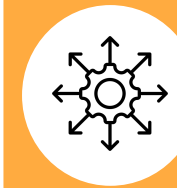
StreamHub: High-performance Managed SciStream as a Service

Seena VazifeDunn, Flavio Castro,
Prof. Ian Foster, Prof. Kyle Chard, Dr. Rajkumar Kettimuthu

November 2025

Overview

Scientific Data Streaming
Challenges and Motivations
RAM-to-RAM Streaming
State of the art Streaming Service



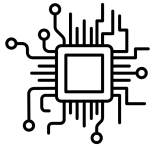
StreamHub
Built on top of SciStream
StreamHub Components
StreamHub Architecture

Results
Streaming Importance
Performance
Conclusion

What is Scientific Data Streaming

High data rate:

Modern scientific facilities (synchrotrons, telescopes, accelerators) can generate **petabytes of data daily**



Requires high performance computing infrastructure:

Lack of computation infrastructure at the scientific facilities requires data computation at HPC facilities

High operational costs:

Large-scale experiments' cost can reach **tens of thousands of dollars per hour**

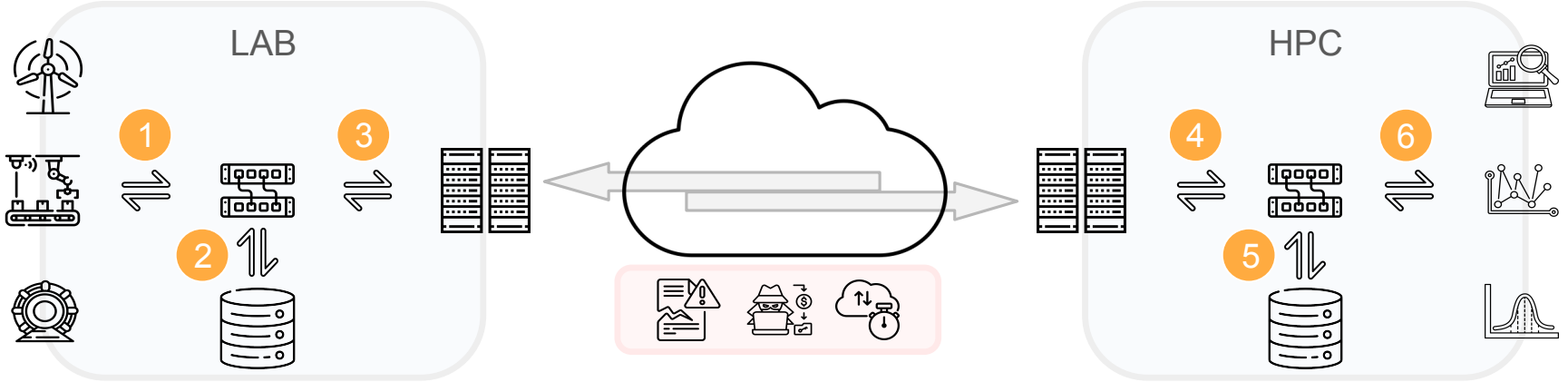
Challenges

Traditional batch-based workflows:

Slow Processing: Data are delayed until the transfer completes

High Storage Costs: Large intermediate storage is required

Limited Real-Time Capability: No immediate processing as data arrives



RAM-to-RAM Streaming:

Data streaming workflows:

