UBA
Universidad de Buenos Aires

ICC Instituto de Ciencias de la Computación

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Software and Computing efforts in Argentina

Including particle tracking and network simulation

Rodrigo Castro
University of Buenos Aires and
ICC-CONICET, Argentina.
rcastro@dc.uba.ar

Latin American Workshop on Software and Computing challenges in High-Energy Particle Physics (LAWSCHEP 2019)

November 20-23, 2019
México City, México

Agenda

1. AthenaMT: multi-threaded framework for ATLAS workflows
   (DF/FCEN/UBA)

2. Rucio: Distributed Data Management for ATLAS
   (IFIP/UNLP)

3. Particle Tracking Simulation for CMS
   (DC/FCEN/UBA)

4. Network and Data Flow Simulation for the ATLAS DAQ system
   (DC/FCEN/UBA)

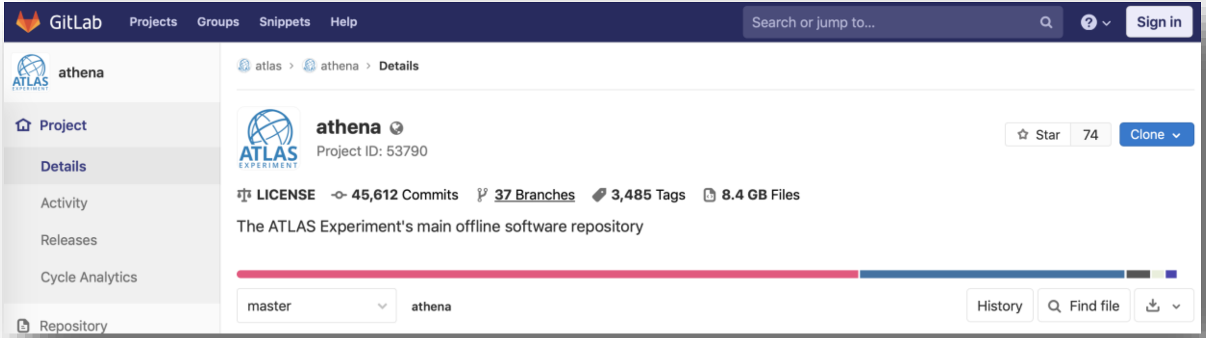
AthenaMT: multi-threaded framework for ATLAS workflows

Physics Department (DF)
School of Exact Sciences (FCEN)
University of Buenos Aires (UBA)
Argentina

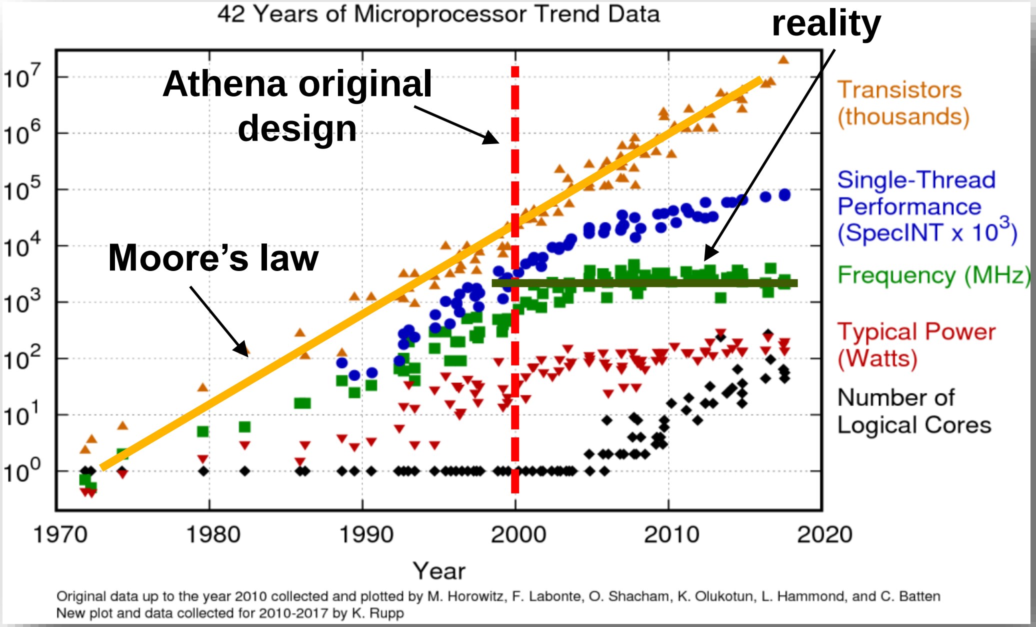
On behalf of Gustavo Otero

Athena

- Athena is the ATLAS offline software
  - Based on the Gaudi framework
    - Open Project
    - Provides interfaces and services to build HEP experiment frameworks for event data processing applications
  - Manages most of ATLAS production workflows (event generation, simulation, reconstruction, high level trigger and derivation production)
  - It is public!

