The PHENIX Event Builder

David Winter
Columbia University
FOCUS Seminar 13 Apr 2004

Outline

- DAQ Overview
- Hardware
  - EvB Machines
  - Gigabit Switch
- Software
- Design Architecture
  - SEBs, ATPs, EBC
  - Level-2
  - Monitoring performance

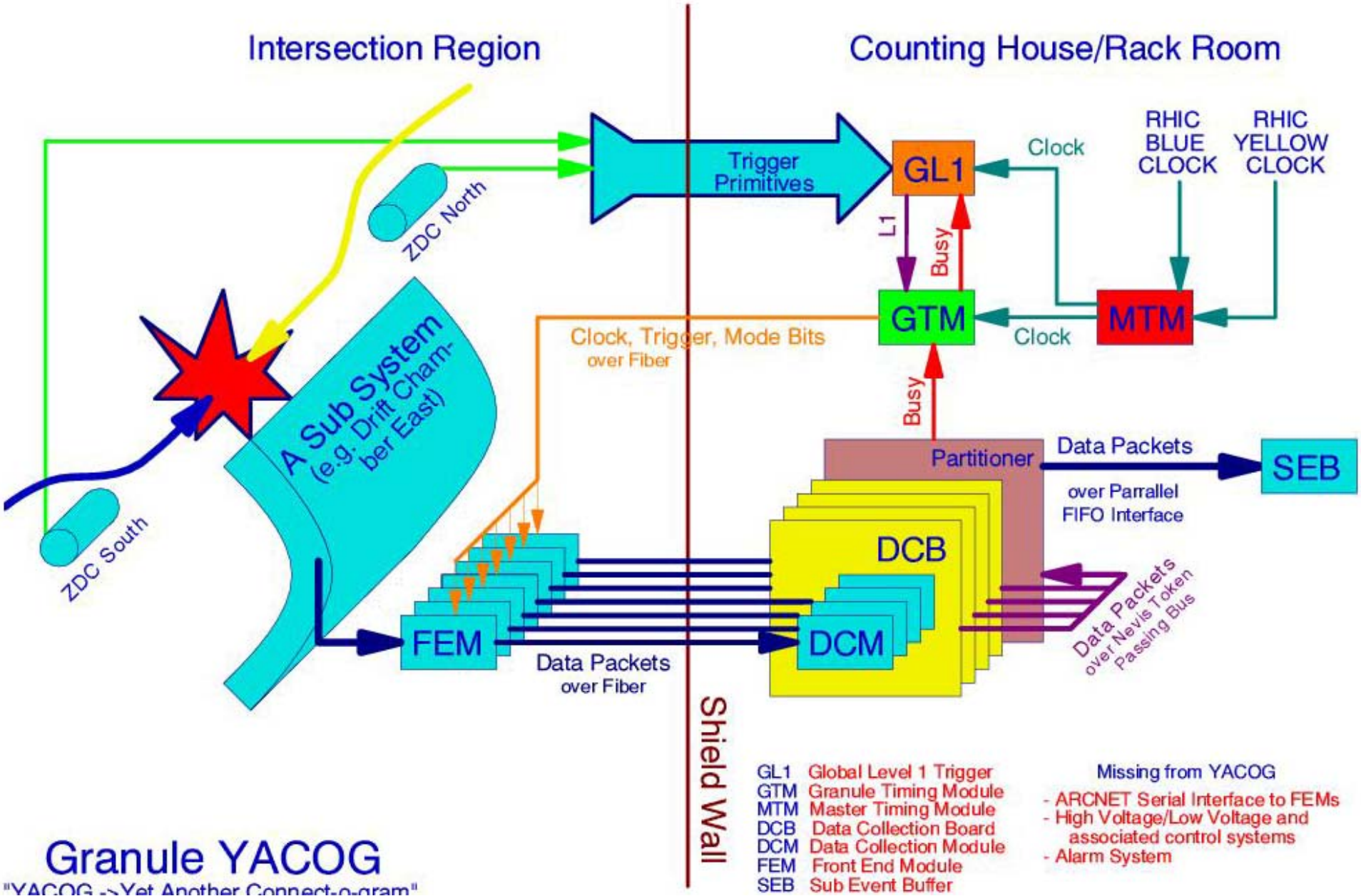
DAQ Overview: Motivation

- High Rates of interaction at RHIC
  - RHIC spec'd to 10MHz crossing/10kHz interaction rate (for AuAu, at least)
- Interest in rare physics processes
- Stringent Design Requirements
  - Need good triggering
  - Buffering
  - High Bandwidth
  - Fast Processing

