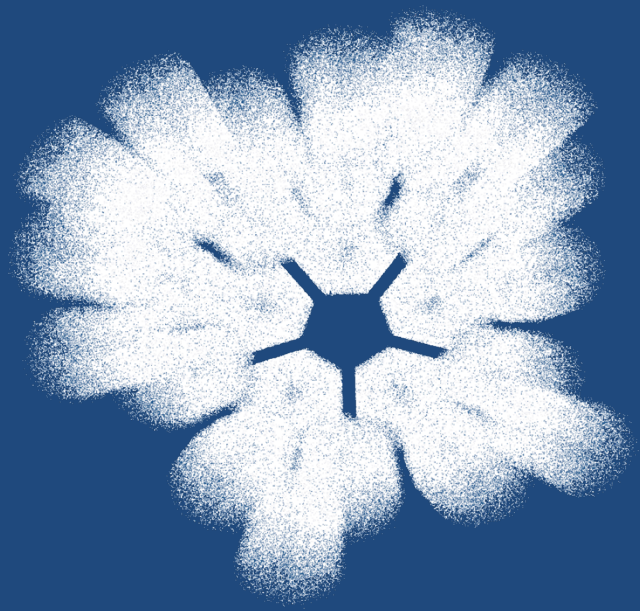
DAQ Overview for GRETA



Mario Cromaz

Nuclear Science Division
Lawrence Berkeley National Laboratory

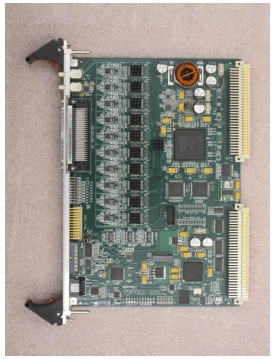
U.S. DEPARTMENT OF ENERGY | Office of Science

Outline

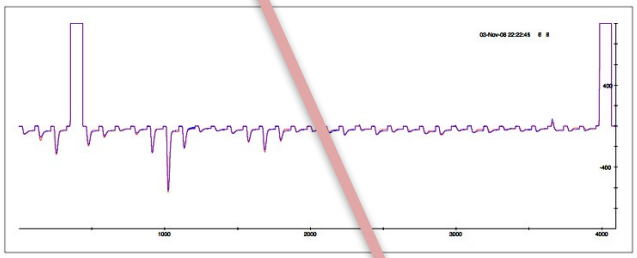
1. Overview of Online Data Processing
2. Requirements, Rates
3. Flow control, Containers
4. Components of HPC system
5. Slow controls, Monitors
6. Possibilities, Summary

High-level Data Processing in Tracking Arrays

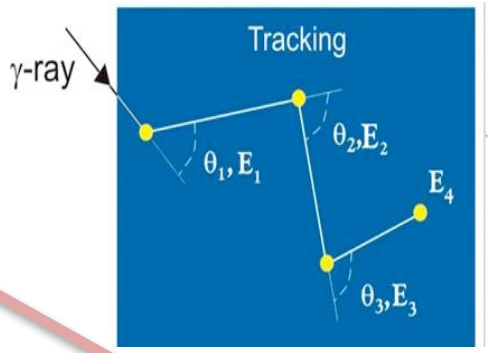
digitize segments, cc (36 + 1)