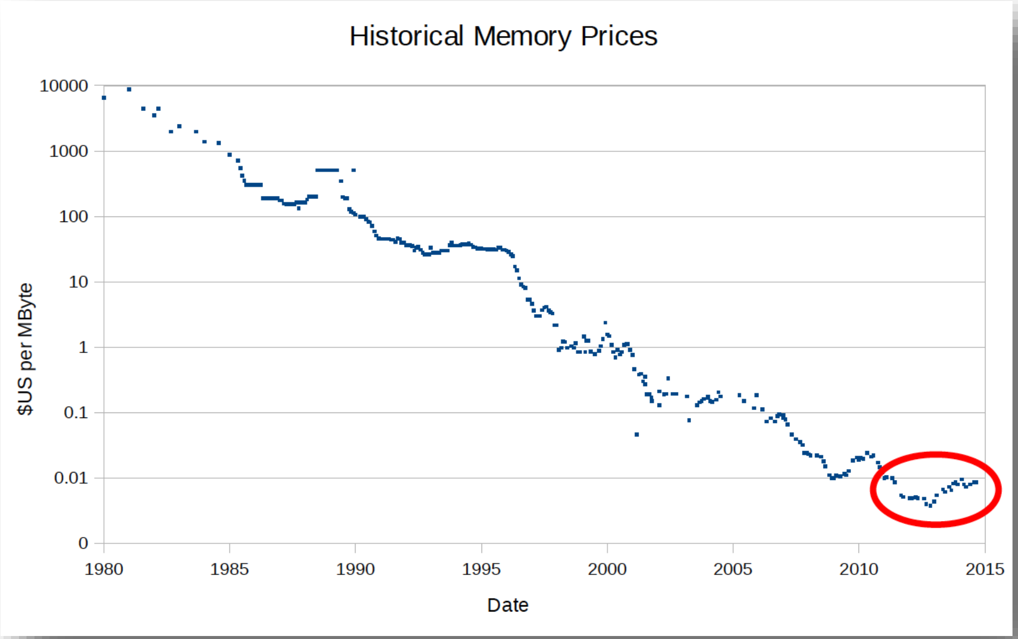


(some) ATLAS Computing Challenges



- **Gaudi/Athena developed in an era of increasing computing performance**
  - **This is no longer the case: clock speed stalled a decade ago due to thermal power density limitations**
- **Manufacturers have tried to compensate by increasing core counts and other features**

Memory usage



- **Clock speed is not the only issue: Athena reconstruction uses a lot of memory**
  - **>3 GB physical memory per job**
- **Memory prices are also in a plateau**
- **5M U$S for 200k grid cores**
- **In general high memory/CPU ratio**
  - **Much harder for Athena to operate**

Involvement from Argentina

- **Many new alternatives to "traditional" architectures**
  - Strong influence from mobile and low-power markets
- **Goal: adapt, leveraging these new architectures**
- **AthenaMT: multi-threaded, concurrent extension to Gaudi**
  - Data flow driven: algorithms declare their data dependencies
  - Multi-threaded: algorithms process events in their own thread
  - Pipelining: multiple algorithms and events can be executed simultaneously
  - Algorithm cloning: multiple instances of the same algorithm executed concurrently

University of Buenos Aires (Argentina) recieved a 30k U$S grant from CERN to train experts for Athena → AthenaMT migration

- Pretty much no changes for the user, LOTS of changes at the framework level
- Currently at CERN contributing to reconstruction migration to AthenaMT

Rucio: Distributed Data Management in ATLAS

La Plata Physics Institute (IFIP)
University of La Plata (UNLP)
Argentina

On behalf of Joaquín Bogado

Rucio Development

- Rucio is the Data Catalog of data in ATLAS
  - keeps track of where the files are in the Worldwide LHC Computing Grid (WLCG)
  - enforces data retention and replication policy

- Rucio was commissioned at the end of 2014 to cope with the higher data throughput of LHC Run 2.
- Rucio is open to the community since 2017
- Rucio ATLAS instance keeps track of 450+ PB of data worldwide

Rucio Development

- Argentinian involvement in ATLAS Rucio development and operation since 2014 (commissioning year of Rucio)
  - Feature development, documentation, operations
  - Automation and metrics instrumentation

- Ongoing, advanced PhD (Joaquin Bogado)
  - Research on Transfer Time prediccion, to provide Distributed Data Management system with more/better information
  - Improve the Data Transfer Scheduling

Dark Data Detection and Monitoring

- A new Rucio module was implemented to detect files in sites that are not in the catalog
- Around 10 PB of data was recovered immediately after the daemon was put in place

- Several efforts put since 2014 into:
  - Documentation, use cases for newcomers, automatic documentation from code
  - Bug fixing, new features across all software layers (200+ commits, 10K+ lines of code)

- Work in progress
  - Automatic site decomissioning (efficient handling of corner cases)

Particle Tracking Simulation

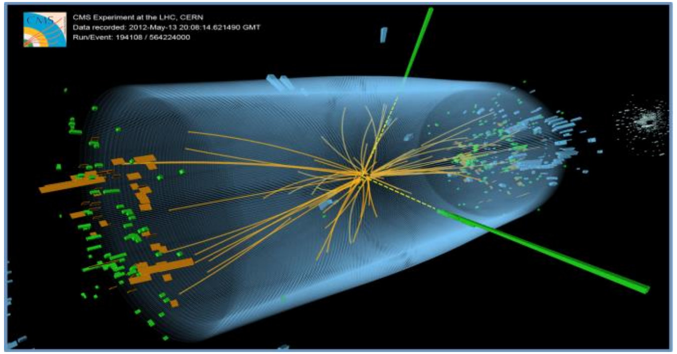
Computer Science Department (DC)
School of Exact Sciences (FCEN)
University of Buenos Aires (UBA)
Argentina