Trigger

- One cannot collect data at every crossing, because
  - Too many empty crossings
  - It would take years to collect the data needed to do any interesting physics
- Trigger system is designed to tell the PHENIX detector to record data when a collision occurred
  - Or when an "interesting" collision occurs
    - One which has muon tracks, high energy or momentum charged or neutral particles.
- Signals which are used to make a trigger decision are taken from the FEMs and fed into the Local Level 1
  - Example: BBCLL1 - Beam Beam paddle multiplicities and hit times

The Hardware Components of A Granule



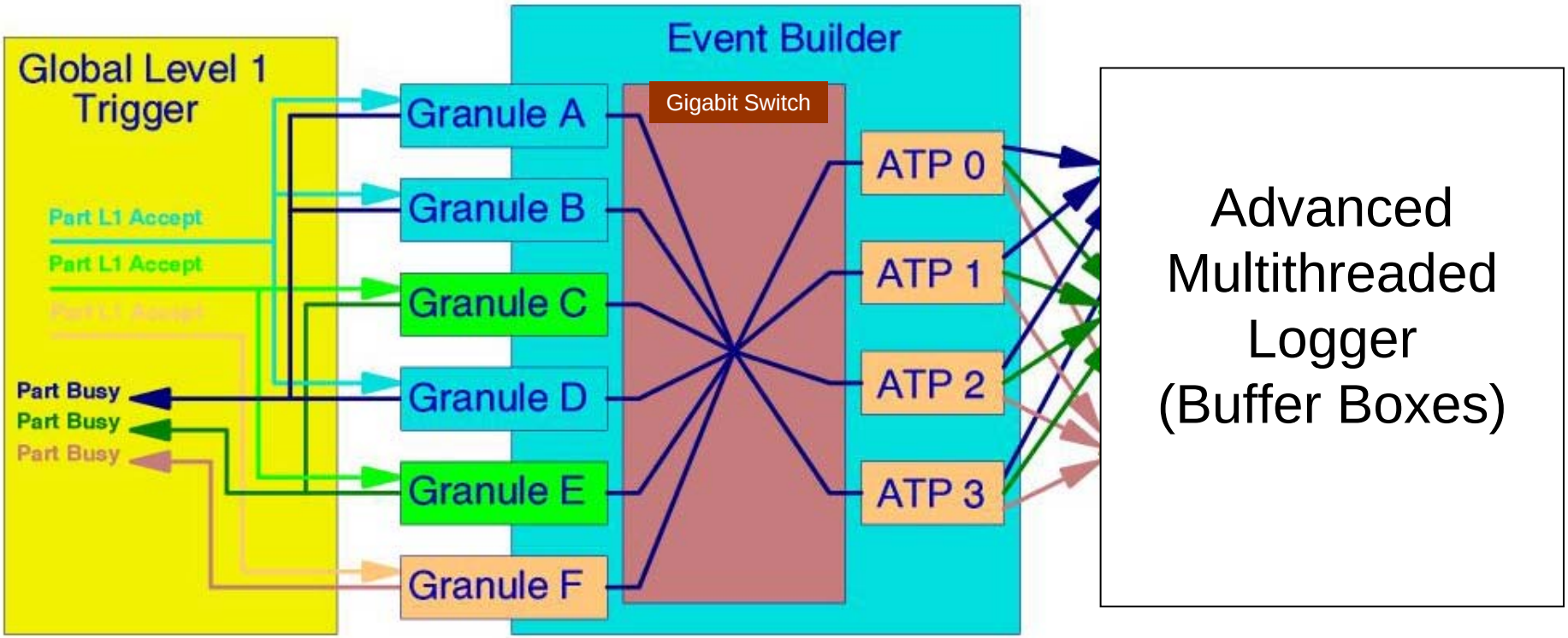
Granules and Partitions

A 3 Partition Example Configuration

Blue Partition
Granule A
Granule B
Granule D

Green Partition
Granule C
Granule E

Tan Partition
Granule F



DAQ overview: Event Builder

Performs the final stage of event assembly in the PHENIX DAQ, assembles fragments from each data stream into complete events and provides an environment in which Level-2 trigger processing can occur.