derive energy from trace



locate interaction points by fitting to crystal simulation



group/order interaction points by fitting to Compton scattering formula

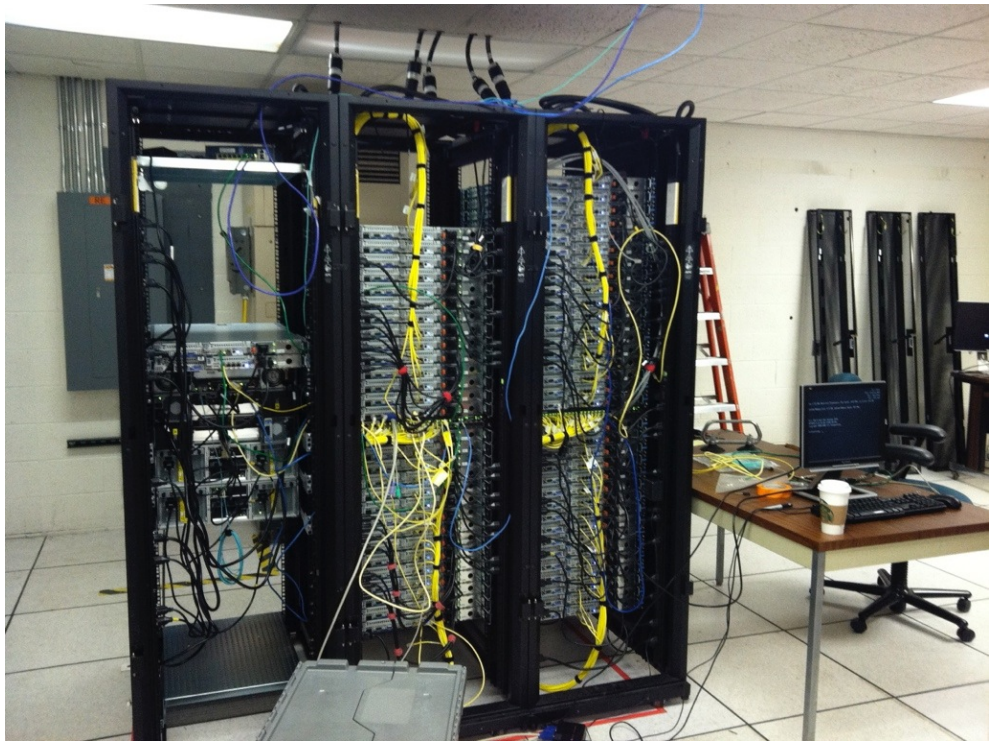
High-Level Computing Requirements

COMPUTING SYSTEMS REQUIREMENTS

- Support an incoming data rate of 32 MB/crystal/s, corresponding to a triggered data readout rate of 4000 γ-ray decompositions/crystal/s.
- Provide computing resources to perform 480,000 signal decomposition calculations per second.
- Support global event building and disk I/O at 500 MB/s.
- Provide local data storage of order 1 PB for experimental operations.
- Provide controls and monitors for approximately 50,000 channels with a nominal latency of 200 ms.

From GRETA Conceptual Design Report

Scaling Up GRETINA to GRETA



GRETINA cluster installation for first NSCL physics campaign

- **×3 crystals (x4 from orig. design)**
- **×4 readout rate / crystal**

- remove bandwidth limitations associated with VME readout bus
- higher performance global event builder, file store

lessons learned from GRETINA will be applied throughout the GRETA design

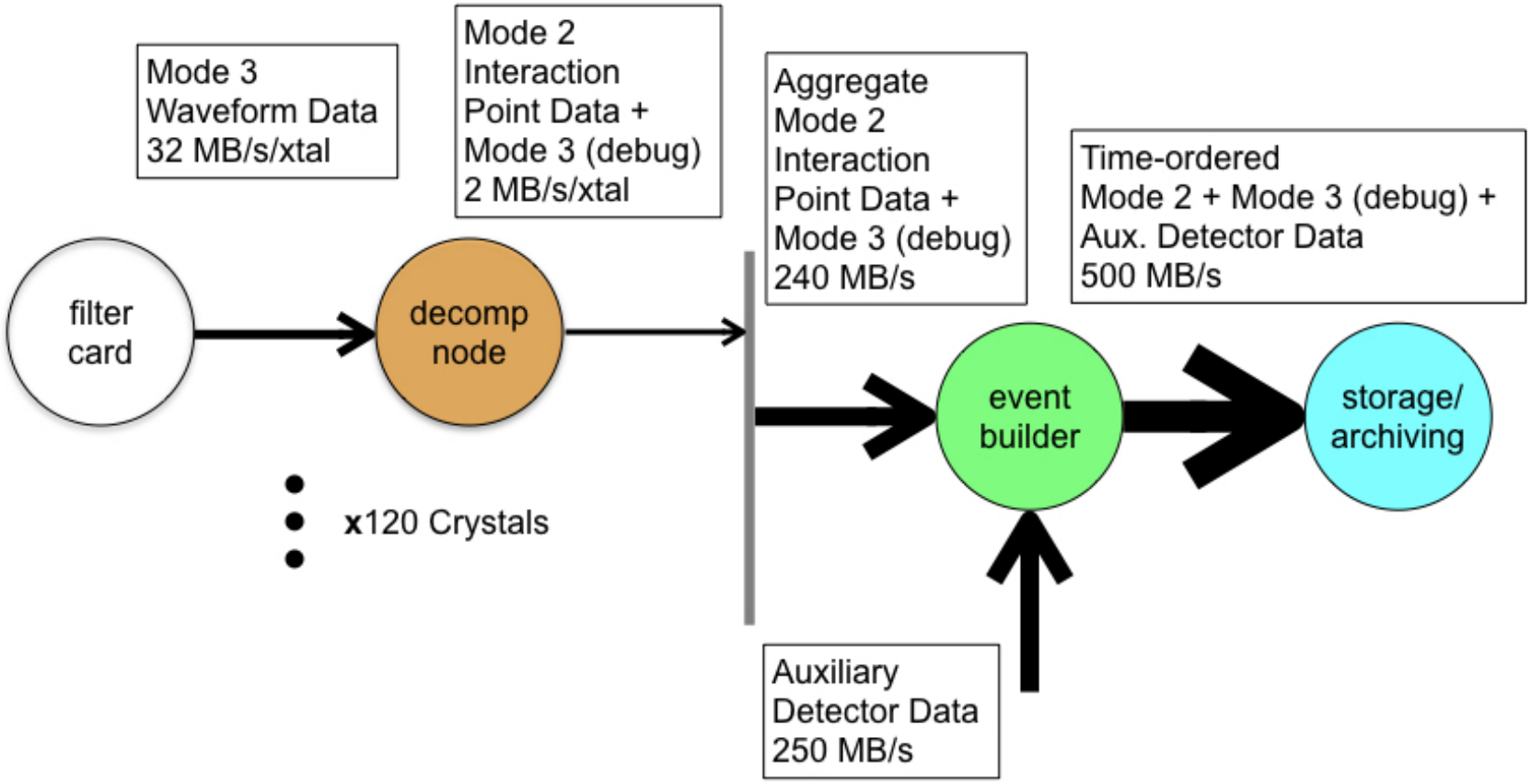
Signal Decomposition is Highly Parallel

- ***Each crystal is independent***
- ***Each event in a given crystal is independent and can be run in a different core/process/thread***
- Only requirements are:
  - events must be timestamped so that interaction point data can be recombined following decomposition (global event builder) for tracking/ physics analysis
  - any given thread must have access to the basis set for the crystal in question (~ 2GB)
- Simplifies system architecture greatly!