Rodrigo Castro
https://modsimu.exp.dc.uba.ar/sed/

Case Study: HEP+CMS

- **Tracking of particles** in physical systems
  - Well known simulation problem
  - Application in many domains
    (e.g. fluid dynamics, crowd systems, etc.)
- We focus on **High Energy Physics (HEP)**
  - Tracking of subatomic particles affected by physics processes within **complex detector geometries**

CMS particle detector at CERN

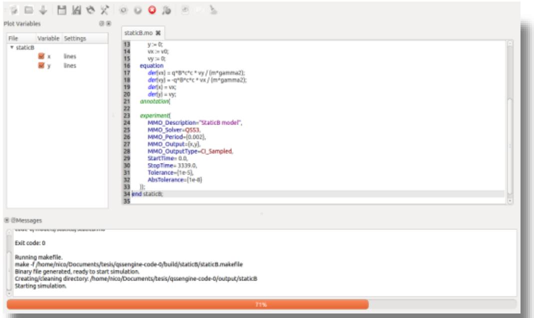


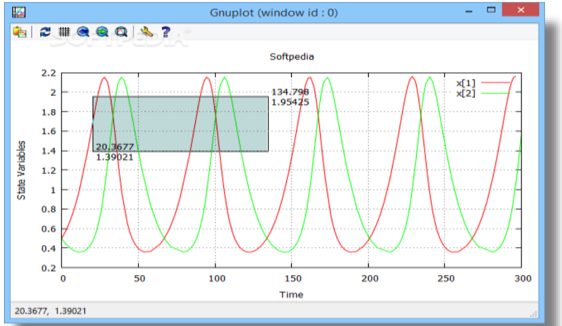
Proton-proton collisions in CMS

Introduction

- **Geant4** is the most widely used **simulation toolkit for HEP**
- **Key issue**: efficient handling of **boundary crossings** in **detector's geometry**
    - Solved by CPU costly **ad-hoc iterative algorithms**
    - **Can we do better ?**
- **Quantized State System (QSS) numerical methods**
  Modern family of **hybrid methods**
    - Attractive performance features for HEP problems
    - **QSS Solver** is a **general purpose simulation toolkit**
        - Features state-of-the-art QSS methods

QSS Solver
toolkit



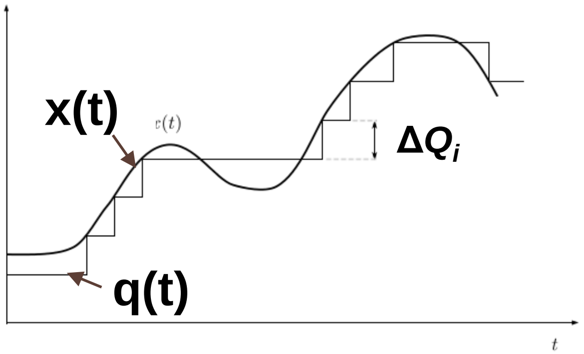


Quantized State System (QSS) Methods

More on QSS:
Ernesto Kofman's talk
Fri 22, 9am

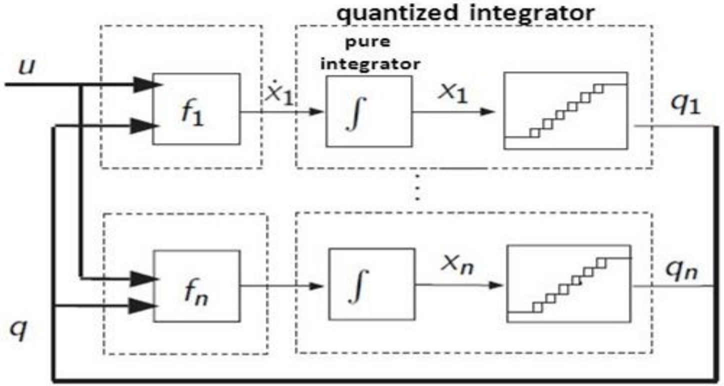
- Based on **state variables quantization**
- QSS methods discretize the system's **state variables** as opposed to traditional solvers which discretize the **time** (e.g. family of Runge-Kutta methods)
- **Continuous state variables** approximated by **quantized state variables**
  - A **quantization function** is in charge of controlling **error and accuracy** throughout the simulation

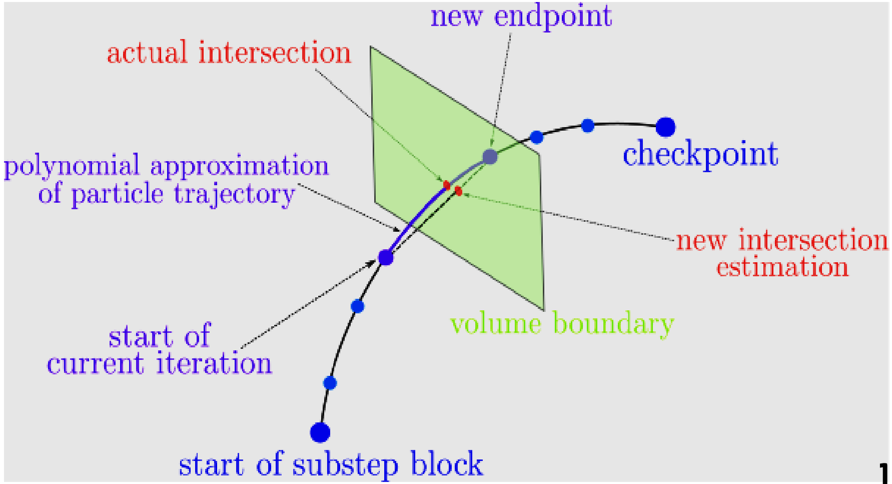
$$\underbrace{\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(\mathbf{t}))}_{\text{ODE system}} \quad \Rightarrow \quad \underbrace{\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{q}(\mathbf{t}))}_{\text{ODE quantized system}}$$



