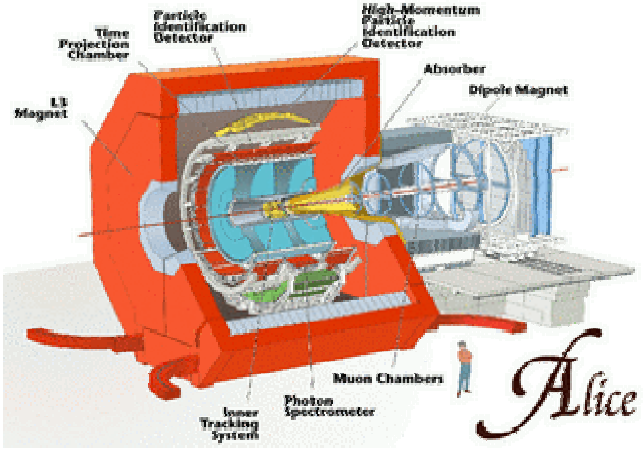
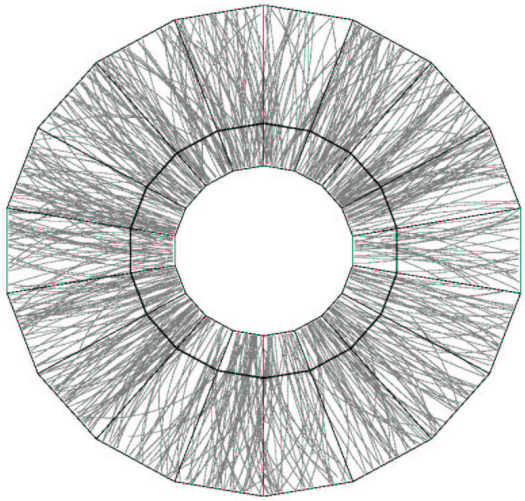
Timm Morten Steinbeck,
Computer Science / Computer Engineering
Kirchhoff Institute f. Physics, Ruprecht−Karls−University Heidelberg

A Software Data Transport Framework for Trigger Applications on Clusters

Requirements

- **Alice: A Large Ion Collider Experiment**
- **Very large multiplicity: >15.000 particles/event**
- **Full event size > 70 MB**
- **Data rate into last trigger stage (High Level Trigger, HLT) up to 25 GB/s**
- **HLT has full event data available**





T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

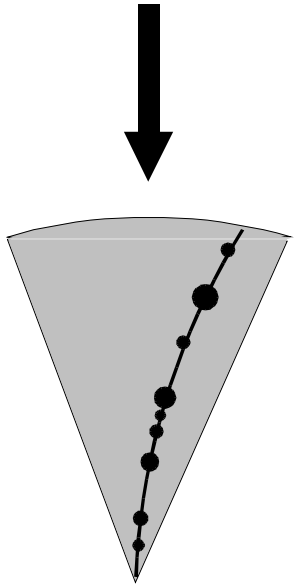
Requirements

HLT primary task is event reconstruction for triggering and storage

From raw ADC values...

1, 2, 123, 255, 100, 30, 5, 1, 4,
3, 2, 3, 4, 5, 3, 4, 60, 130, 30, 5,
.....