Decomposition Rate Sets Scale of GRETA Computing

- Signal decomposition is a online process - interaction point data required to monitor/debug experiments - ***Need to keep up!***
- It is the most computationally intensive aspect of array
- Cores required:
  - Number of crystals - **120**
  - Rate (/ crystal, post trigger) - **4 kHz**
  - Time / crystal event - **10 ms / core** (2016 Xeon)
  - Implies **4800 core cluster** (40 cores / crystal)

Data Flows (GRETINA-like architecture)

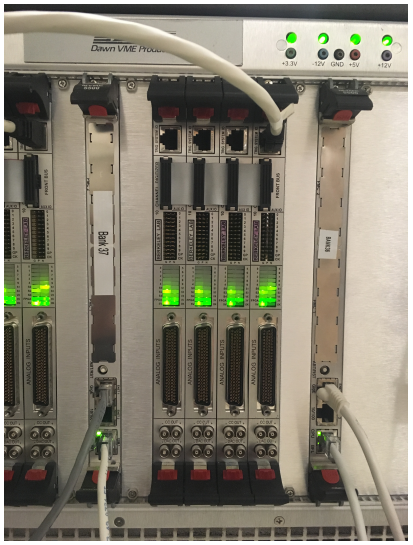


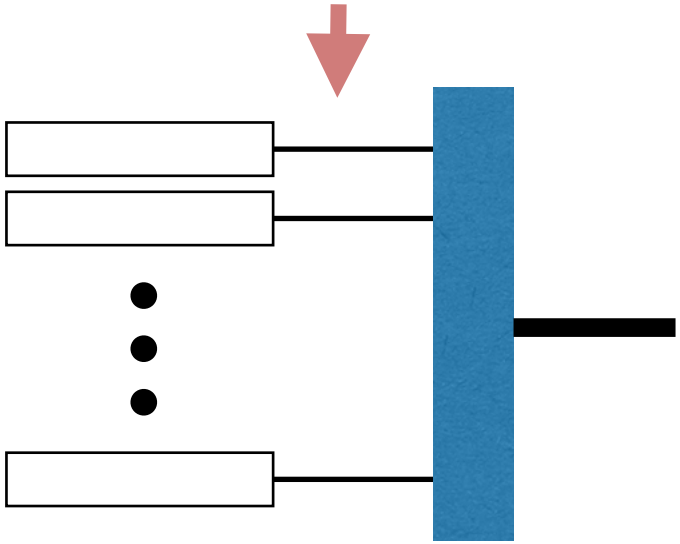
Data Rates In/Out of Cluster

- Input bandwidth:
  - event size - trace (mode 3) - (single crystal) - **8kB**
  - input bandwidth/crystal from filter boards - 4 kHz * 8kB **= 32 MB/s/crystal**
  - aggregate bandwidth (switch fabric, link between electronics room & cluster) - 32 MB/s * 120 = **3.8 GB/s**
- Output bandwidth:
  - event size - interaction points (mode 2) - **512 bytes**
  - output bandwidth - 512 bytes * 4 kHz = **2 MB/s/crystal**
  - aggregate output bandwidth (mode 2) = **240 MB/s**

Network-based Electronics → Architecture Changes

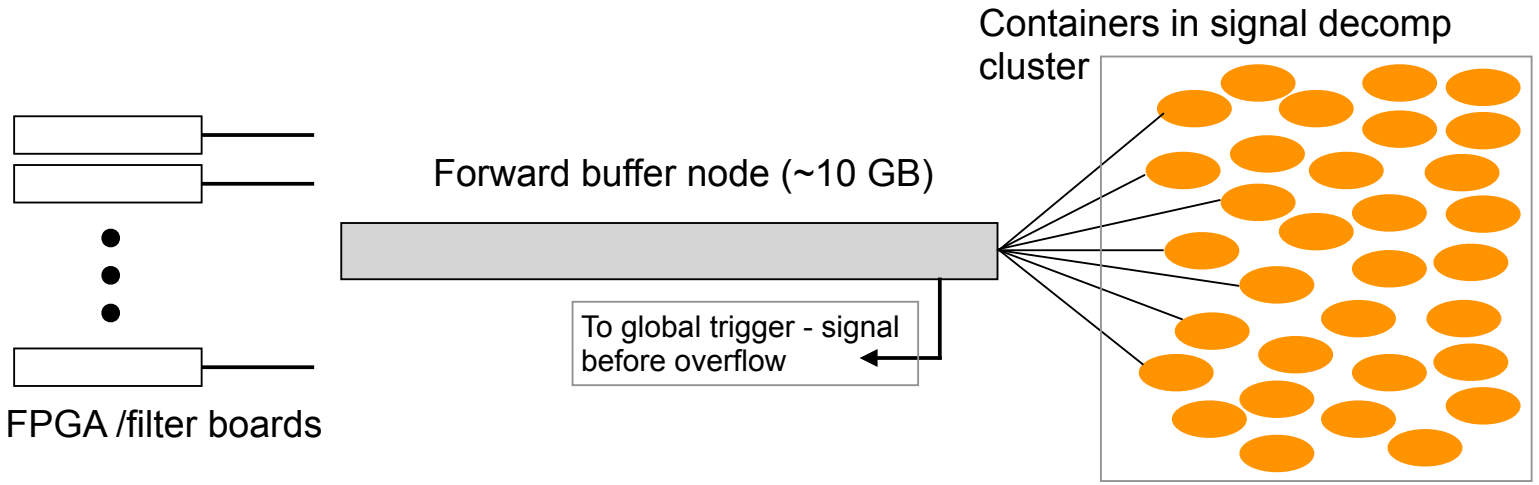
- GRETINA:
  - Bus-based readout
  - IOC (MVME5500) for data transfer and electronics controls
  - "Pull" model - signal decomposition nodes request data from IOCs as they become free
- GRETA:
  - Network-based ~~readout~~ "writeout" - on-board FPGA's send data via UDP
  - "Push" model - datagrams addressed by electronics, sent to 1000's of cores, 100's of processes
  - Load balancing??? Flow control???