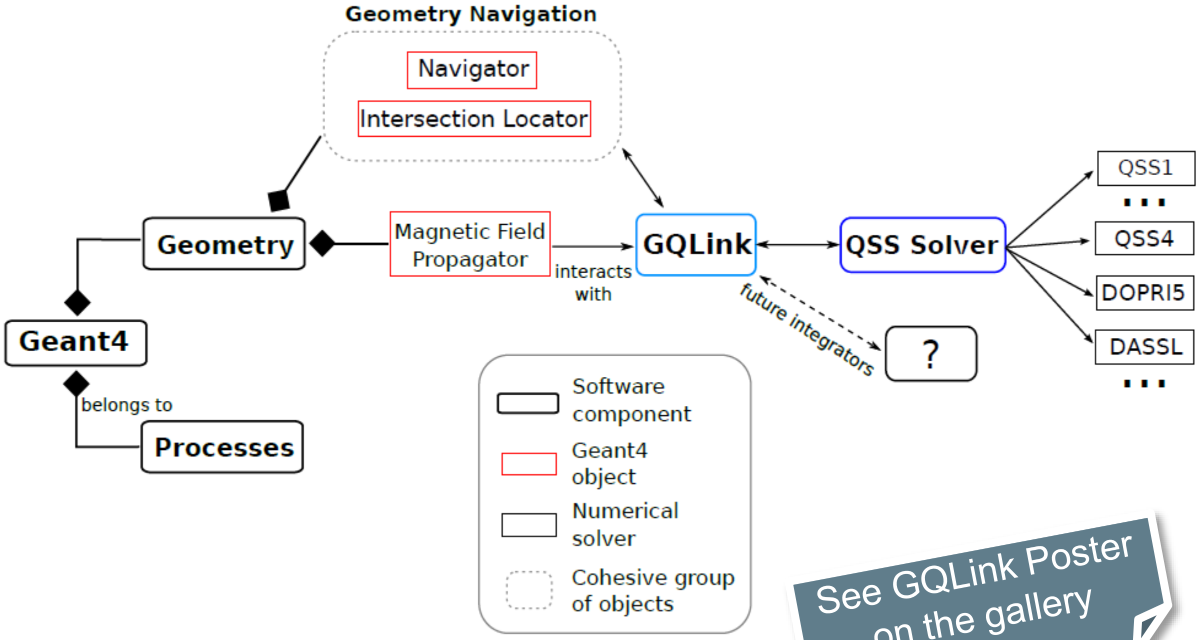
Attractive QSS features for HEP problems

- **Inherent asynchronicity**
  - Decoupled, independent computation of changes in states variables (no "global clock")
- **Dense trajectory output**
  - Supported by piecewise polynomial approximations of trajectories
- **Lightweight discontinuity handling**
  - Boundary crossings detected by finding roots of simple polynomials

