



Scalable autonomic
sTreaming middleware
for **REAL** time processing
of **MASSIVE** data flows

Scalable Autonomic Streaming Middleware for Real-Time Processing of Massive Data Flows

Ricardo Jimenez-Peris
Universidad Politecnica de Madrid
Project Coordinator

Project Data

- Start: February 2008.
- Duration: 3 years.
- Partners:
 - UPM – Spain (coord.).
 - FORTH - Greece.
 - TU Dresden - Germany.
 - Telefonica - Spain.
 - Exodus - Greece.
 - Epsilon - Italy.

Background

- Data streaming is a new paradigm developed in the database community to process large data flows in memory in an online fashion.
- It allows to perform continuous queries over flowing data.
- Most existing platforms are centralized, and a few distributed, and perform 1-2 orders of magnitude better than relational DBs.

Scope

- Many potential applications in Internet today require to process huge amounts of information in an online fashion:
 - Mitigation of DDoS attacks.
 - Spam filtering.
 - Processing the output of sensor networks.
 - Detecting fraud in cellular telephony.
 - Financial applications.
 - QoS monitoring for enforcing SLAs.
 - Real time data mining.
 - Etc.

Objectives

- Stream aims at developing a highly scalable middleware infrastructure to process massive data flows in real time.
- The innovation lies in the **sheer scale targeted** by the project 1-2 orders of magnitude higher than current technology.

Innovation

- Parallelizing data streaming operators:
 - Currently a query operator can be deployed on a single site and it has to process the full data flow thus becoming the bottleneck.
 - Stream is developing distributed versions of query operators that enable to run individual query operators in a cluster of sites.
- Exploiting leading edge high performance networks and IO systems:
 - Reaching 40 gbs for both networking and IO.
 - This results in high throughput communication among sites and very low latency.

Innovation

- **Self-healing:**
 - Able to tolerate failures.
 - Able to online recover new nodes.
- **Self-configuring:**
 - Dynamic load balancing.
- **Self-provisioning:**
 - Nodes are added and removed as needed depending on the load.

Expected Outcome

- Highly scalable and autonomic infrastructure to process massive data flows.
 - 2 orders of magnitude more scalable than current distributed data streaming platforms.
- Application to 3 different markets:
 - Telco: Fighting fraud in cellular telephony.
 - Services: Real-time checking of SLAs fulfillment.
 - Financial/banking: Detection of laundry financial operations/Fraud detection in credit card payments/Real time data warehousing.