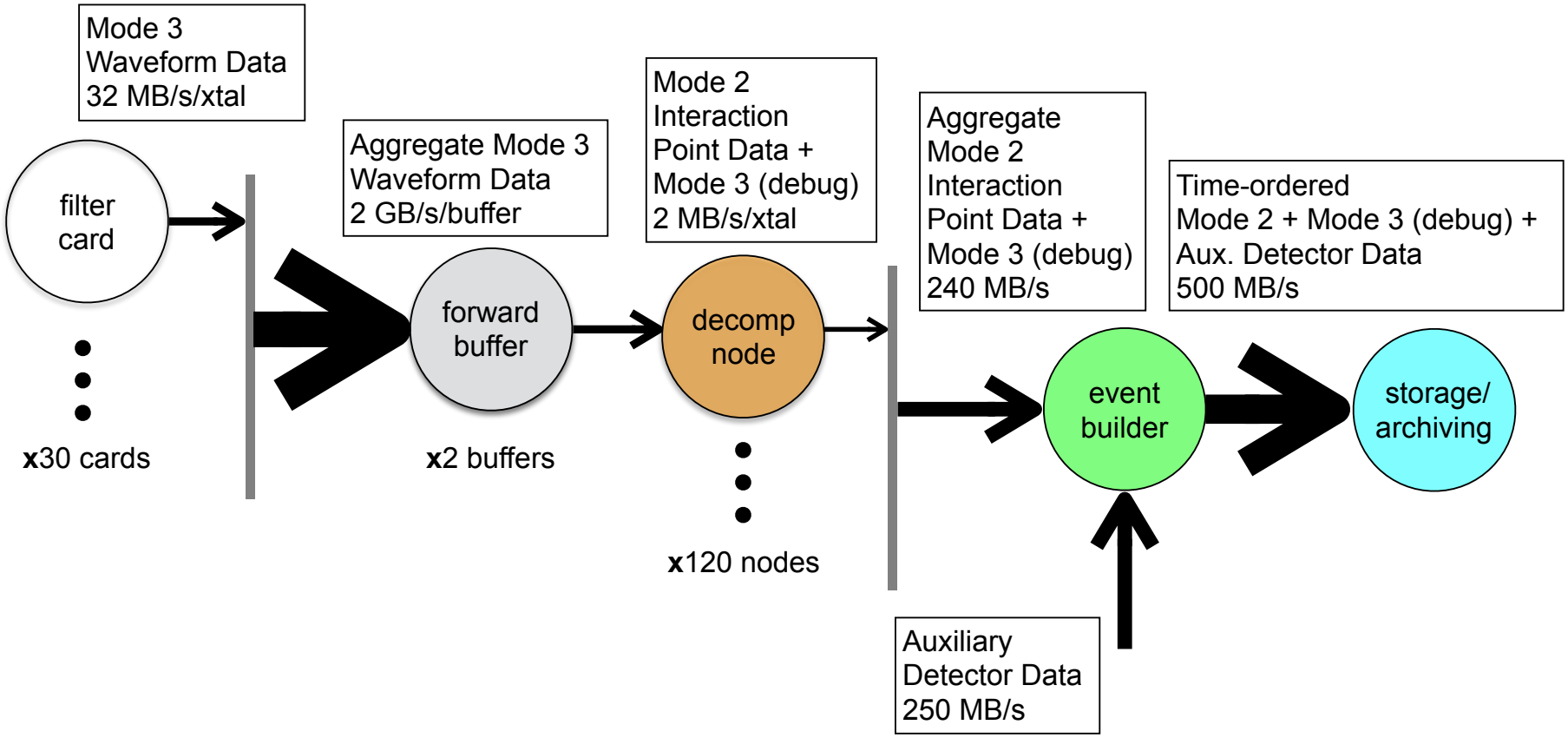
FPGA /filter boards

Forward buffers



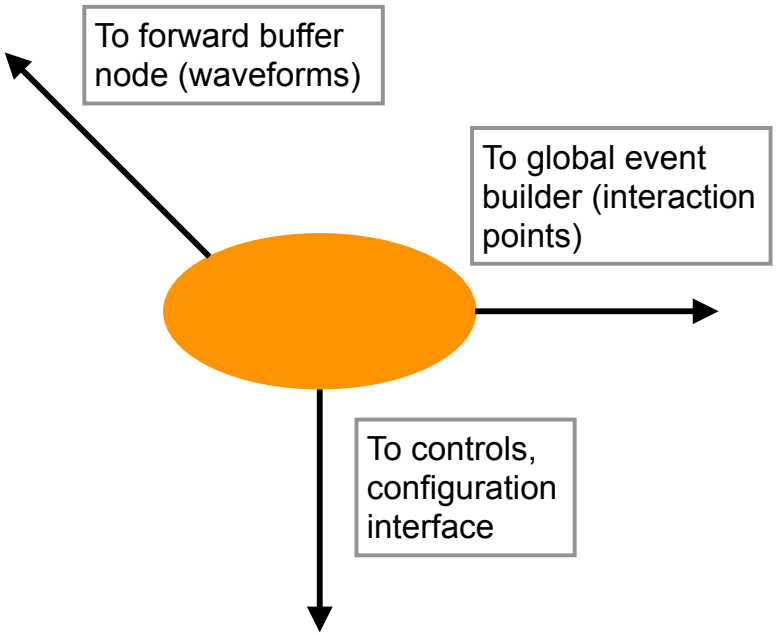
- Introduce a compute node that acts as a memory buffer between electronics and signal decomposition nodes
- Filter boards sends to a common buffer (static address), decomposition processes read from that buffer
  - Restore ‘pull’ model for decomp processes - load balancing
  - Allows for easy flow control, synchronous throttling of the trigger
  - UDP only necessary between filter board and forward buffer

