IMPLEMENTATION OF THE ATLAS TRIGGER WITHIN THE MULTI-THREADED ATHENA'MT FRAMEWORK

TENG JIAN KHOO
ON BEHALF OF THE ATLAS COLLABORATION
OVERVIEW

• Computing resources don’t scale with challenge of growing LHC luminosity — SW evolution necessary

• Large project to implement GaudiHive & AthenaMT framework upgrades
  • Framework elements common with LHCb and other experiments [Gaudi webpage]
  • Multithreading uses Intel Threading Building Blocks (TBB)

• Here cover adaptation of ATLAS software trigger
WHY MULTITHREADING?

• “Embarrassingly parallel” no longer enough
  - Processor speed plateau —> growth of multicore
  - Memory price floor —> need for memory sharing
• Run 2 approach: multiprocess execution
  - Fork workers after initialisation (or first event)
  - Share large static data structures
  - Slow memory growth unavoidable
  - Intra-event parallelisation limited
• Pathway to future computing architectures
  - E.g. GPUs or FPGAs for hardware acceleration
PARALLELISM

AthenaMT

Event 0
Thread 0
Thread 1
Thread 2
Thread 3
Event 1
Event 2
Event 3
Event 4

Inter-event
Intra-event
Sub-algorithm
Possible
PERFORMANCE SCALING

Close to ideal CPU scaling per core with minimal memory growth
Some bottlenecks still to be addressed

ATLAS Trigger Architecture

Event rates (peak):
- Level 1 Accept
- Calo/Muon
- Pixel/SCT
- Other
- ROD

DAQ:
- FE
- FTK
- Readout System

Data Flow:
- O(100)
- Data Logger
- ROD

Detector Readout:
- O(60 TB/s)
- ~ 160 GB/s

CERN Permanent Storage:
- ~ 1.5 GB/s

Regions of Interest:
- HLT
- HLTsv

Processing Unit:
- ~ 40k

Custom Hardware:
- 40 MHz

Full event:
- ~ 1.5 kHz

Fragments:
- ~ 40k

Data rates (primary physics):
- ~ 25 GB/s
ATLAS TRIGGER ARCHITECTURE

Data AcQuisition from ATLAS

- Level 1 (Hardware) Trigger
- High-Level (Software) Trigger
- Selected events to permanent storage

Event rates (peak):
- 40 MHz
- 100 kHz
- ~ 1.5 kHz

Regions of interest:

- Data Flow:
  - O(10)
  - ~ 160 GB/s
  - ~ 25 GB/s
  - ~ 1.5 GB/s

- Detector Readout:
  - FE
  - ROD

- Peak data rates (primary physics):
  - O(60 TB/s)

- Data Logger:
- CERN
- Permanent Storage
High-Level (Software) Trigger
HLT COMPUTING CONSTRAINTS

**L1 input rate**

- 40 MHz
- 100 kHz

**Event rates (peak)**

**Maximum rate to storage**

- ~1.5 kHz

**Average 400-500 ms / event processing time** — compare with O(30s) offline reconstruction

**Event size ~1MB**

**Peak data rates (primary physics)**

- O(60 TB/s)
- ~160 GB/s
- ~25 GB/s
- ~1.5 GB/s

**Regions of interest**

- Custom Hardware
- Level 1 Accept
- FTK
- O(100)
- Data Logger
- O(10)
- Data Flow
- Storage
- Full event
- Fragments

**Event size ~1MB**
Core HLT Elements

- Algorithmic code
  - Four-momentum reconstruction & Particle ID
  - Decision-making (“Hypothesis”)

- Regional reconstruction
  - Local detector readout where L1 trigger fired
    - “Regions of Interest” (all detector slices)

- Data flow & scheduling (“Control Flow”)
  - Reconstruct once, cache data for reuse
  - Early termination when event rejection is established