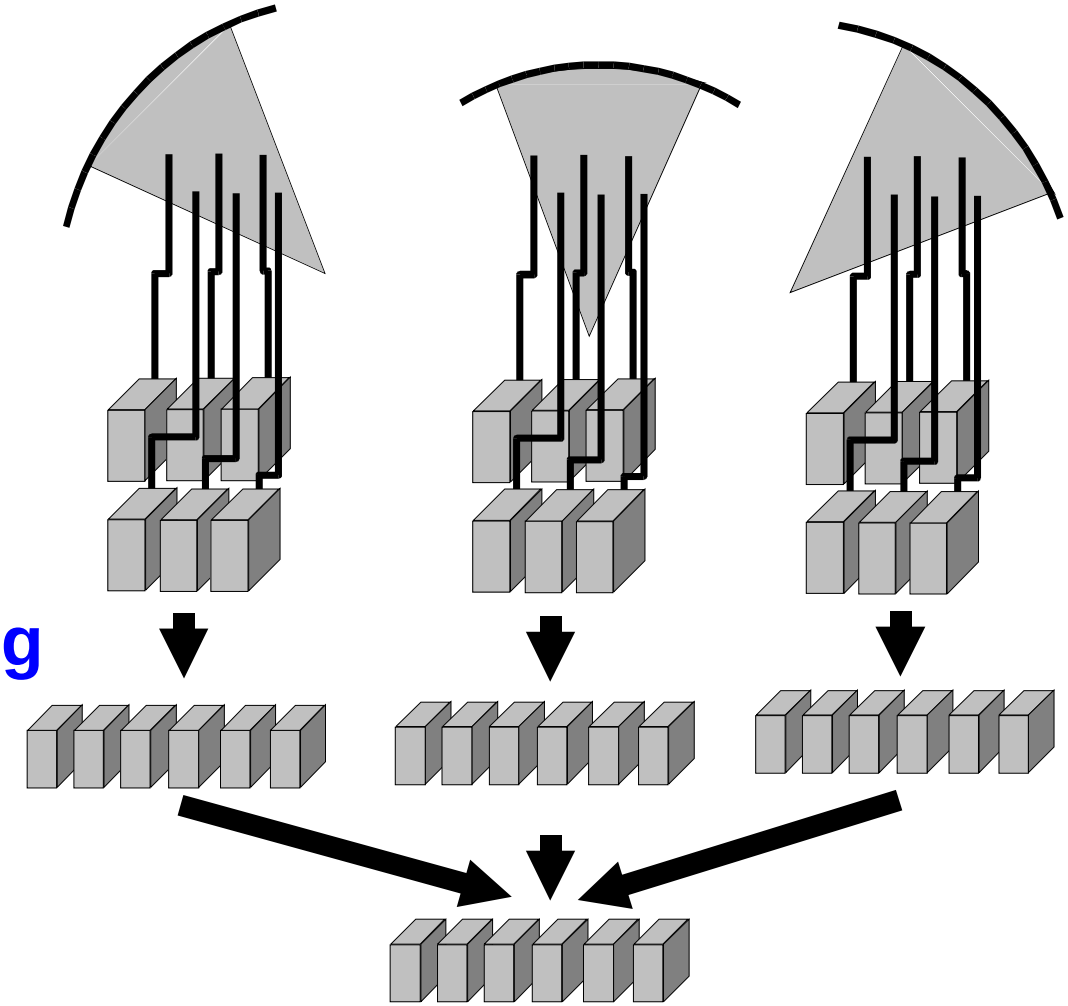
to



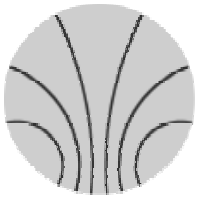
particle tracks

High Level Trigger Dataflow

- HLT consists of Linux PC farm w. roughly ≈1000 nodes
- Analysis is performed hierarchically
- Several stages for data processing and merging
- Natural mapping of dataflow, cluster topology, detector geometry



