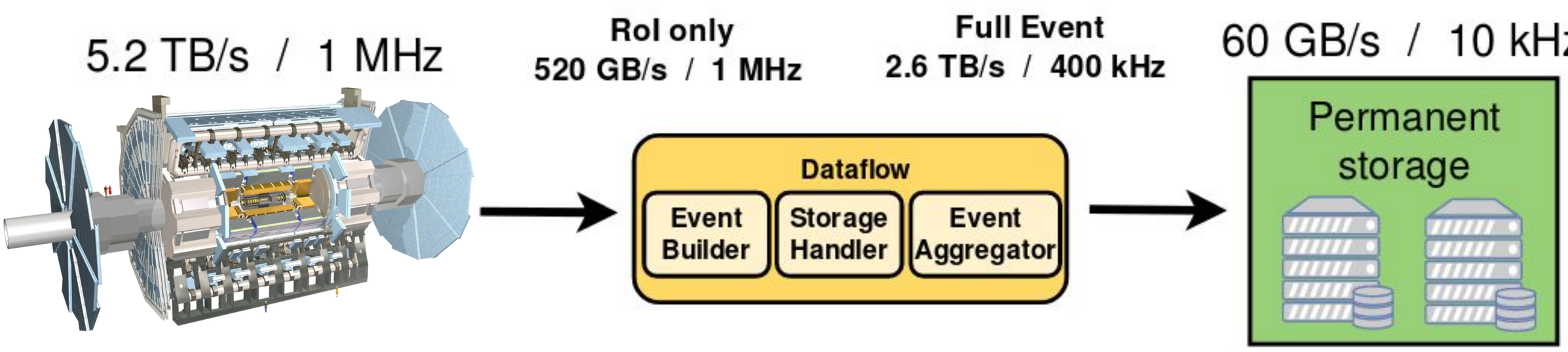
Performance evaluation of distributed file systems for the phase-II upgrade of the ATLAS experiment

Adam Abed Abud (Univ. Pavia & CERN)
Fabrice Le Goff (CERN), Giuseppe Avolio (CERN)

Phase-II Dataflow Bandwidth and Event Rates



Objectives

- Evaluate feasibility of using a commodity software to manage the large storage volume at the heart of the Phase-II Dataflow system
- Performance evaluation of **Distributed File Systems**
- Understand trade-off between computing power, network consumption and storage resources

Results

- Tested throughput vs IO block size
- Access patterns: seq writing, seq/random reading
- Tested throughput vs concurrency (parallelism)
- Effect of metadata traffic on the system
- **Setup**: 1 Gbit network, 3 storage nodes
- **Results**
  - Different behaviours
  - Optimize the solution to ATLAS use case. Minimize metadata traffic
  - Getting experience with deployment and configuration
- **Future:** Scale up number of nodes for large-scale testing

Dataflow

Event Builder
- Builds events provided by the Readout system
- Logical interface between Readout and Dataflow

Storage Handler
- O(45 PB) multi-host buffer
- Decouples Readout and Event Filter

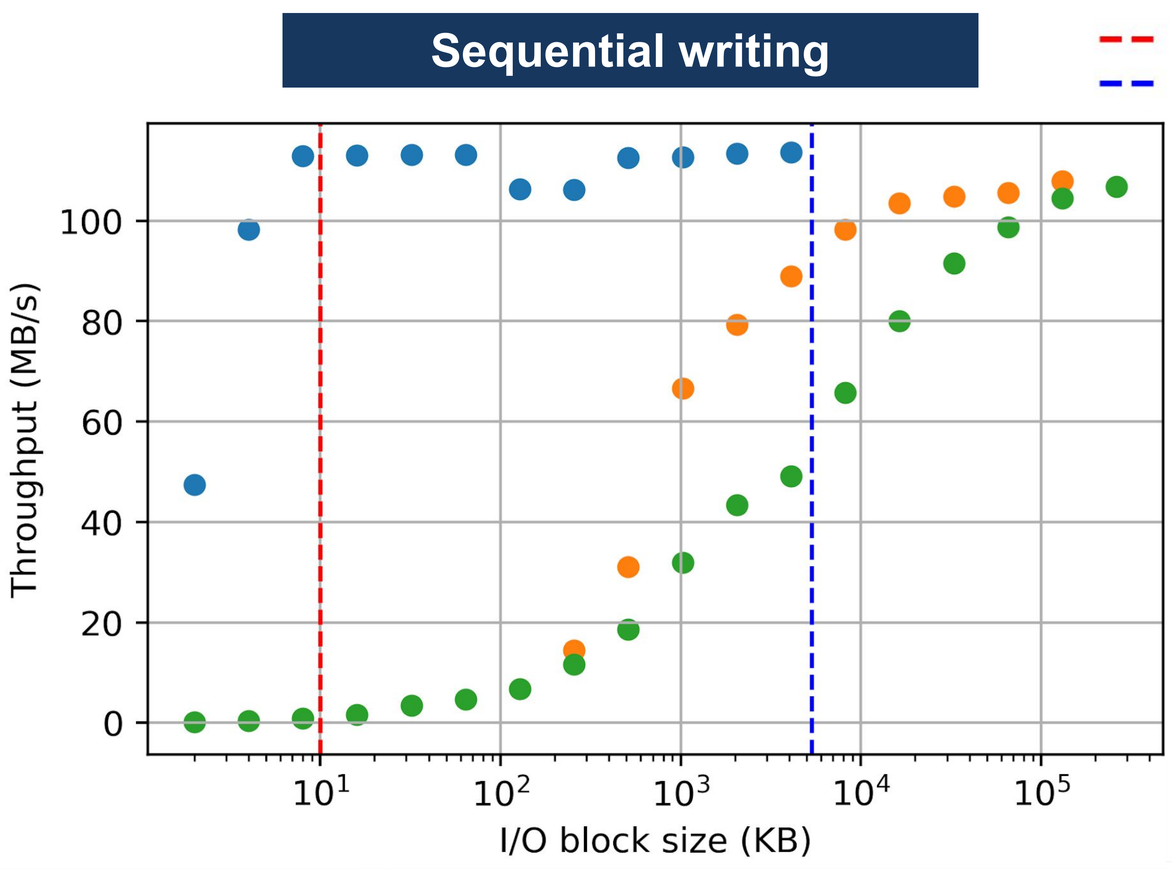
Event Aggregator
- Aggregates selected events
- Transfers to permanent storage

Distributed File System

- Definition: Data and file structure hosted across multiple storage nodes
- Possible Dataflow implementation (Storage Handler)
- **Advantages over other software-defined storage solutions**
  - Third-party software
  - Backed by industry
  - Load and storage balancing
  - Self-healing
  - Topology aware

Performance comparison between GlusterFS, Hadoop and CephFS

Sequential writing



Sequential reading