Data Flows (w forward buffer)

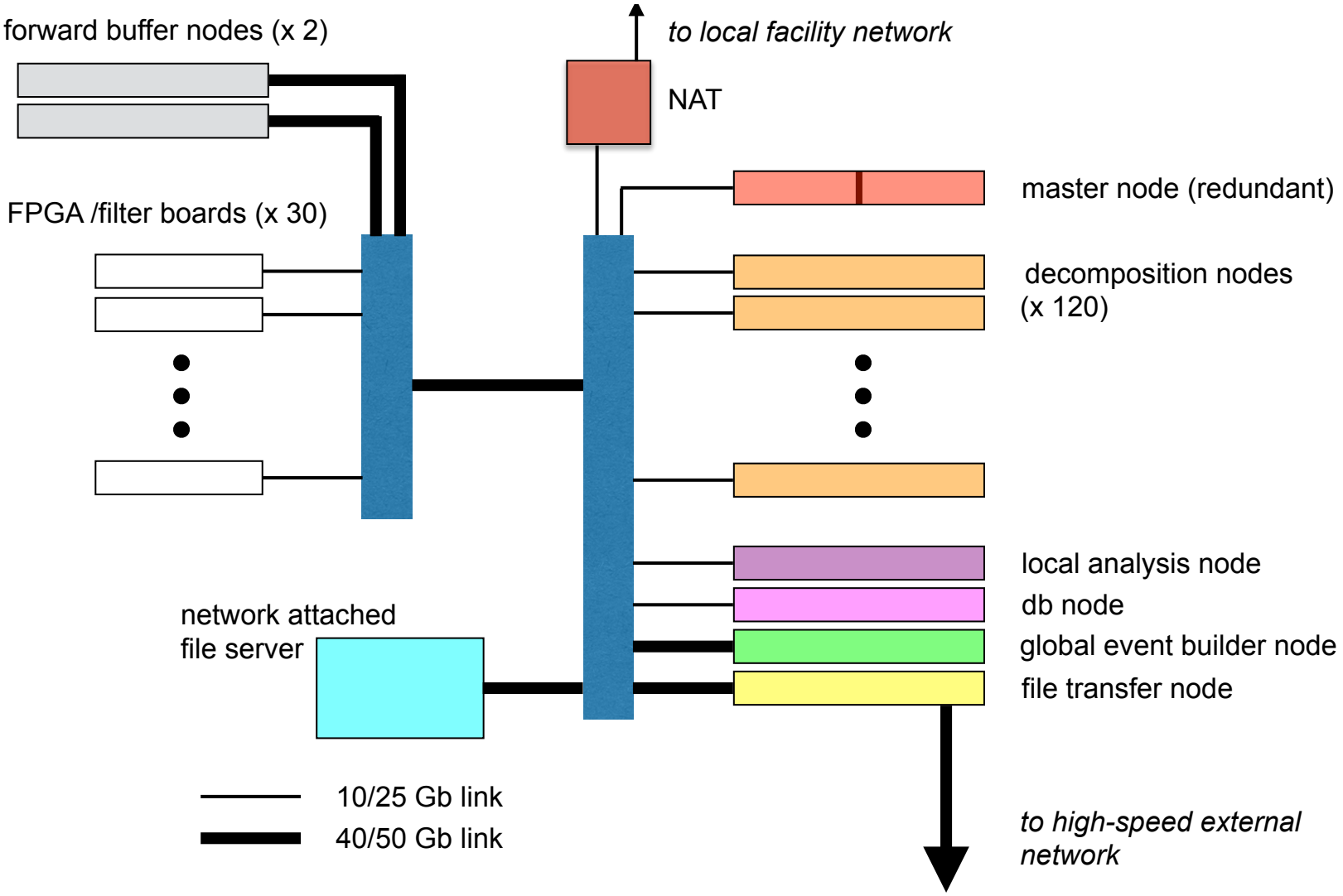


Signal Decomposition / Software Containers

- Encapsulate signal decomposition algorithm in a software container
- Each container:
  - Serves one crystal only
  - Implements thread-level parallelism
  - Has 3 (static) interfaces
- Expect of order 1000 such containers on local GRETA cluster
- Advantages:
  - Dynamic load-balancing in cluster (new container instances can be created on-the-fly)
  - Portability/deployability - development laptop, local cluster, AWS, your favorite HPC
  - Modularity - forces well-defined interfaces