T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration



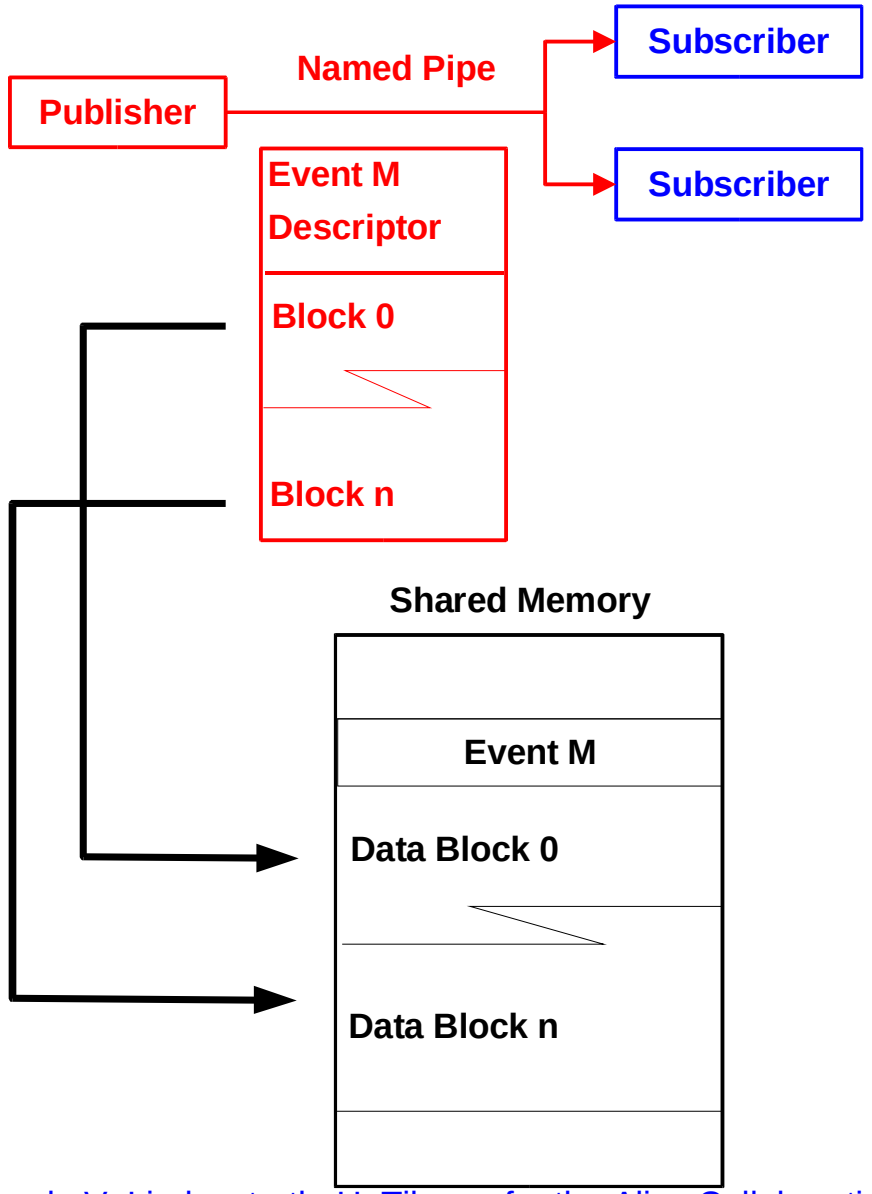
High Level Trigger

Framework software required to transport data in HLT

- **Flexible**
  - Components communicating via standardized interface
  - Pluggable components to support different configurations
  - Support for runtime reconfiguration
- **Efficiency**
  - Minimize CPU usage to retain cycles for data processing (primary)
  - Transport data as quickly as possible (secondary)
- **C++ Implementation**

Component Interface

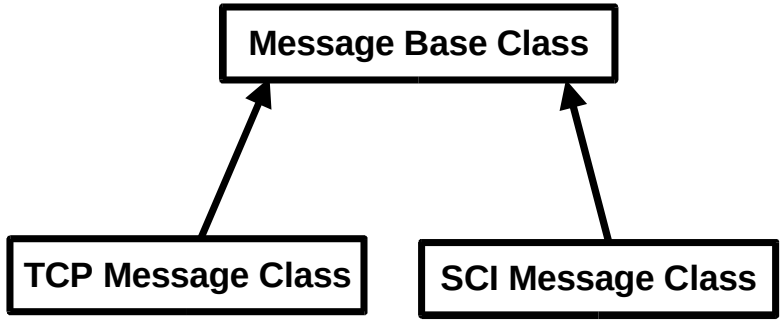
- Only locally on a node
- Uses shared memory for data and named pipes for descriptors
