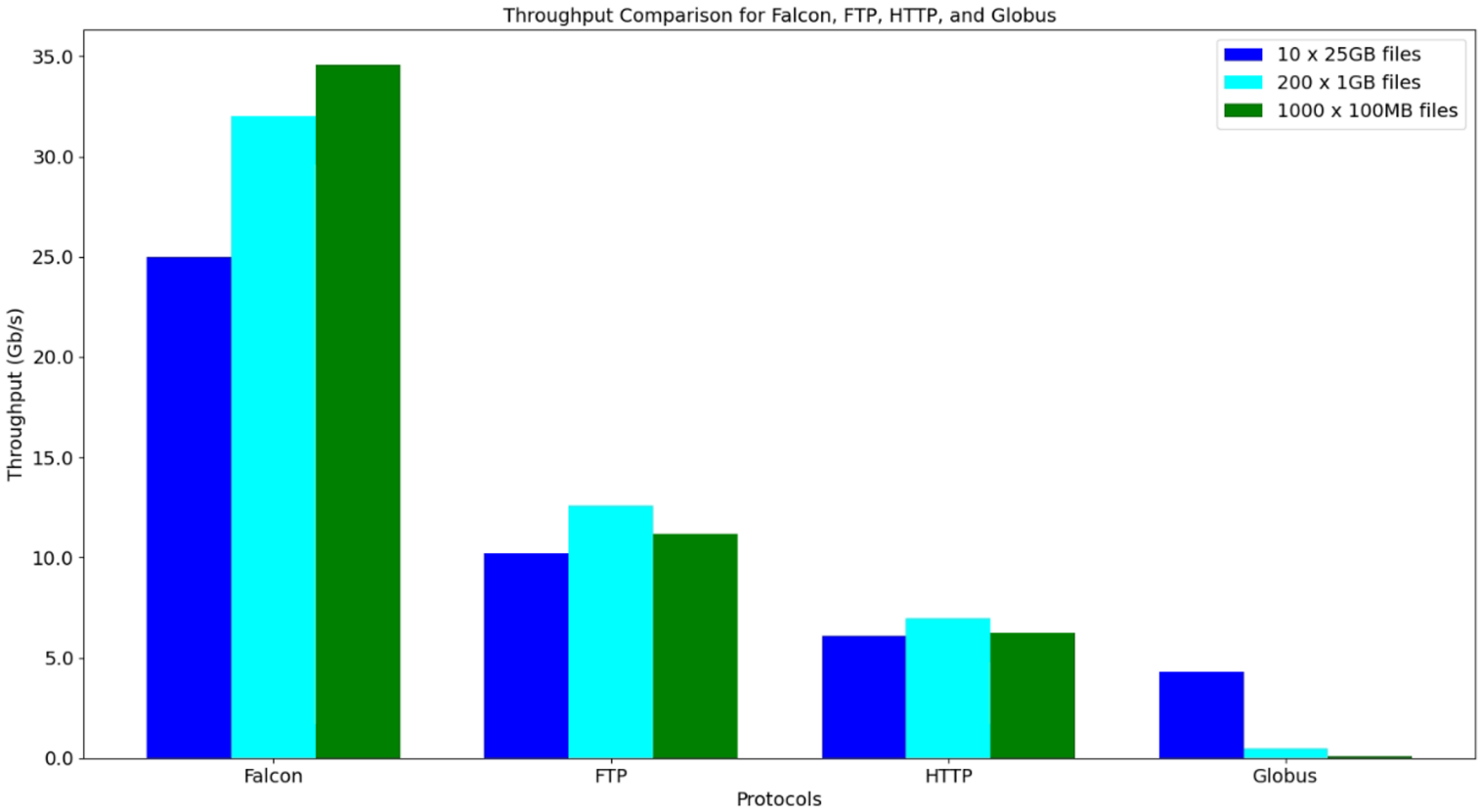
This approach include transferring an entire directory in Parsl using Falcon as a data-provider

ex: Input = Dir('falcon://{netloc}{directory path}')

In this approach we use a different of parsl “File” object and create a similar object “Dir” which serves the same purpose and almost have the same properties with the difference being our object as the name infers is an entire directory rather than a file.

We take use of the python package falcon-datamover to achieve this transfer.

Falcon as Parsl Data Provide - Version 2 Throughput



Falcon as Parsl Data Provide

- Both versions need to have a falcon-sender command running on the source host.
- Version 1 requires a falcon-receiver command running on the destination host (in Version 2 Parsl data-provider takes care of that).
- Version 2 provided a higher average throughput than version 1 mostly because of the bottleneck created by the wait of acknowledgment.
- Improvement for both versions is still being tested.

Thank You!

Email: mabughosh99@gmail.com

LinkedIn: https://www.linkedin.com/in/mousa-abughosh-72061518a/