Fast Processing: Data are processes as production rate

No Storage Costs: No intermediate storage is required

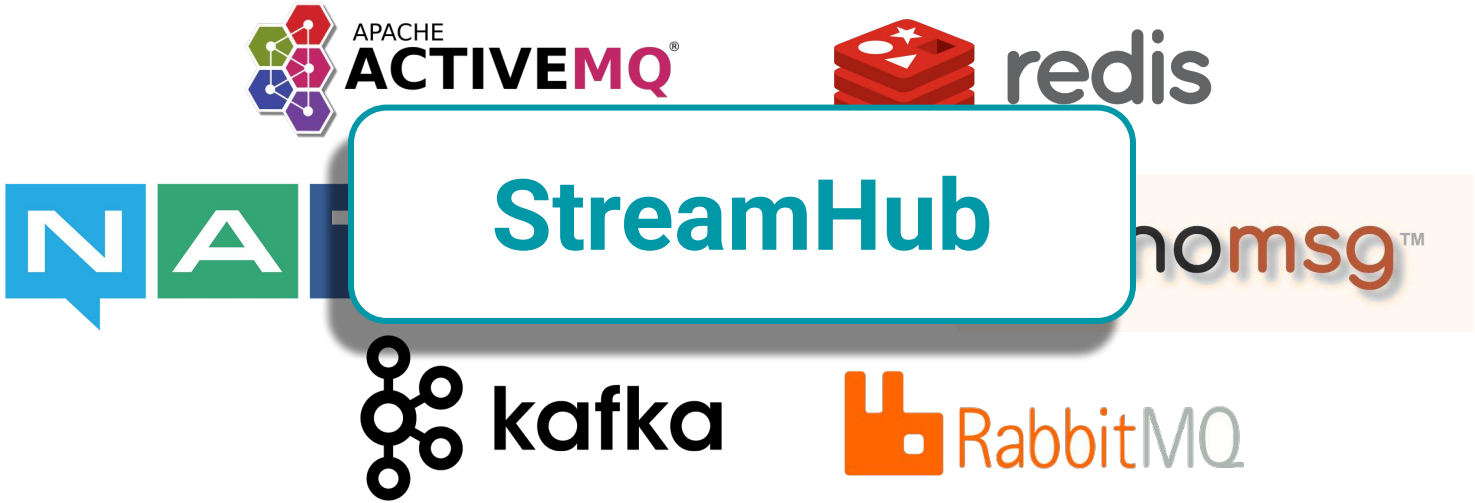
Real-Time Capability: Immediate processing with no I/O delay

Streaming Services & Applications

No federated security model, managed and automated orchestration

Streaming Services & Applications

No federated security model, managed and automated orchestration



Built on top of SciStream

Middlebox Architecture, Allocates Resources, Establishes Streaming Channel

User Control (S2UC): Interacts with user and software to initiate the service

Control Server (S2CS): Interacts with S2UC and S2DS to setup the service

Data Server (S2DS): Streams data between the distributed networks

SciStream Limitations

Separate User Authentication in each domain

→ Creating **inconsistent security policies**

Manual User Intervention for establishing secure tunnels

→ Increasing **risk of misconfiguration** and setup time

Firewall Modification at each institutions to permit streaming sessions

→ Adding **overhead** and potential **delays**

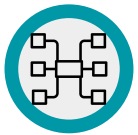
Lacking Integrated Monitoring Support and Centralized access control

→ Limiting operational **visibility**

StreamHub

A **High-performance**, **Scalable**, **Secure** Scientific Data Streaming Service