- Decisions recorded as event data
TRIGGER DECISION SCHEMATIC

- Calo Cells
- Tracker hits
- L1 RoI

Regional data access based on L1 seed (with exceptions)
TRIGGER DECISION SCHEMATIC

Staged reconstruction with early rejection
TRIGGER DECISION SCHEMATIC

Calo Cells → Fast e step → Precision e step

Tracker hits → Fast \( \tau \) step → Precision \( \tau \) step

L1 RoI

Reject

Data for local region unpacked only once, shared by multiple chains (caching)

Accept

Reject

Accept
TRIGGER DECISION SCHEMATIC

- Calo Cells
- Fast e step
- Precision e step
- Fast \( \tau \) step
- Precision \( \tau \) step
- L1 RoI
- Tracker hits
- Accept
- Reject

Data
Reco
Decision
TRIGGER DECISION SCHEMATIC

~1500 active chains in 2018
Constraints & Solutions

• Ensure code is thread-friendly
  • Const data access
  • State-free execution

• Asynchronous conditions access
  • Conditions object containers in data store
  • Intervals Of Validity mapped to data objects
  • Conditions Algorithms populate store for new IOV

• Removal of trigger-specific steering wrapper
  • Integration of Control Flow elements into framework allowing execution to be stopped early
Core HLT Elements in AthenaMT

• Algorithmic code
  • Shared where possible with offline domain

• Regional reconstruction
  • Views in Event Store — restrict geometric acceptance transparently to algorithmic code
  • Parallel reconstruction of multiple Regions of Interest

• Data flow & scheduling
  • GaudiHive graph-based scheduler
    • Declarative data access
    • Control Flow sequences

• Decisions recorded as event data
ATHENA-MT DATA ACCESS

Event Store “Whiteboard”

Data “collection”

L1 RoI “View”

Read

Write

Full-detector Reco

Hypo

Regional Reco A

Hypo

Regional Reco B

Hypo

Full-detector Reco

Hypo

Regional Reco A

Hypo

Regional Reco B

Hypo
AthenaMT Scheduling

Data in/outputs specified during initialisation. GaudiHive scheduler resolves execution graph before launching event loop.

Sub-algorithm components (tools) propagate data inputs/outputs to parent.
TRIGGER SCHEDULING

- Reco alg: prepare data
- Hypo: make decision on data
- Filter: gate execution based on hypo

**Step 1a**
- Pre-execute filters
  - Algs execute only once in entire graph
- Failed filter blocks exec

**Step 1b**
- Execute all steps (e.g. for combined chain)

- Execute all children in parallel, return logical OR
- Execute children sequentially, return early if fail

Time

Step 2, ...
MIGRATION STATUS

• Core infrastructure largely in place

• "Mechanical" migration of reconstruction elements in progress
  • Jointly with offline code

• $O(100)$ Run 2 chains in test trigger menu, some fraction of which fully implemented
SUMMARY

• Common Gaudi framework extended to support multithreading
• AthenaMT extensions permit ATLAS HLT operation for $O(500\text{ms})$ reconstruction and event filtering
  • Regional reconstruction with Event Views
  • Early rejection with Control Flow gate nodes
• Some Run 2 chains fully implemented
• Validation & performance assessments to come
REFERENCES

- AthenaMT: ATL-SOFT-PROC-2017-019
- GaudiHive/Avalanche scheduler: http://concurrency.web.cern.ch/GaudiHive
BACKUPS
PHYSICS PERFORMANCE

ATLAS Preliminary

Data 2017, $\sqrt{s} = 13$ TeV, 21.9 fb$^{-1}$

Optimum:
Online = offline

Offline selection:
$\geq 1$ jet with $|\eta| < 2.8$

HLT, $p_T > 450$ GeV

- 2016 calibration
- 2017 calib., calorimeter only
- 2017 calib., with tracks

Improvement from implementing offline corrections at cost of CPU

ATLAS-DAQ-PUB-2018-002
OVERVIEW OF ATHENA ARCHITECTURE