- There are three types of Event Builder Components: SEB, ATP, EBC
- Sub-Event Buffer (SEB) receives data from Partition Module (PM)
  - Event fragments + clock signals received and buffered in a PCI card (JSEB)
  - Data transferred over 32-bit parallel cables (allows JSEB to raise a hold against PM)
  - JSEB has two 1 MB buffers
  - PCI bus driver delivers data to SEB's memory from one JSEB buffer while the other receives data

DAQ overview: Event Builder

- Assembly Trigger Processor (ATP)
  - An ATP requests data from all SEBs that belong to a partition
  - Upon receipt of data from the SEBs, concatenates the event fragments and then executes Level-2 algorithms (if enabled)
  - Built events are then sent to the Buffer Boxes
- Event Builder Controller (EBC) receives notification that data has arrived at GL1 SEB
  - Assigns events to ATPs and tells them to fetch the data from the SEBs
  - Keeps track of completed assignments and sends "flush" messages to the SEBs when assembly has been completed

"Pull Architecture"

- All communication via Gigabit Ethernet
  - Data transferred from SEB to ATP using UDP
  - (LZO compressed) Data transferred from ATP to Buffer Boxes using TCP
  - Control messages sent via Multicast protocol

Hardware: EvB machines

Num		Manufacturer	CPU	RAM	FSB	Disk	OS	Network
28	11	Dell	Dual 1GHz P3	256 MB ECC	133 MHz	8 GB SCSI	NT4	Onboard Dual Fast Ethernet
	17						plus 1 W2K	plus Add-on Gigabit PCI card
45		Microway	Dual 2.4GHz P4 Xeon	1 GB ECC	400 MHz	37 GB IDE	W2K	Onboard Fast Ethernet/Gigabit
32		Microway	Dual 2.4GHz P4 Xeon	1 GB ECC	533 MHz	37 GB IDE	W2K	Onboard Dual Gigabit
105								

- Rack-mounted 1U x86 dual-CPU servers
- NT4 SP 6, Windows 2000 Pro SP3 and SP4
- 39 SEBs
- 52 ATPs
- 1 EBC
- All connected to CRT display in CR via cascaded Raritan Master Console II KVM switches