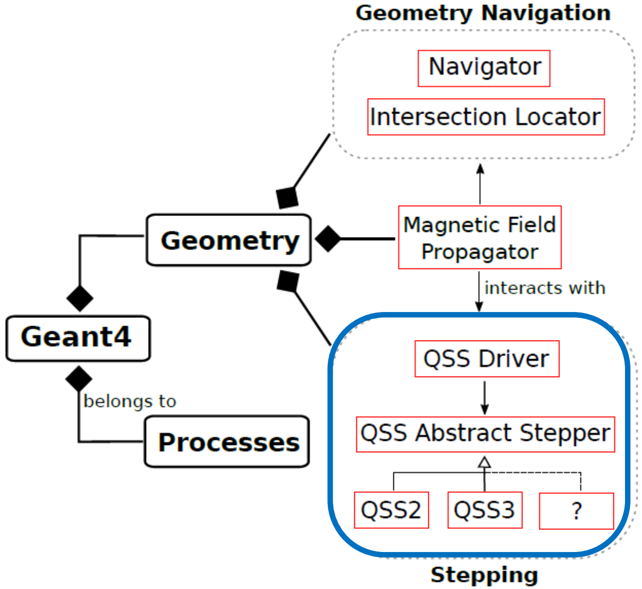




High Level Software structures and strategies



See GQLink Poster on the gallery



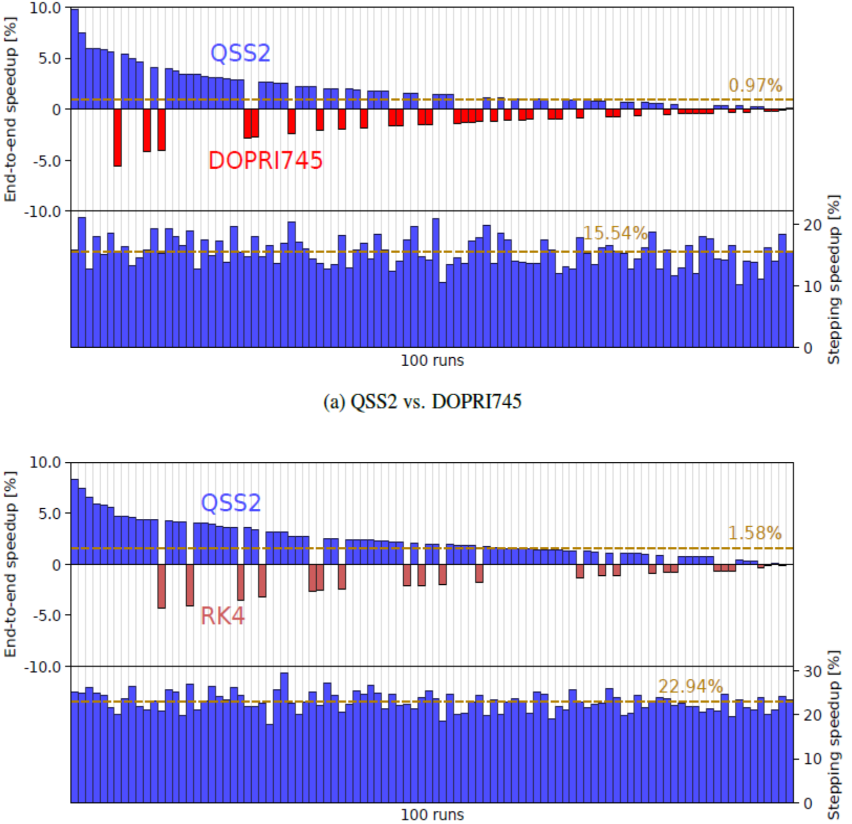
Strategy 1: Co-Simulation
- **GQLink** (Geant4 to **QSS Solver** Link)
- Toolkits preserve their resposibilities

Strategy 2: Embedded QSS
- **QSS Driver** for QSS
- New Native G4 Steppers

QSS speedups for a CMS Benchmark [1]

- Experiment:
  - 100 independent runs, 2000 particle gun events CMS full Run 1 geometry, single π^- particles
  - Physics list FTFP/BERT
- QSS2 vs. DOPRI745 (default since 2018)
  - 62 runs favorable for QSS; 38 for Geant4
  - End to End speedup**: ~1% (max ~10%)**
  - Stepping speedup: **~15% (max ~20%)**
- QSS2 vs. RK4 (default until late 2017)
  - 77 runs favorable for QSS; 23 for Geant4
  - End to End speedup: **~1.5% (max ~8%)**
  - Stepping speedup: **~23% (max ~30%)**

- Past researchers:
  1 Masters
- Current researchers:
  1 PhD
  1 Masters



(a) QSS2 vs. DOPRI745

(b) QSS2 vs. RK4

QSS solvers for HEP applications

- Research in collaboration with the Detector Simulation Group in Fermilab
- Since 2015
- 1 completed Master's Thesis
- 1 PhD Thesis and 1 Master's Thesis ongoing (Defense in mid 2020)
- 5 published papers
- Succesfull case study for a HEP-Computer Science interdisciplinary collaboration
- Results are relevant and innovative both for the Physics community and for the Computer Science community

- Preliminary results presented to the Geant4 21st Collaboration Meeting, September 2016, Ferrara, Italy. (Lucio Santi)

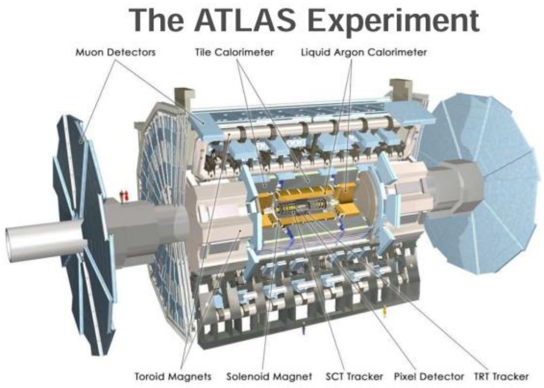
Network and Data Flow Simulation for the ATLAS DAQ system

Computer Science Department (DC)
School of Exact Sciences (FCEN)
University of Buenos Aires (UBA)
Argentina

Rodrigo Castro
https://modsimu.exp.dc.uba.ar/sed/

Data Network simulation - Why?

- Simulation supports: **network design, sizing, fine-tuning, failure prediction, etc.**
- System under Study: The **TDAQ system**, **ATLAS** Detector @ CERN

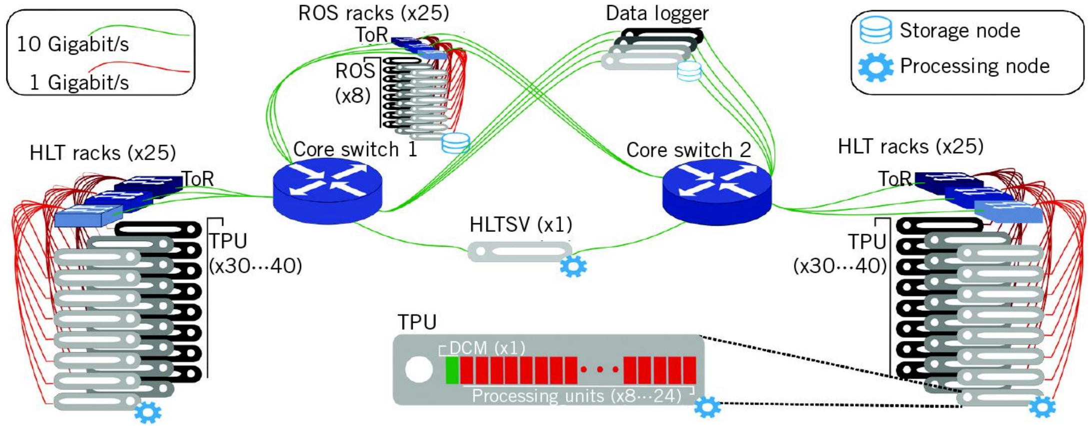


- Detector: 40Mhz Proton collisions generate ~**64 TB/s**
- Level1: Hardware filters down to **~160 GB/s**
- High Level (HLT): Software filters down to **~1.6 GB/s**