- Multiple consumers attached to one producer (Publisher−Subscriber paradigm)
- Buffer management has to be done in data producer

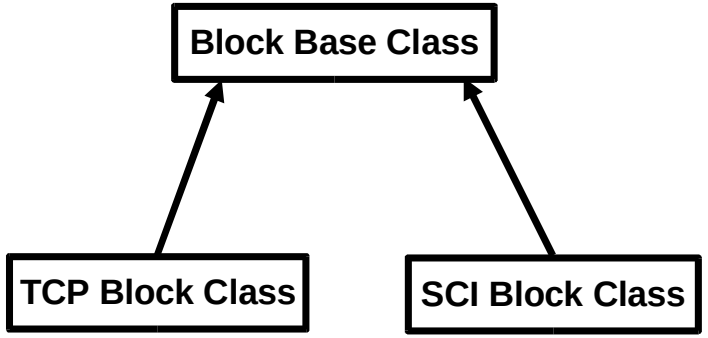


T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

Network Communication

- Uses class library
- Abstract call interface
- Classes optimized for
  - Small message transfers
  - Large data blocks
- Implementations for multiple network technologies/protocols possible
- Currently supported: TCP & SCI



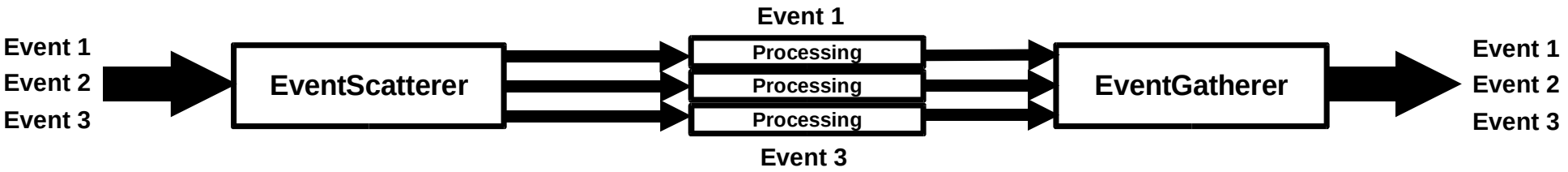


Components

Framework contains components to configure dataflow

- To merge data streams belonging to one event

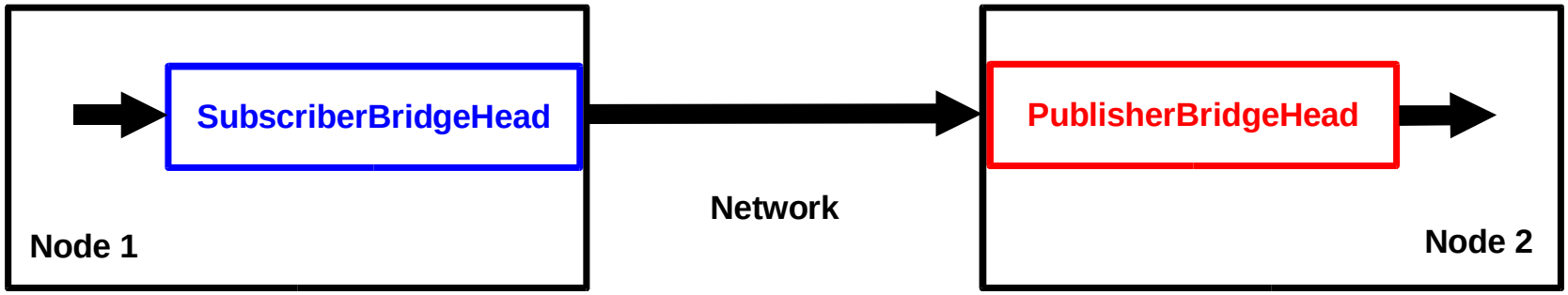
- To split and rejoin a data stream (e.g. for load balancing)