The secure authentication and authorization fabric of **Globus Auth**



The remote function execution capabilities of **Globus Compute**

Leverages the high-performance data streaming capabilities of **SciStream**

System Design

Separated **Control Plane** & **Data Plane**

- **StreamHub Client:**

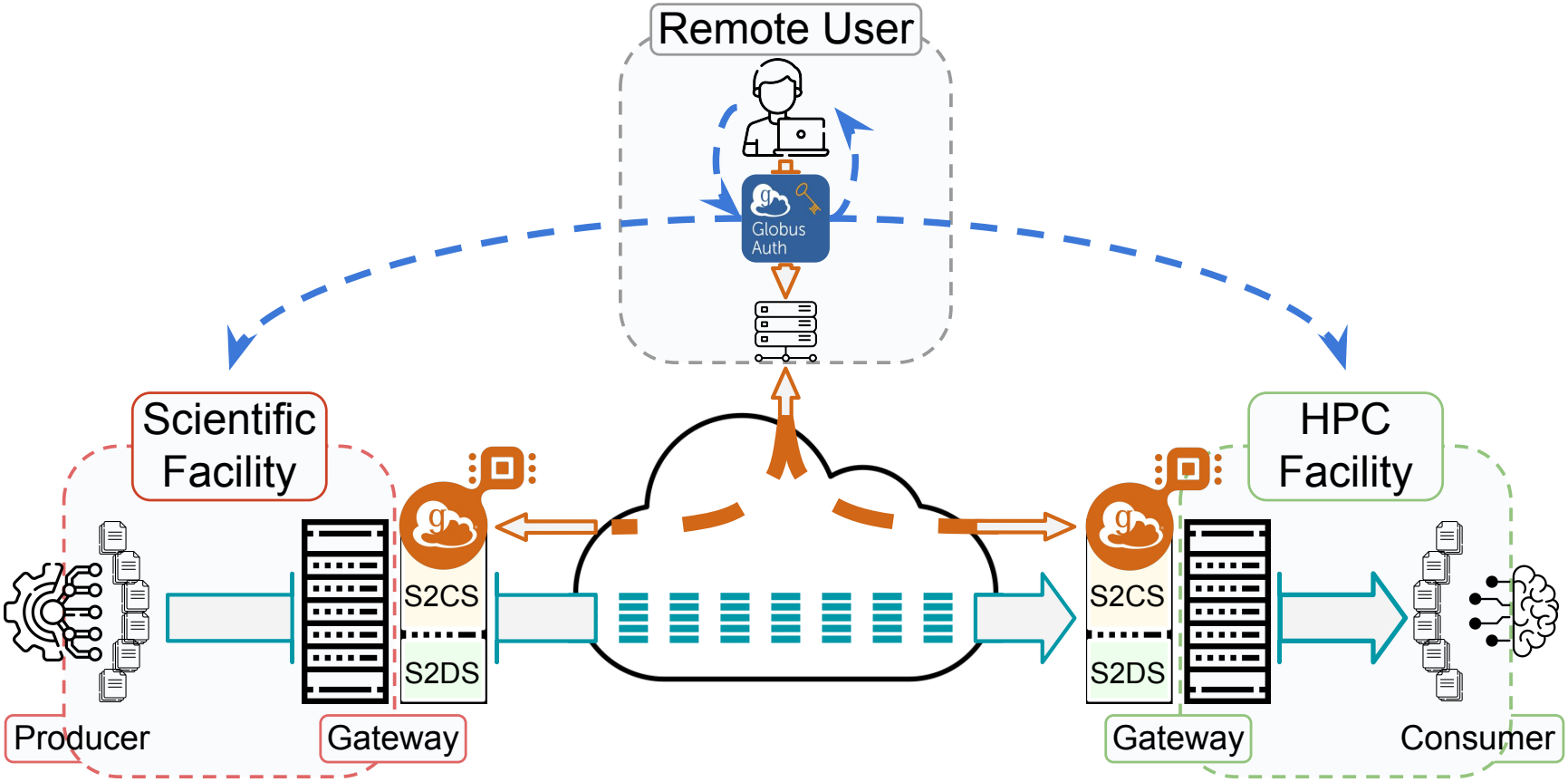
User Authentication, Parameter negotiation, Managing the life cycle

- **Stream Control Agents:**

Execute the task, generate session credentials, lunch the components

- **Streaming Data Plane:**

Streaming encrypted data over streamings channels

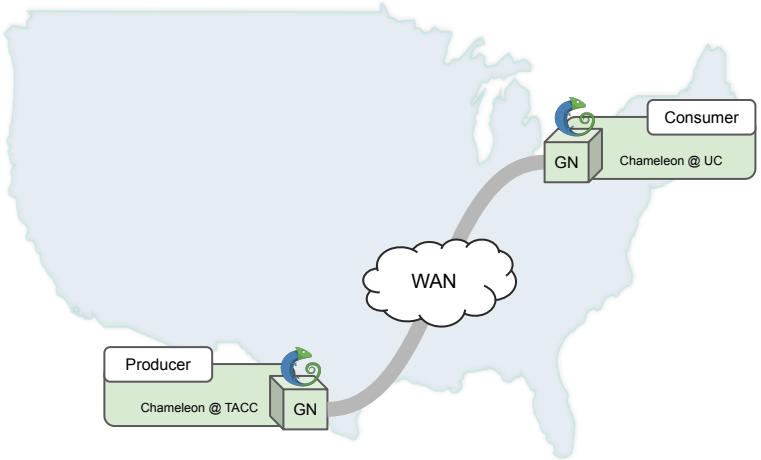

Remote User
Globus Auth
Scientific Facility
HPC Facility
S2CS
S2DS
S2CS
S2DS
Producer
Gateway
Gateway
Consumer