~2000 multicore servers

1-10 Gbps links

ATLAS HLT Network



More on ATLAS TDAQ:
Wainer Vandelli's talk
Thu 21, 3pm

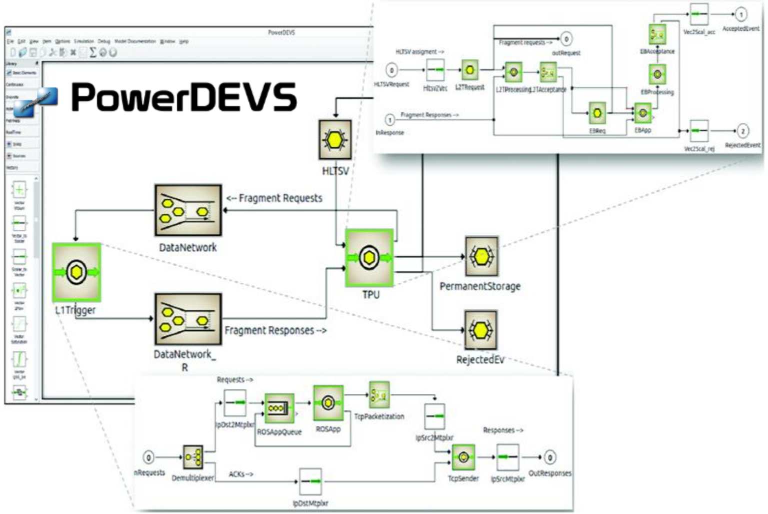
Large-scale Data Network simulation

- Packet-Level simulation performance is **proportional to #packets in the system**
  - Increases with #nodes, throughput and link bandwidth
- **Challenge**: very difficult for simulation of very large networks
- WallClock/Virtual simulation time **ratio: 240x**
- Example:
  - Simulate Phase I system
  - 1 rack (40 servers)
  - Virtual: 30s, **WallClock: 2 hs.**
- Full System: ~50 racks

Orientative TDAQ figures:

Phase I:	**Future Phase II:**	
Frequency: 100 KHz	**Frequency: 1 MHz**	**(10x)**
Event Size: 1.2 MB	**Event Size: 5.2 MB**	**(5x)**
Links: 10 Gbps	**Links: 100-200 Gbps**	**(10~20x)**

Packet-level TDAQ HLT Network Model



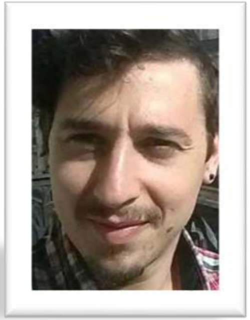
- **General purpose discrete event simulator**
- **New, reusable models tailored for TDAQ**

Some applications (since 2014)

- Characterization of event building latency for different hardware options
  (e.g. shallow and deep buffer switches)
- Optimal tuning of dataflow parameters
  (e.g. traffic shaping credits to avoid TCP incast effect)
- Evaluate alternative dataflow control algorithms and impacts on the system
  (e.g. new Join-Shortest-Queue load balancing strategy)
- Evaluate the use of advanced networking features
  (e.g. ECN, RED, etc.)
- Bottleneck analysis for proposed upgrade topologies
  (e.g. Felix network for Phase-I, using SDNs)

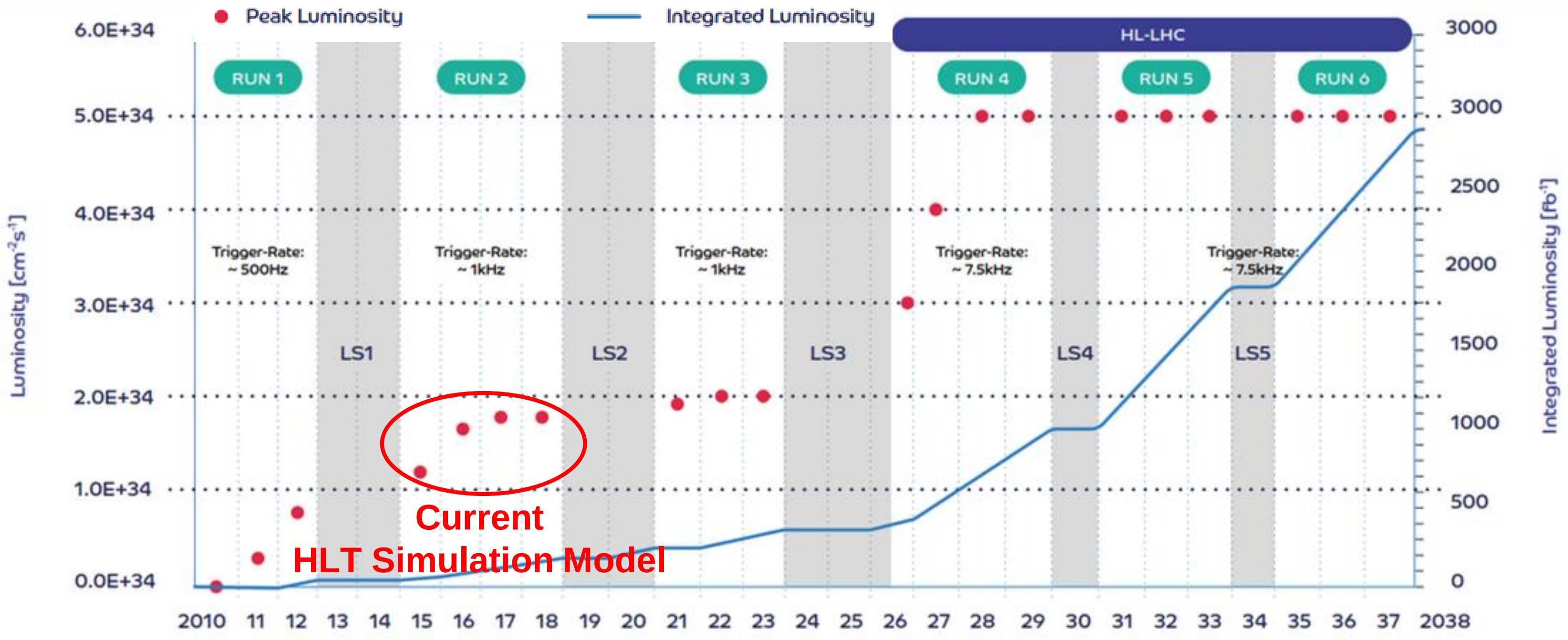
- Past researchers:
  3 PhD (1 fulltime)
  3 Masters
  (1 PhD Thesis)
- Current researchers:
  1 Postdoc
  1 Masters