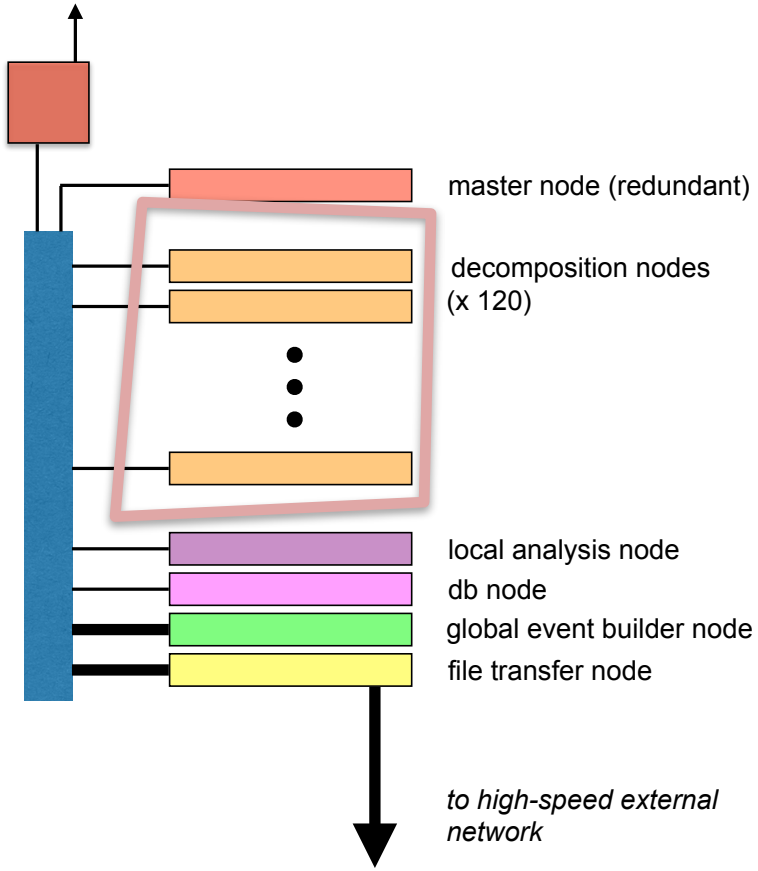


Cluster Components



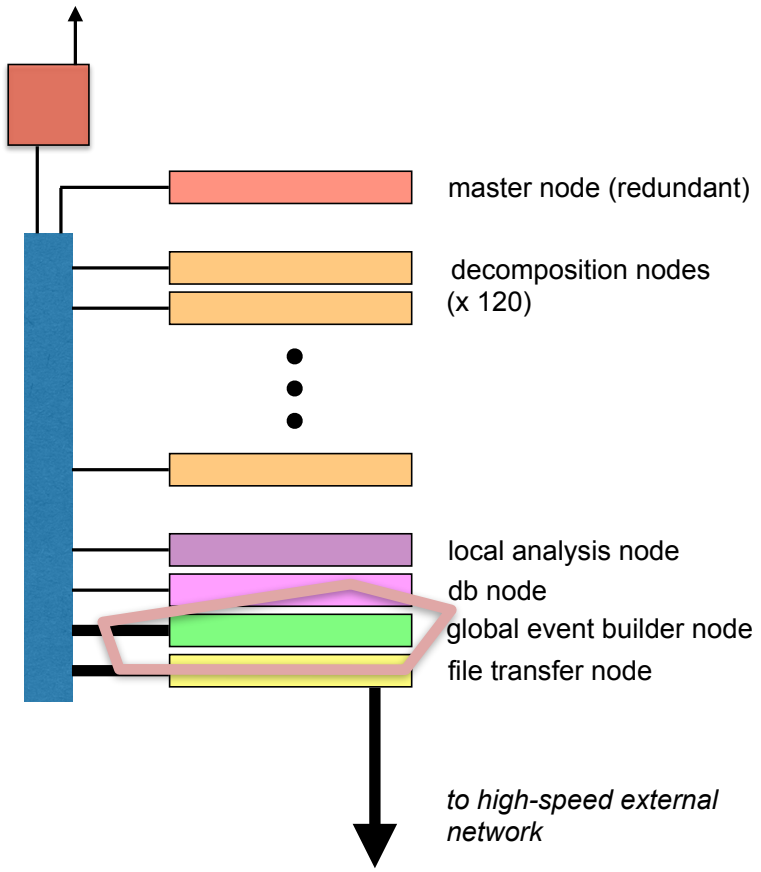
Signal Decomposition Nodes

- Signal decomposition processes implemented on a linux-based cluster
- 4800 cores (2016 - Xeon) required to keep up with data rate
- Estimates based on:
  - Dell 6220 - 2 processors/node - Broadwell 2860v4, 14-core 2.4GHz
  - 120 nodes, 30 chassis
- 60U rack space ('twin format'), 44KW
- Future .. more highly integrated processors may allow smaller form factor, reduced power
- Use standard LBNL cluster management scheme for stateless operation