Evaluation

Evaluating over **6,700 experiments** on Chameleon Cloud

Experimental parameters across all configurations

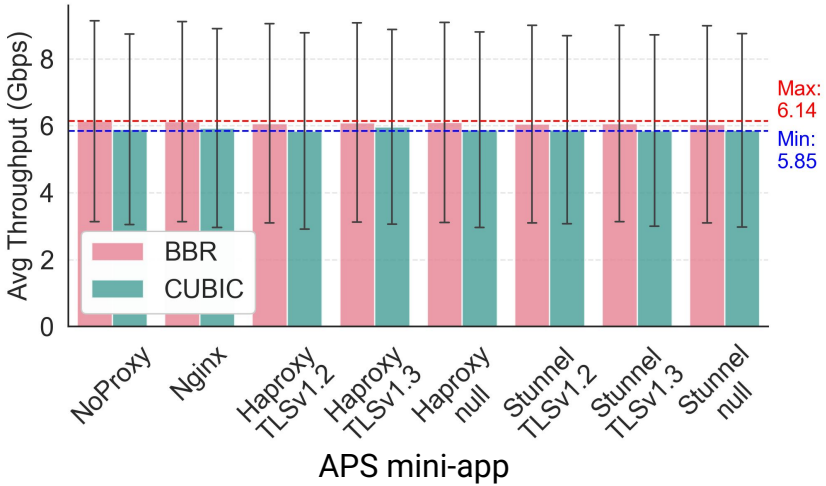
- **Baselines:** Nginx, No-Proxy
- **Applications:** iPerf3, APS mini-app
- **Proxy:** Stunnel, Haproxy
- **Encryption:** Plain TCP, TLSv1.2, TLSv1.3
- **Congestion Control:** CUBIC, BBR
- **Parallelism:** 1, 3, 5 concurrent stream
- **Duration:** 10s, 20s, 30s, 60s
- **Iteration:** 10 runs per configuration