Large-scale Data Network simulation

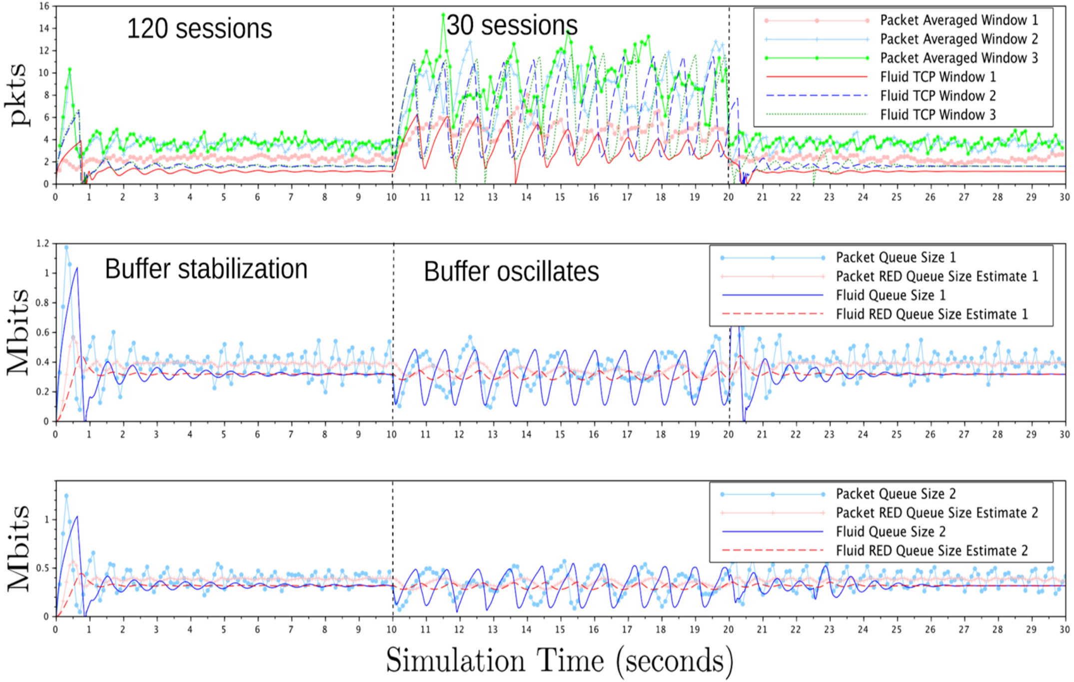
- The future promises more luminosity: more challenges for network simulation.



Research: Packet Level + Fluid Flow simulation

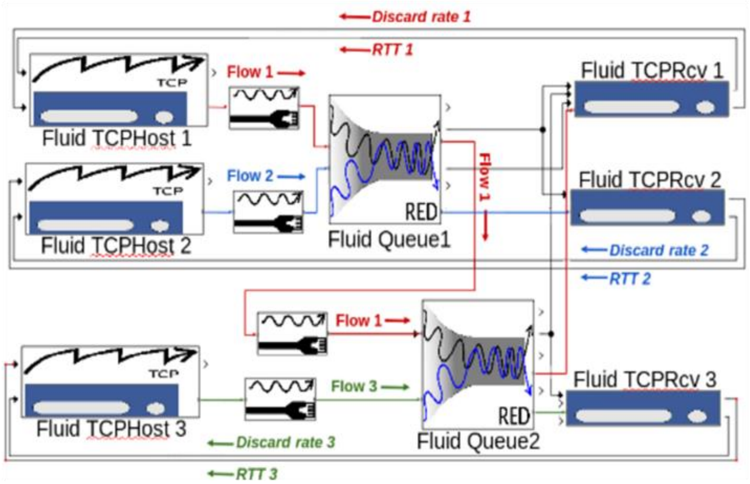
- Example: 2 Cascade queues, 120 TCP sessions, 3 flows