Components

Framework contains components to configure dataflow

- **To transparently transport data over the network to other computers (Bridge)**

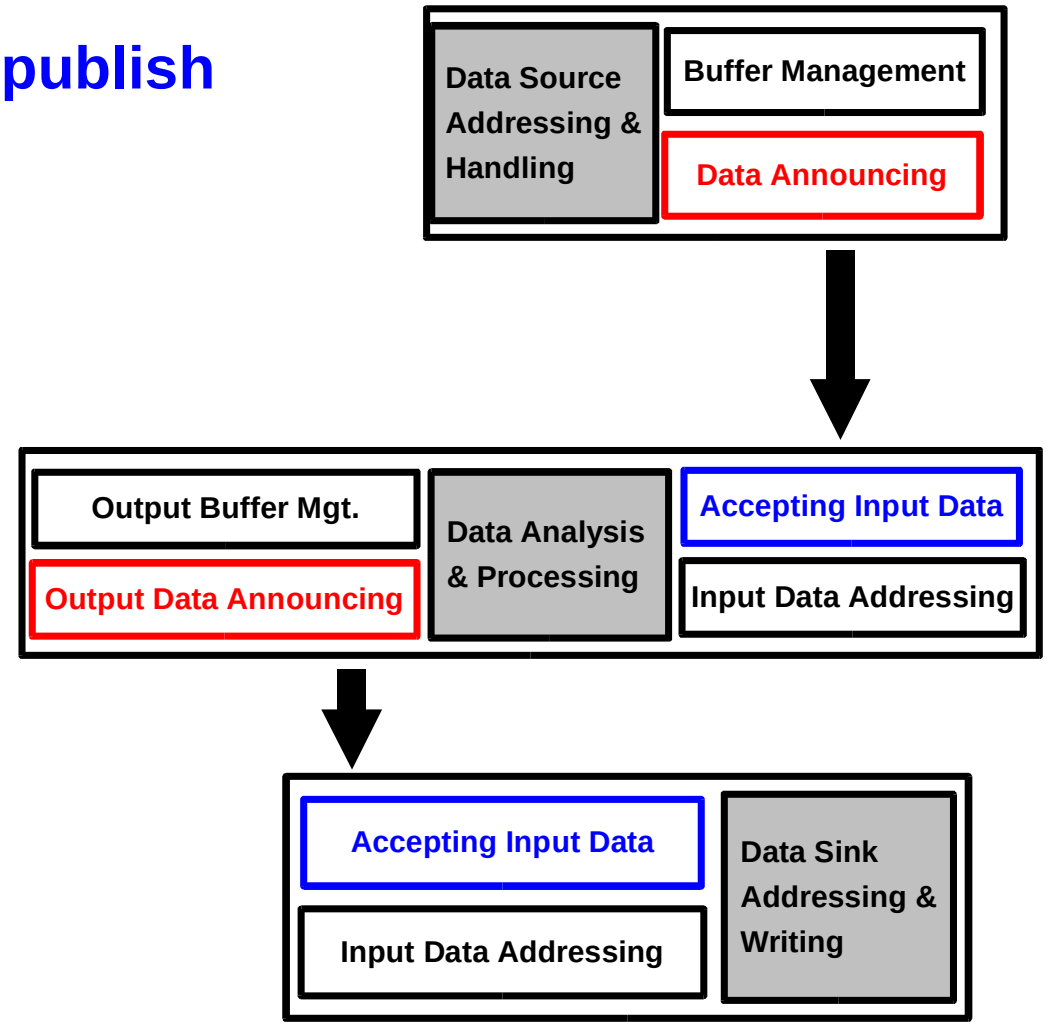


- **SubscriberBridgeHead has subscriber class for incoming data, PublisherBridgeHead uses publisher class to announce data**
- **Both use network classes for communication**

Components

Templates for user specific components

- **Read data from source and publish it (Data Source Template)**
- **Accept data, process it, publish results (Analysis Template)**
- **Accept data and process it, e.g. storing (Data Sink Template)**