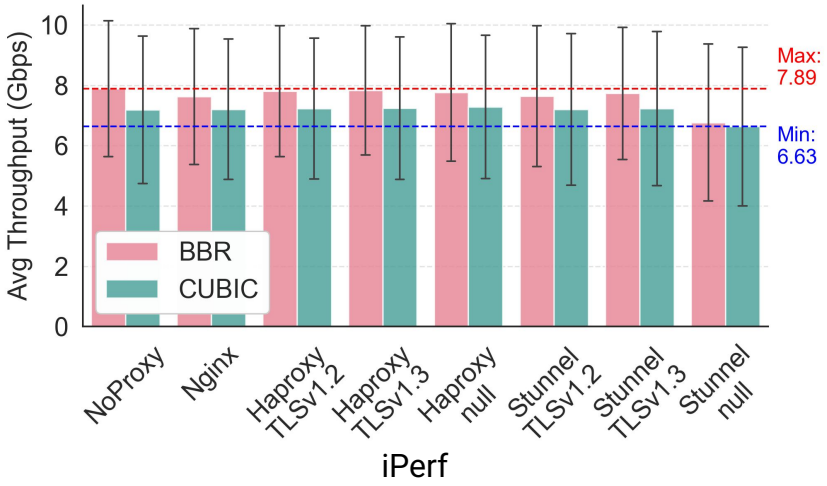
Proxy & Congestion Algorithm Comparison

Near data acquisition speed secure streaming

Higher throughput: **BBR** compared to **CUBIC**



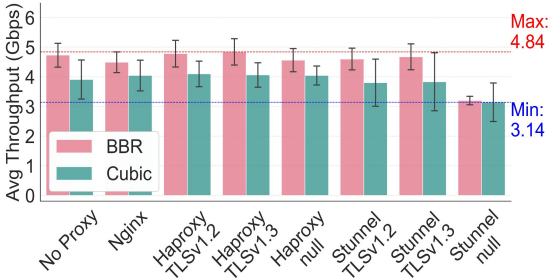
APS mini-app

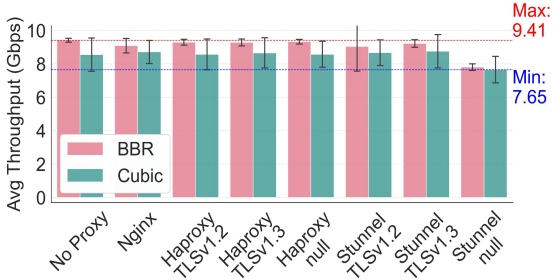


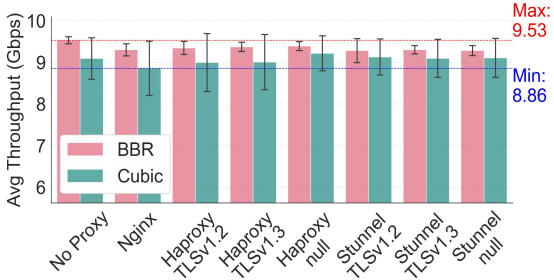
iPerf

Application & Parallelism

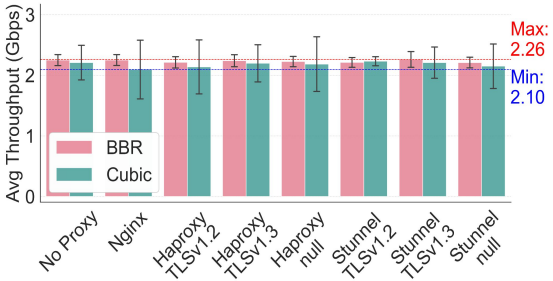
Max throughput: iPerf with **3 streams** & APS mini-app with **5 streams**