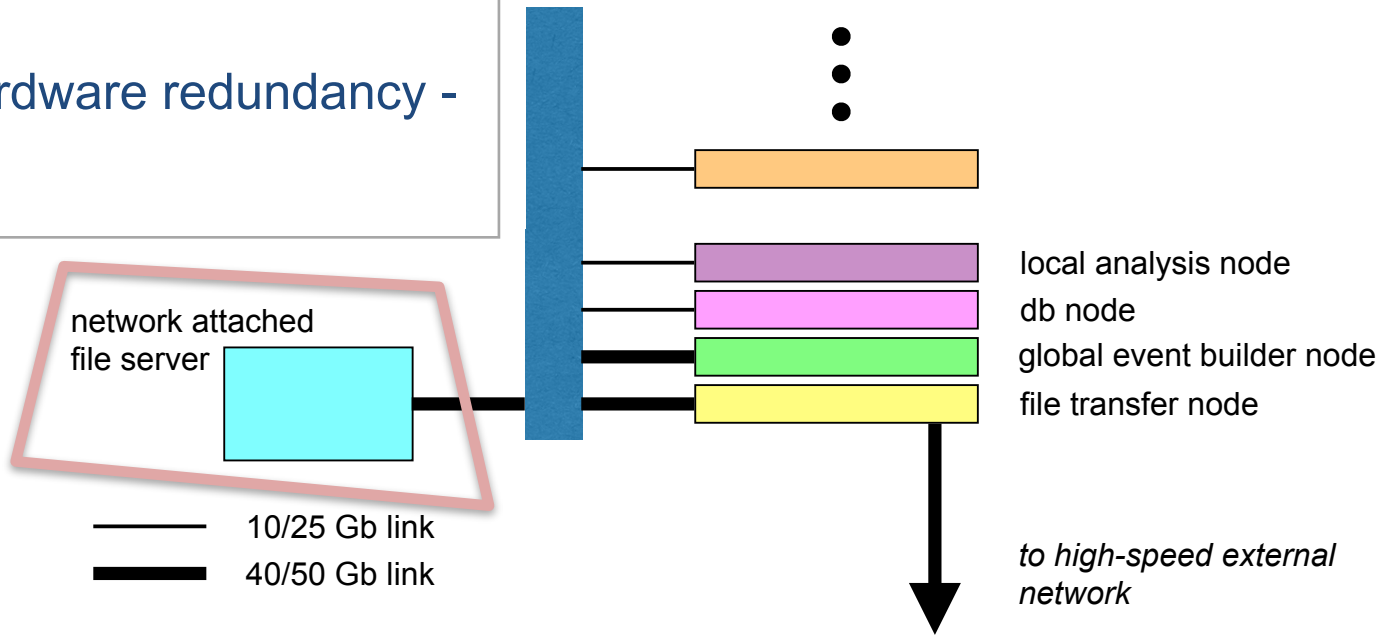
Global Event Builder

- Combines event data according to global timestamp - essentially a sort algorithm
- Will share codebase with forward buffer
- 120 crystals (≤ 240 MB/s)
- Auxiliary detector data (≤ 250 MB/s)
- Aggregate IO - 500 MB/s in, 500 MB/s out (to filestore or file transfer node)
- Seconds of latency → GB's of buffering



1 PB Data Store; high IO bandwidth

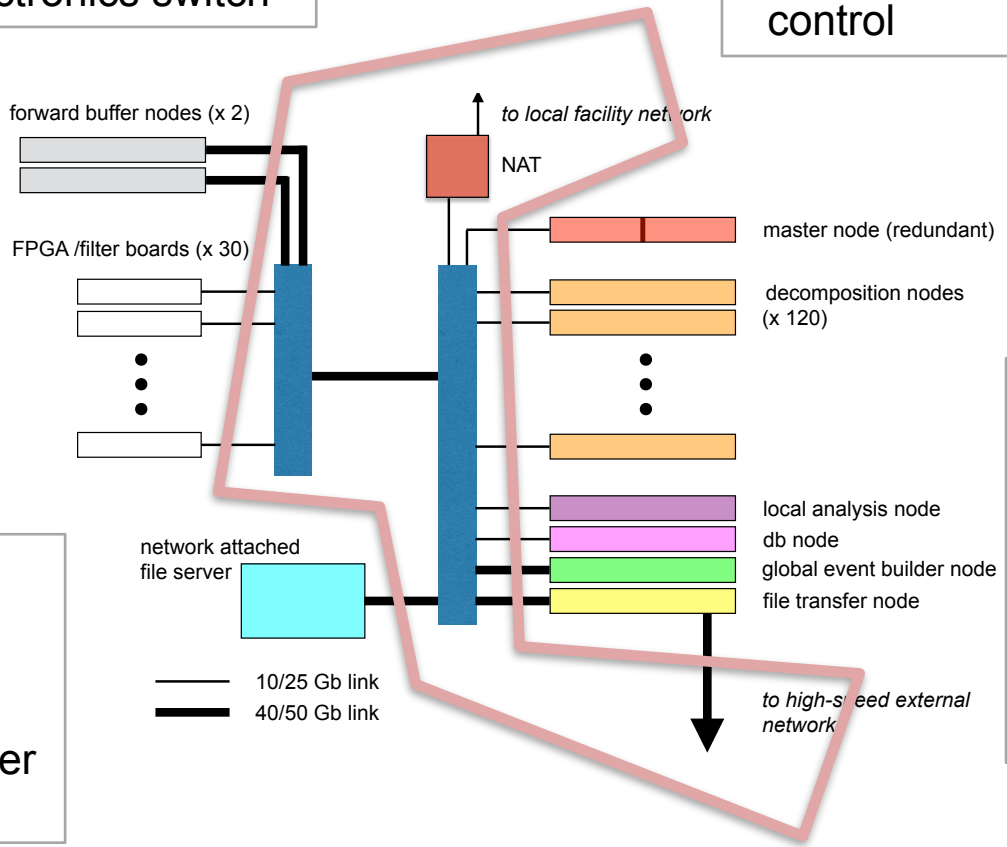
- Storage of experimental data, calibration data, basis, configuration files
- Key requirements:
  - 1 PB capacity: three 5-day mode-2 (interaction points) sets + mode-3 (waveform) sets [1 day equivalent]
  - 1 GB/s I/O bandwidth (×2 event-builder) .. this is probably very conservative
  - high level of hardware redundancy - vital subsystem