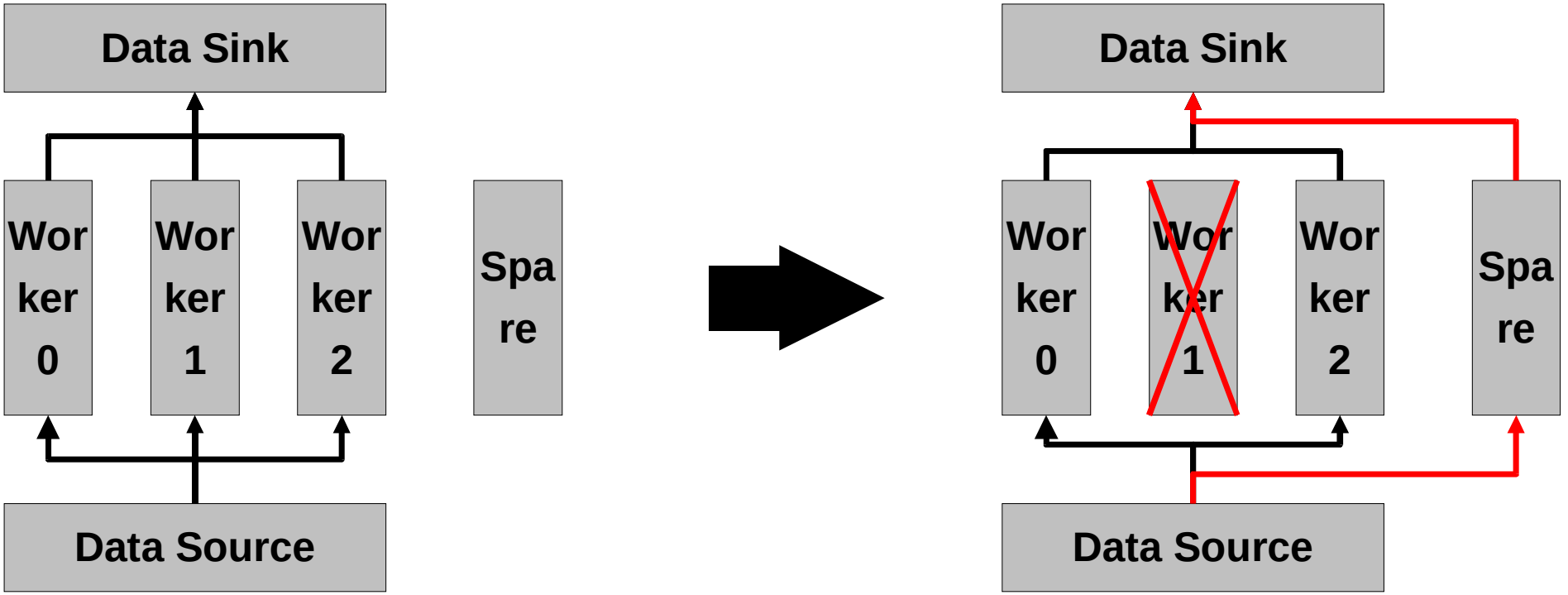


T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

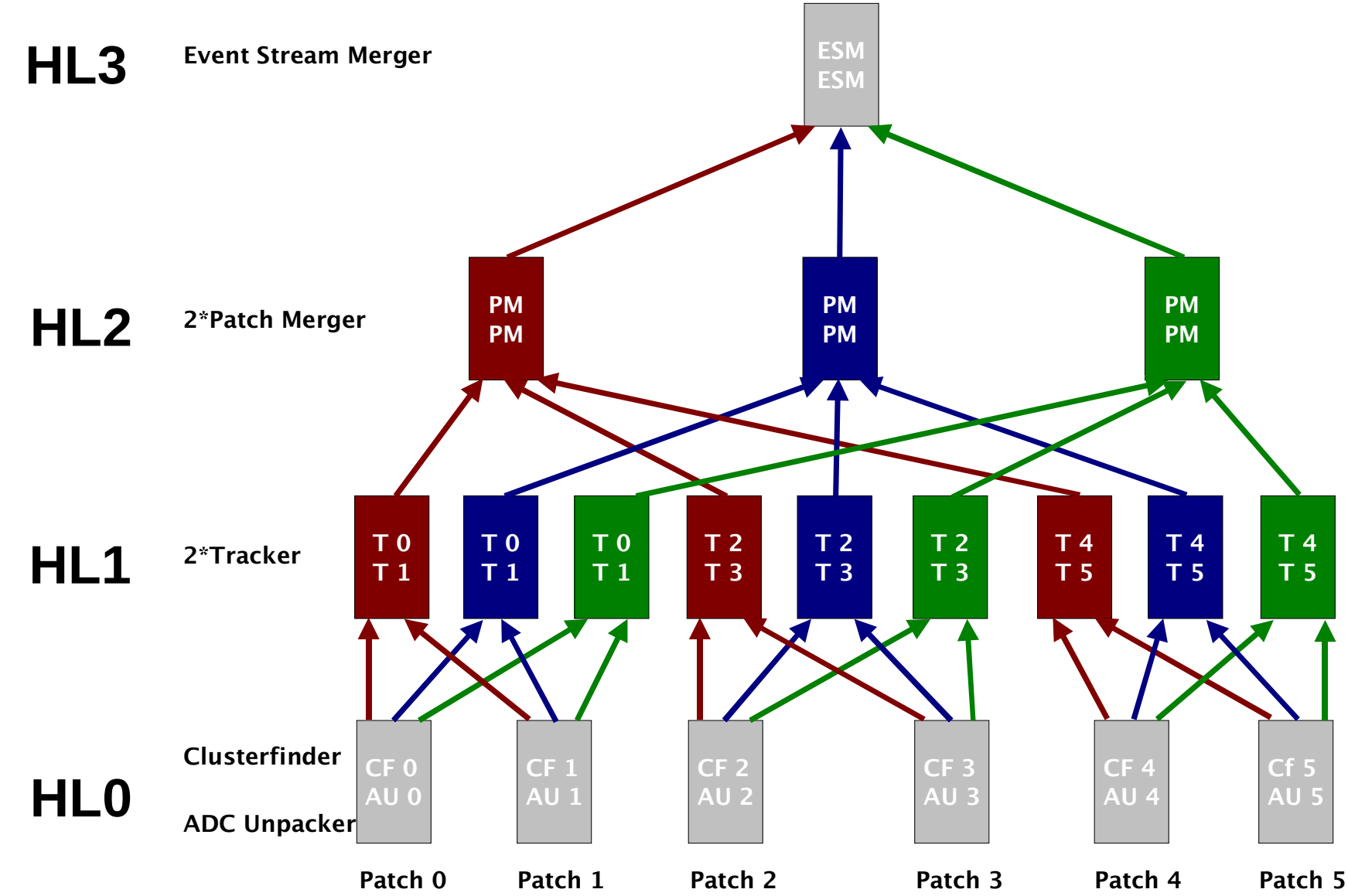
Fault Tolerance

Components to handle software or hardware faults

- Processing distributed for load balancing + redundancy
- Upon failure reschedule events and activate spare node



„Real-World-Test“



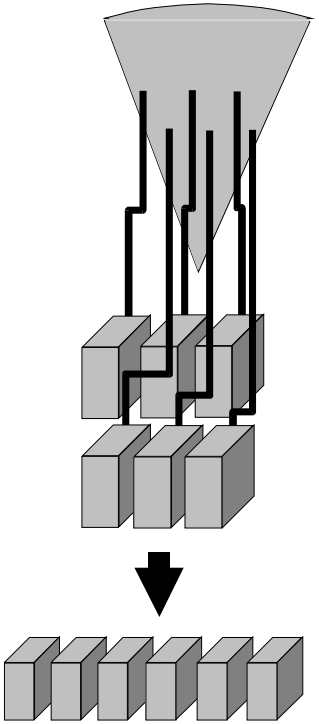
T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

„Real−World−Test“

Task:

- **Tracking of simulated Alice pp events,**
- **Pile up of 25 events**
- **Simulate one sector of Alice TPC**

  (1/36 of detector)