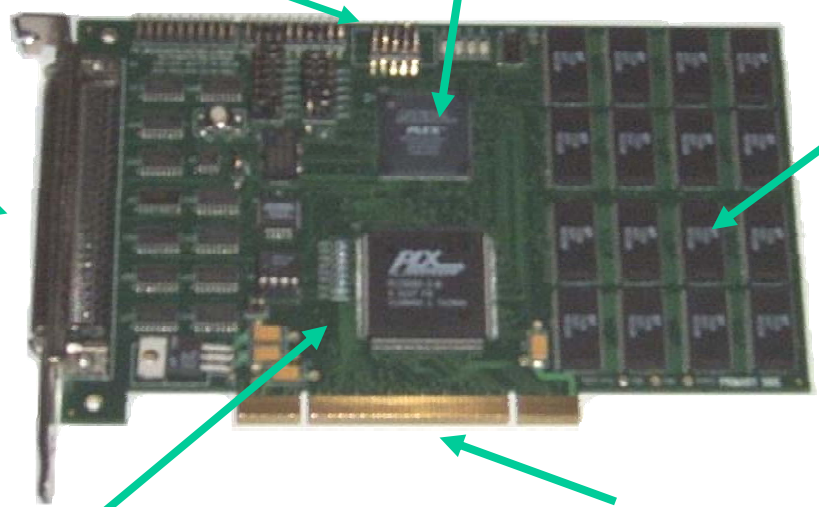
JSEB Interface Card to SEB

JTAG port for programming FPGA

Altera FLEX10k FPGA

2 banks of 1 MB static RAM

Input for JSEB cable from Partitioner Module



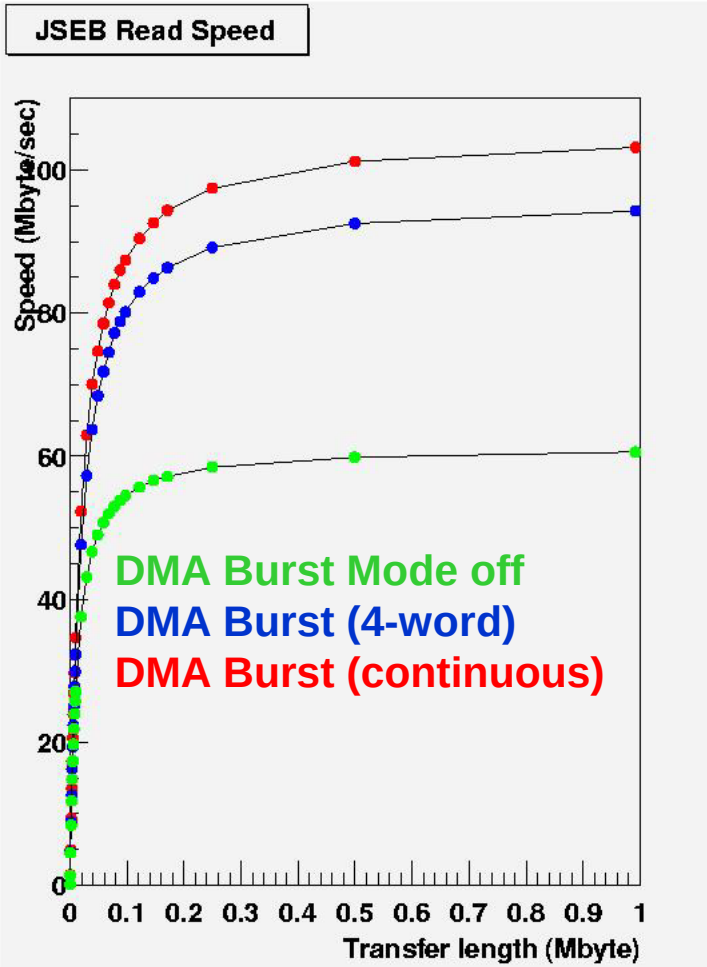
PLX 9080 PCI controller

PCI 2.1 interface: 32-bit 33 MHz with 3.3v or 5v signaling

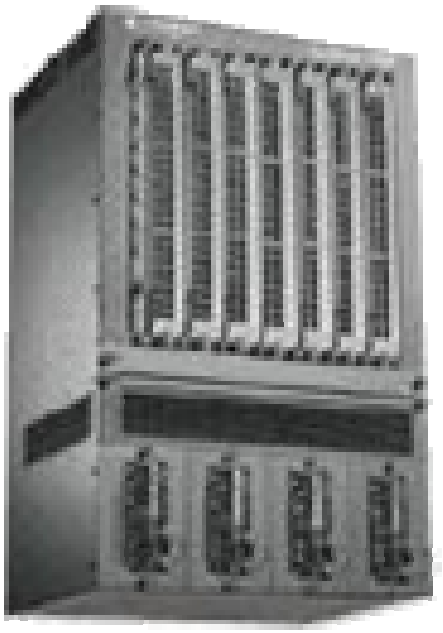
JSEB cont'd

- Interfaces the Partitioner Module (PM) to the SEB
- Data from DCMs transmitted via 50-pair 32-bit parallel cable
- Firmware contained in FPGA
  - Versions 0x11 and 0x13
- (Pseudo) Driver provided by Jungo
- DMA burst mode not stable

  (Firmware v. 0x27 looks promising)