Network

- Two switch mainframes:
  - electronics shack
  - cluster racks
- switches linked by ≥ 2 40/50Gb links
- deep-buffering in electronics switch

- GRETA connects to local facility network through a NAT unit - abstracts address space, provides access control



- High bandwidth connections:
  - forward buffers
  - file server
  - global event builder
  - file transfer node

- Large data transfers to external sites done though dedicated file transfer node to a high-speed network

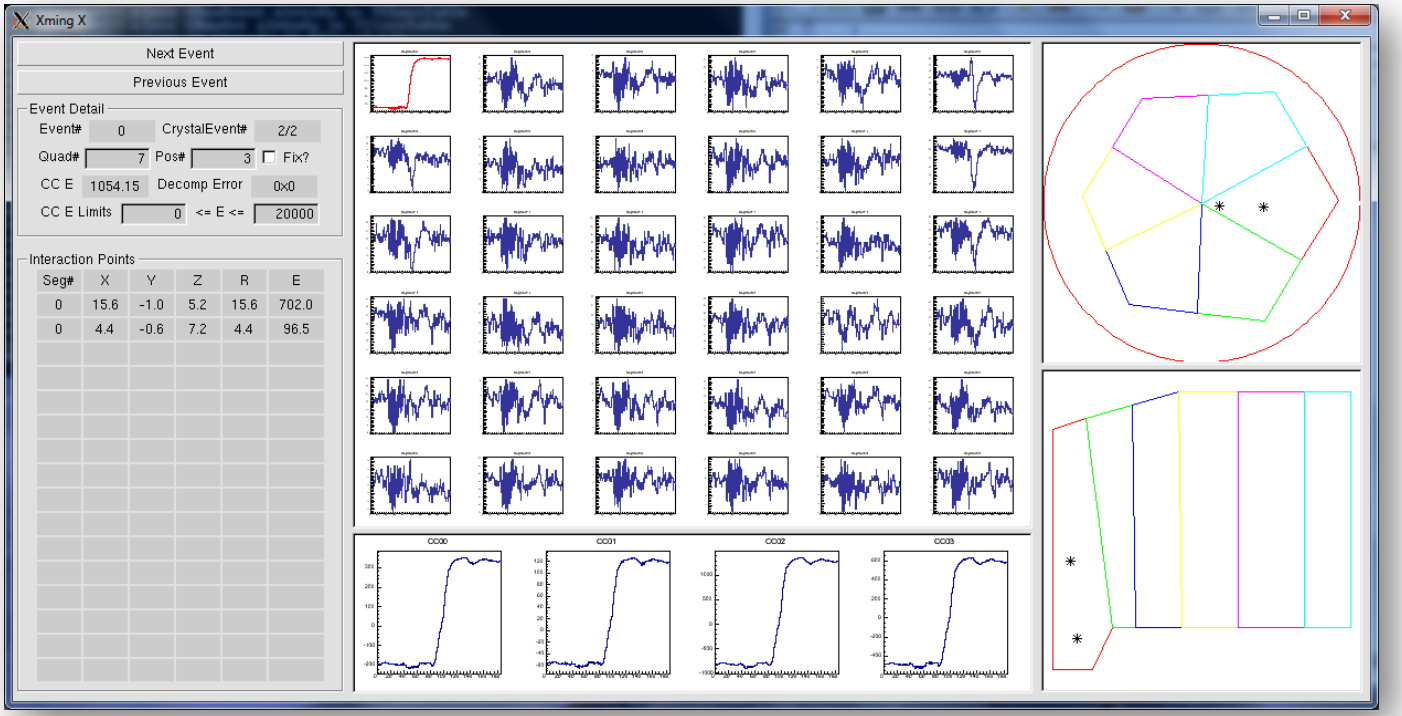
Controls and Monitors

- Large # of control monitor points (~ 50k) but most replicated on per channel or per crystal basis, latency ~200 ms
- Supports archiving, system restoration from checkpoint
- ***Electronics***
  - EPICs based interface using UDP
- ***Computing Cluster***
  - Currently in GRETINA, state machines (run by EPICS soft IOCs) implement controls on real-time processing codes at process level
  - Does not map well to large # of processes, software containers - will need another method
- ***Liquid nitrogen fill system***
  - to be based on commercial PLCs
  - remotely accessible monitors, alarms

Online Monitoring

- Allows for prompt monitoring and diagnostics to facilitate the optimization of experiments in real-time
- Includes the gamma-ray tracking package
- To be based on GRETINA analysis packages supported by the user community

ROOT-based visualization tool for signal decomposition in GRETINA



Future Possibilities

- Performance of modern computing platforms gives us new possibilities .. and new questions:

- Implicit architecture assumption - single-pass signal decomposition algorithm - could this change?
- Waveforms (mode 2) to HPC facilities? - 4 GB/s through DTN .. very possible (even today)
- Allow for energy filters in cluster? (not fpga's) - should have sufficient network/IO bandwidth

Summary

- The GRETA computing system provides online processing to determine the location of interaction points in the array, global event building, and controls/ monitors for all major subsystems
- The scope/requirements of the GRETA computing system are well understood from the experience with the GRETINA system
- Network-based readout of electronics changes computing architecture
- Take advantage of modern computing platforms and techniques