Processing Streaming Data In High Energy Physics Workflows

NATHAN TALLENT, DARREN KERBYSON, MAHANTESH HALAPPA NAVAR
MALACHI SCHRAM, KEVIN BARKER, LUIS DE LA TORRE, RYAN FRIESE, JIAN YIN
ERIC STEPHAN, KERSTIN KLEESE VAN DAM

Pacific Northwest National Lab

STREAM ’16 Workshop

March 22-23
High Energy Physics: Belle II Analysis Workflow

International effort to advance particle physics

Credit: Malachi Schram

KEK
High Energy Accelerator Research Organization
Belle II: Geographically Distributed Analytics

- Belle II Workflow: Extensive data analysis
- Data! 25 PB/year of raw data
  - Stored data expected to reach 350 PB
- Many analysis pipelines run concurrently
  - Normalize raw data
  - Physics analysis
  - Monte Carlo simulations
  - Data storage/archiving

Contention! Many independent data accesses in small window.
IPPD’s ‘Enhanced’ Belle II Workflow Execution

Challenge
- Mitigate contention
- Consider power/energy

1. Scheduler avoids I/O & network contention
2. Accurate task predictions enable good schedules
3. Optimized Data Transfer
4. Provenance feeds back performance data

IPPD: Integrated End-to-End Performance Prediction and Diagnosis
Hierarchical Scheduler Avoids I/O Contention

Approximate, two-level solution for NP-hard problem

**Challenge**
- Demand & supply vary considerably
- Hard to estimate task times
- Congestion dilates execution time

**1** Most *cost-efficient subset* of compute resources that meets the tasks’ demand
  - unit commitment (power grid)
  - mixed int/linear programming

**2** Best assignment of tasks to compute resources
  - bi-objective: energy & time
  - semi-matching: tasks $\leftrightarrow$ resources

**Figure:**
- Time (months)
- Energy (J)
- Provision exceeds demand by 3%
- Near-optimal solutions
- V: resources
To model each function, combine static and dynamic analysis

- Compose models of relevant functions
- Binary-based analysis
  - source code unavailable for key libraries, system I/O
- Enable “what if” studies
  - What if the memory system was 20% faster

Modeling Belle2 is hard
- Most functions latency bound (memory ops; dynamic dispatch)
- Costs widely dispersed (Profiling identifies key routines)

Analytical Modeling Predicts Task Execution Time
Prefetch data to reduce I/O blocking time

- Overlap remote data transfer and computation
- Retrieve only the needed part of a file
- Split data transfer across multiple internet connections
- Dynamically adjust given load on each connection
- Pace I/O request to improve end-to-end performance

Challenge: I/O Requests Create Blocking Time

Optimize Data Transfer via ‘Paced’ Prefetching

![Graph showing time vs. data size for 'No Prefetching' and 'Prefetching' scenarios. The 'Prefetching' line is significantly lower than the 'No Prefetching' line, indicating a 10x improvement.](image)
Provenance delivers execution statistics to scheduler & modeling.

Provenance (Provenance Environment) collects:
- Time series-based information for system/host
- Performance metrics for application/workflow

Predictive Analytics
IPPD’s ‘Enhanced’ Belle II Workflow Execution

Challenge
- Mitigate contention
- Consider power/energy

1. Scheduler avoids I/O & network contention
2. Accurate task predictions enable good schedules
3. Optimized Data Transfer
4. Provenance feeds back performance data

IPPD: Integrated End-to-End Performance Prediction and Diagnosis