Hardware: Gigabit Switch



- Foundry FastIron 1500
- Total switching capacity = 480 Gbps
- Jumbo Frames (we use 9014 byte MTU)
- 15 Slots (10 in use)
  - 2 control modules (includes 24 100 Mbps ports each)
  - 6 16-port Gigabit modules (95 ports in use)
    - 39 SEBs
    - 52 ATPs
    - 1 EBC, phnxbox2, phnxbox3, corba

Software

- Development environment: Visual C++ 6.0 SP5
- Code developed and compiled on dual-CPU workstation SEBDEV in CR
- Executables distributed to EvB machines from SEBDEV (via FTP)
- Control via CORBA (Iona ASP 6.0)
- Run Control written in Java
- Performance monitor written in Java

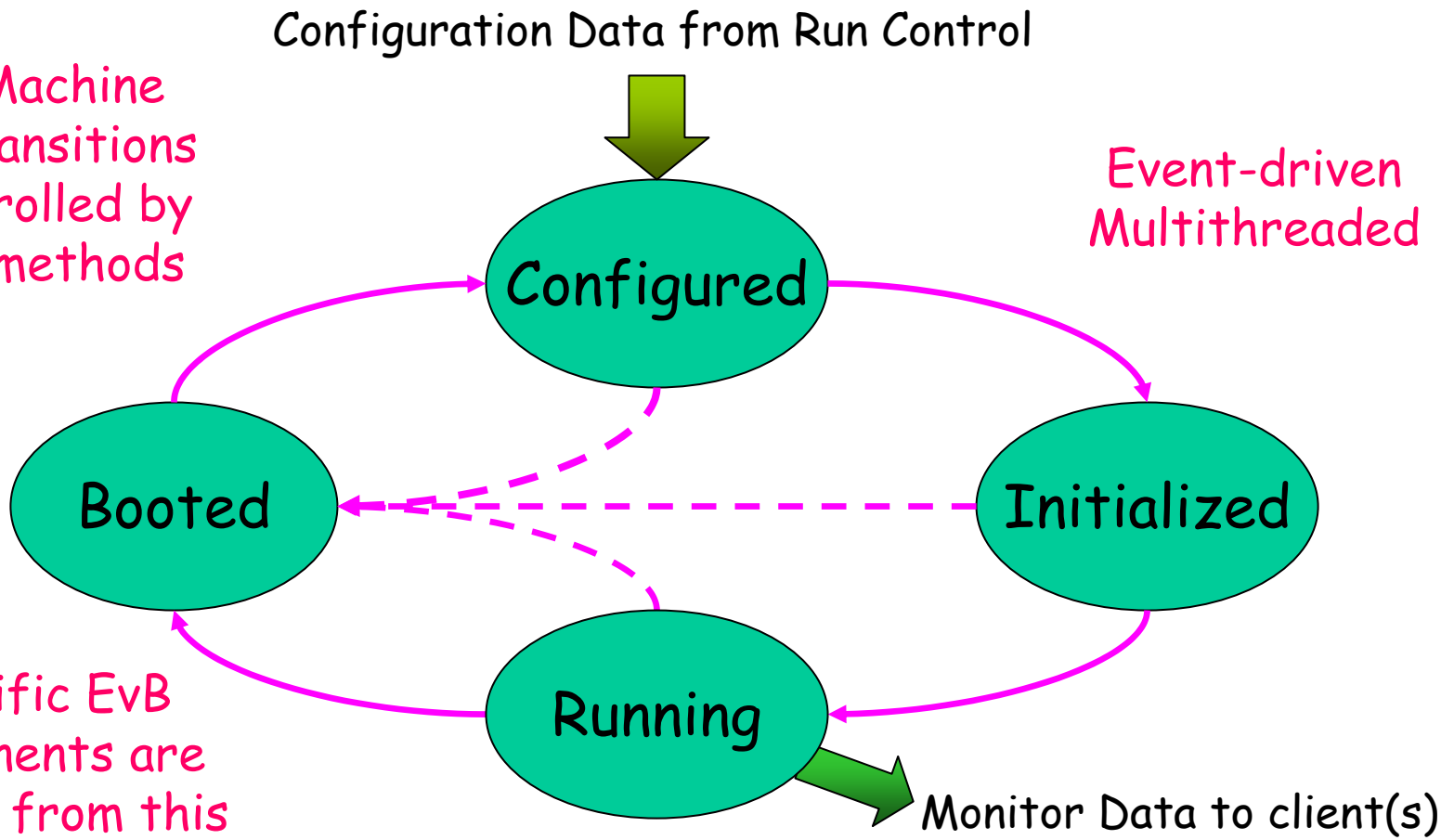
CORBA

- Common Object Request Broker Architecture