Example simulation results



Fluid Flow equations capture both transient and stable phases of discrete Packet Level traffic

Example topology



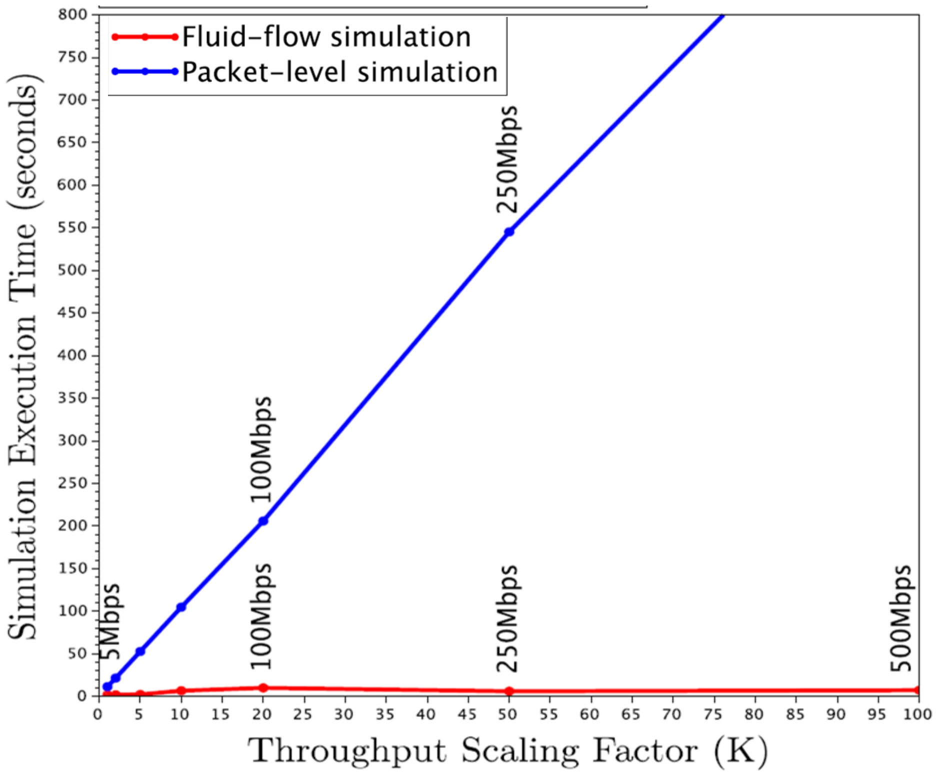
Combination of:

- Solving differential equations (flows)
- Handling discrete events (packets)

Lots of CS/Math research involved

Packet-Level vs. Fluid Flow Simulation Performance

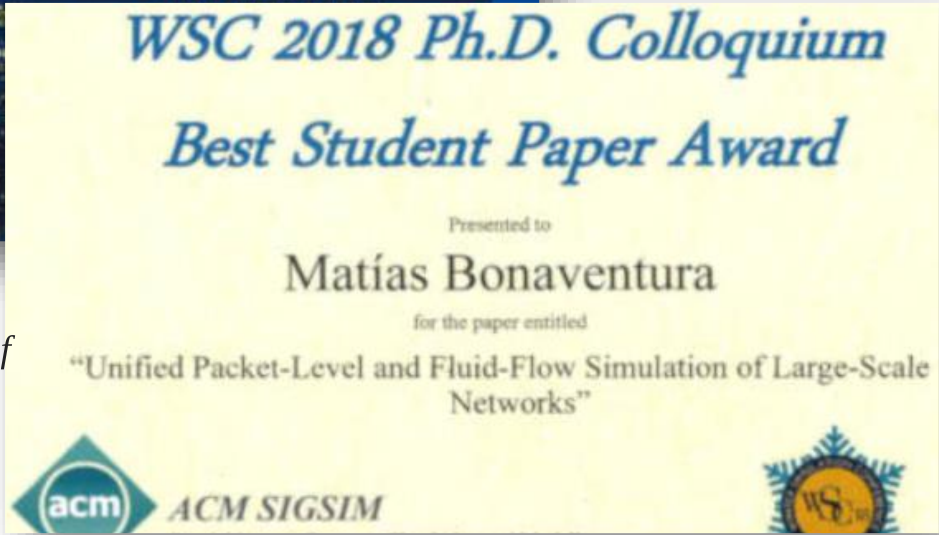
- Proportionally Increasing: link bandwidth, buffer capacity, RED thresholds and number of TCP sessions by a scaling factor K.



- Achieved speedups: up to 3 orders of magnitude

- Scaling with link speeds
  - Packet-Level: ~linear O(n)
  - Fluid-Flow: ~constant O(1)

Unified Packet-Level and Fluid Flow Simulation



Bonaventura, Matías. "Unified packet-level and fluid-flow simulation of large-scale networks." Proceedings of the 2018 Winter Simulation Conference. IEEE Press, 2018.

Fully cradled in ATLAS TDAQ, CERN.

Conclusions

- Collaboration between HEP experiments and CS labs has proven successful

- HEP as a provider of challenging use cases for CS research

- Complement a model based on good programmers for one-time projects with a model based on computer scientists (incl. good programmers) towards more sustainable, robust, and disruptive S&C platforms for HEP

- Challenge: Need to deal with different concurrent time scales (research deliverables vs. coding deliverables)