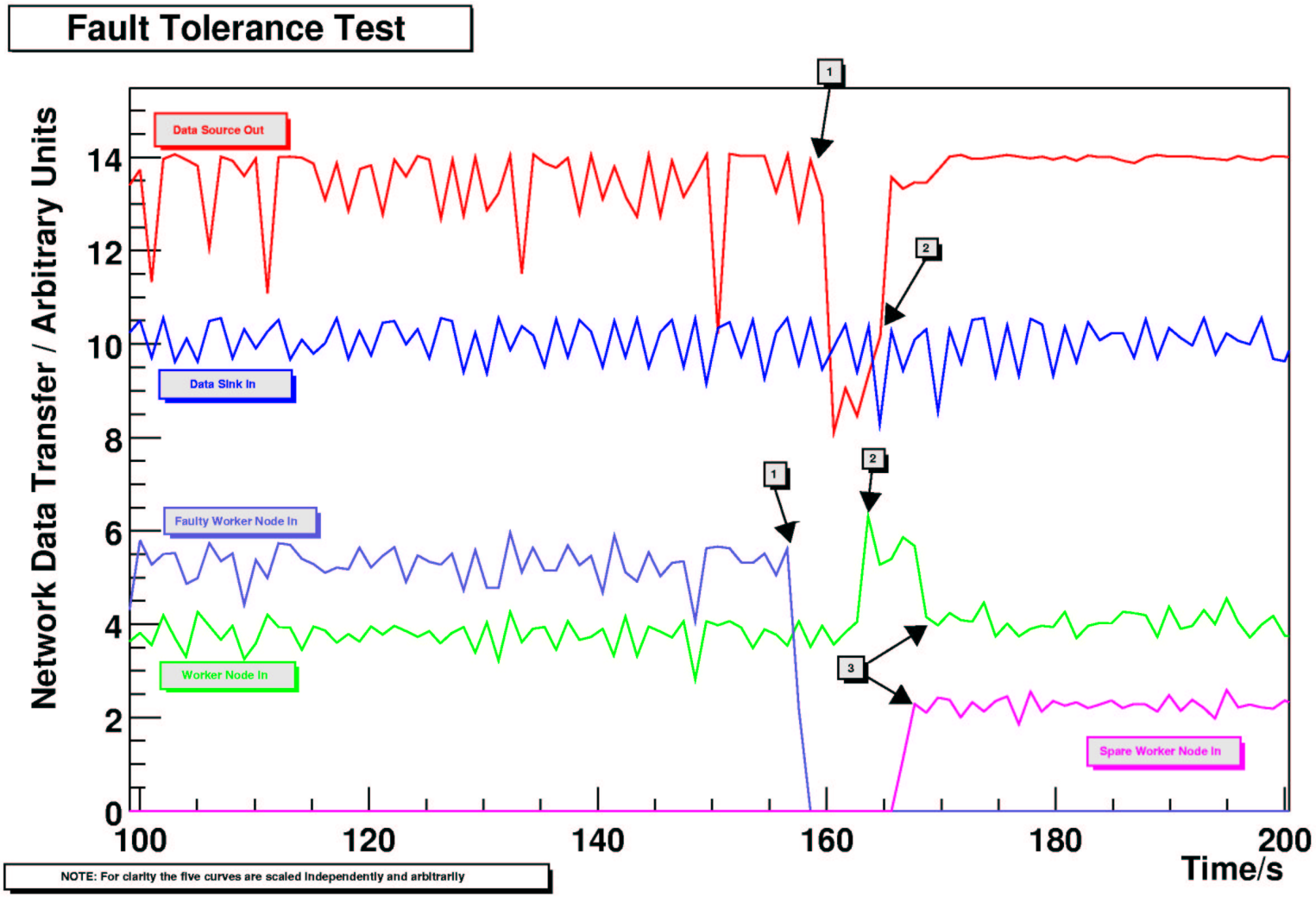


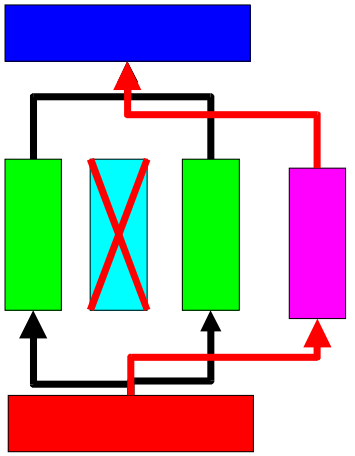
Performance on mix of 800 MHz and 733 MHz systems:

Processing rate of more than 420 Hz

T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

Fault Tolerance Test



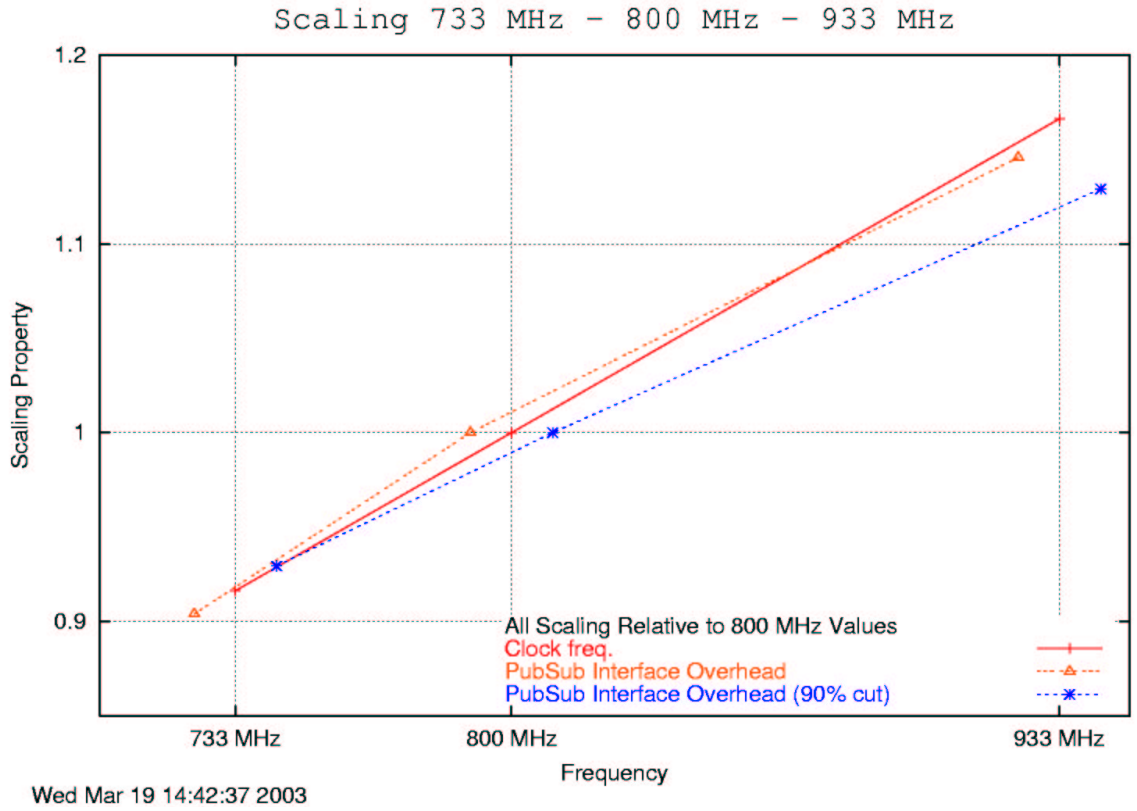


Curves are scaled independantly and arbitrarily

T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

Interface Performance

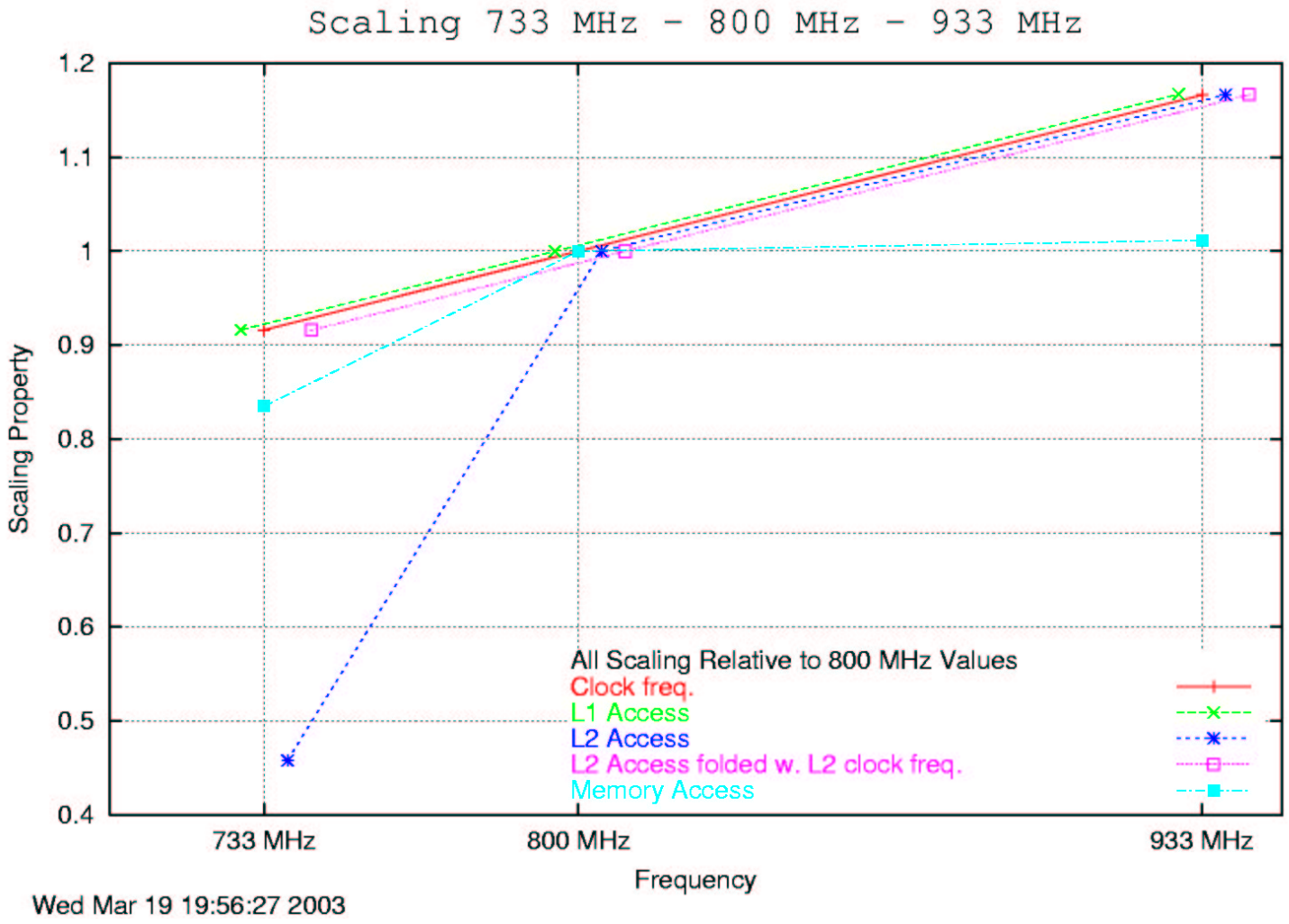
	733 MHz PC	800 MHz PC	933 MHz PC
Average Event Rate [kHz]	11.86	12.73	14.41
Average Time Overhead [μs/event]	168.7	157.1	138.8



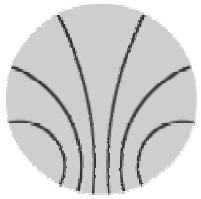
T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration

Interface Performance

Reference PC memory benchmark scaling



T. M. Steinbeck, V. Lindenstruth, H. Tilsner, for the Alice Collaboration



Conclusion

- Working framework
- Flexible configuration w. fault tolerance abilities
- Can already be used in real applications

To Do:

- Tool for easier configuration and setup
- Fault tolerance control instance/decision unit
- Fine−grained fault tolerance and recovery
- More tuning