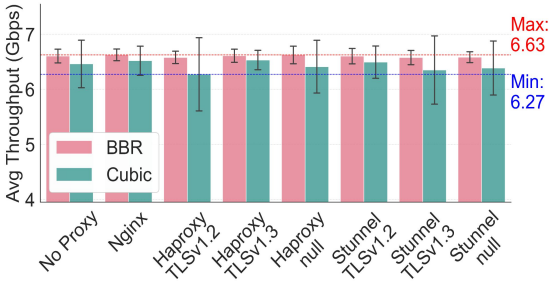


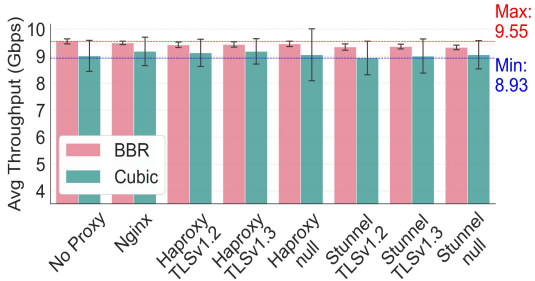




iPerf



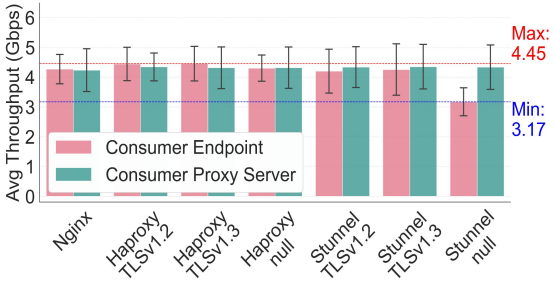


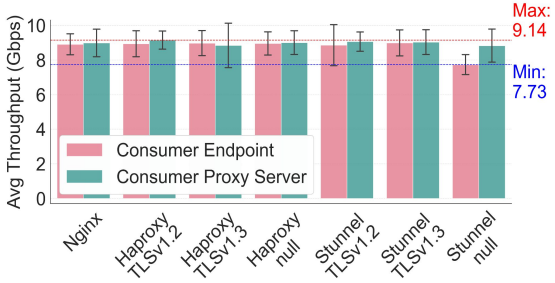


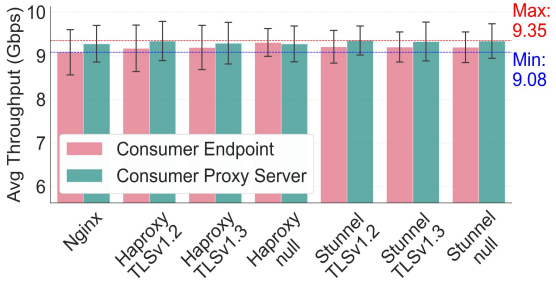
APS mini-app

End-to-End vs. Gateway-to-Gateway

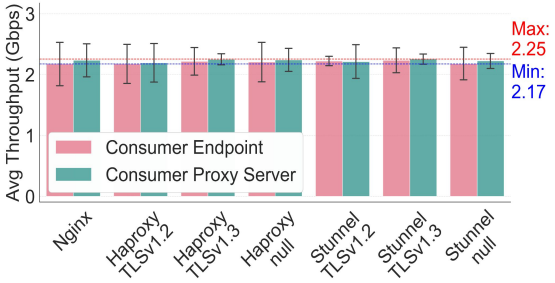
Negligible throughput reduction of only **less than 2%**

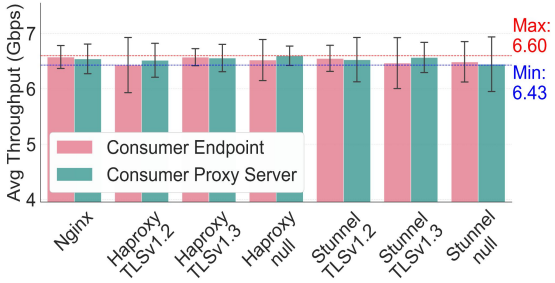


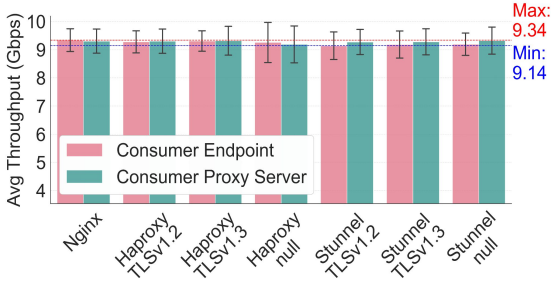




iPerf







APS mini-app

Control Plane Orchestration Overhead

Negligible Overhead of only **~4.3s** for the entire end-to-end orchestration

Operation	Time (s)	Notes
Authentication	~ 0.7	Access token verification
Service Discovery	~ 0.12	Queried Globus services
Function Registration	~ 0.19	Registered remote functions
Task Submission	~ 0.21	Submitted control tasks
Session Setup	~ 0.65	SciStream session initiation
Inbound Setup	~ 0.41	Producer gateway configured
Outbound Setup	~ 0.47	Consumer gateway configured
Additional delays	~ 1.55	Scheduler, network jitter, overlaps
Total Duration	~ 4.3	End-to-end orchestration overhead

Conclusion

Scientific discovery depends on real-time, secure streaming

Traditional approaches **add latency**, are **costly** and **inefficient**

Collect data → Transfer to storage → Later analyze on HPC

Requires storage → High latency → No online steering → Costly

Current approaches **lack federate security**, **automation**, or **scalability**

Federated security model • Manual intervention • Admin-level permissions

Network modification • Remote execution

StreamHub: **Secure End-to-End, High-Throughput, Scalable, Efficient**

Thank you.