- The networking protocol by which the run control software components talk to each other.
- Based on a client/server architecture through a heterogeneous computing environment.
  - VxWorks, Linux, Windows NT/2000
- Servers implement "CORBA Objects" and execute the functionality described in the member functions of the object.
- Clients get a reference to the Servers "CORBA Object" and execute their member function. This causes the server to execute the code of the member function.

Basic EvB Component Design

State Machine whose transitions are controlled by CORBA methods

Event-driven Multithreaded

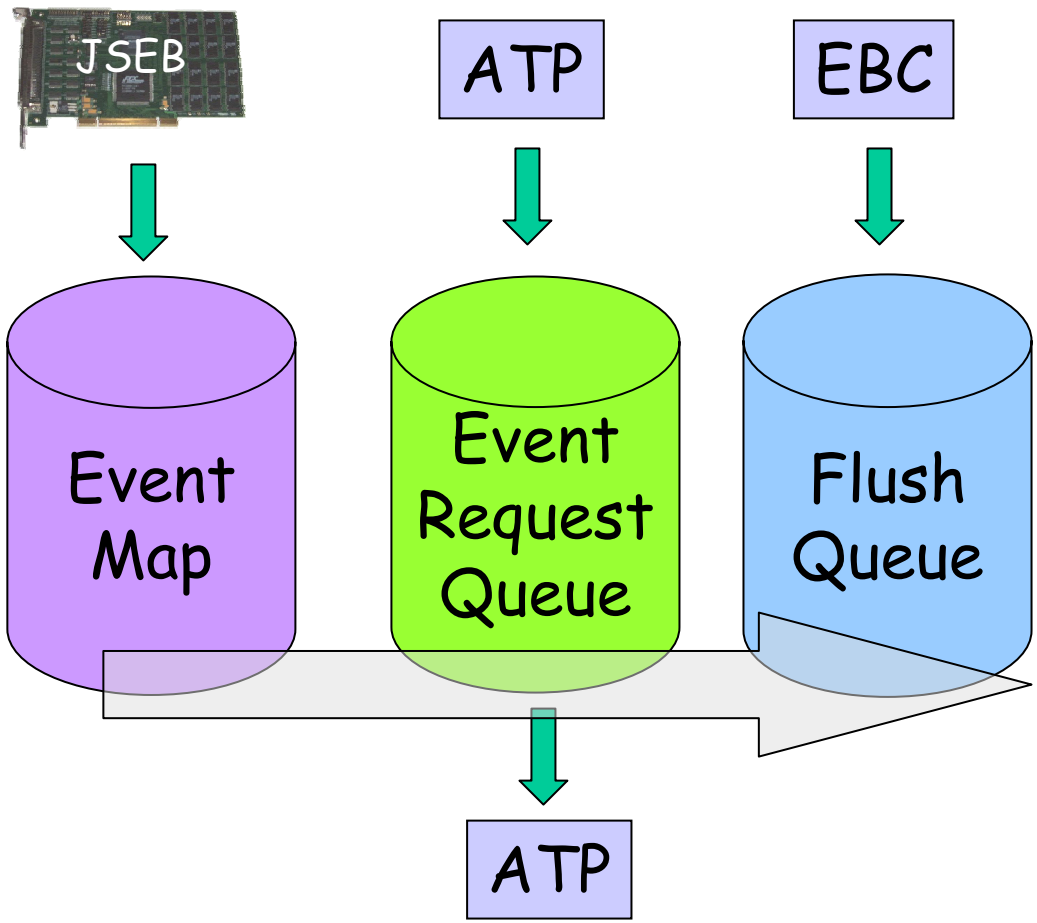


Specific EvB components are derived from this model

Communication

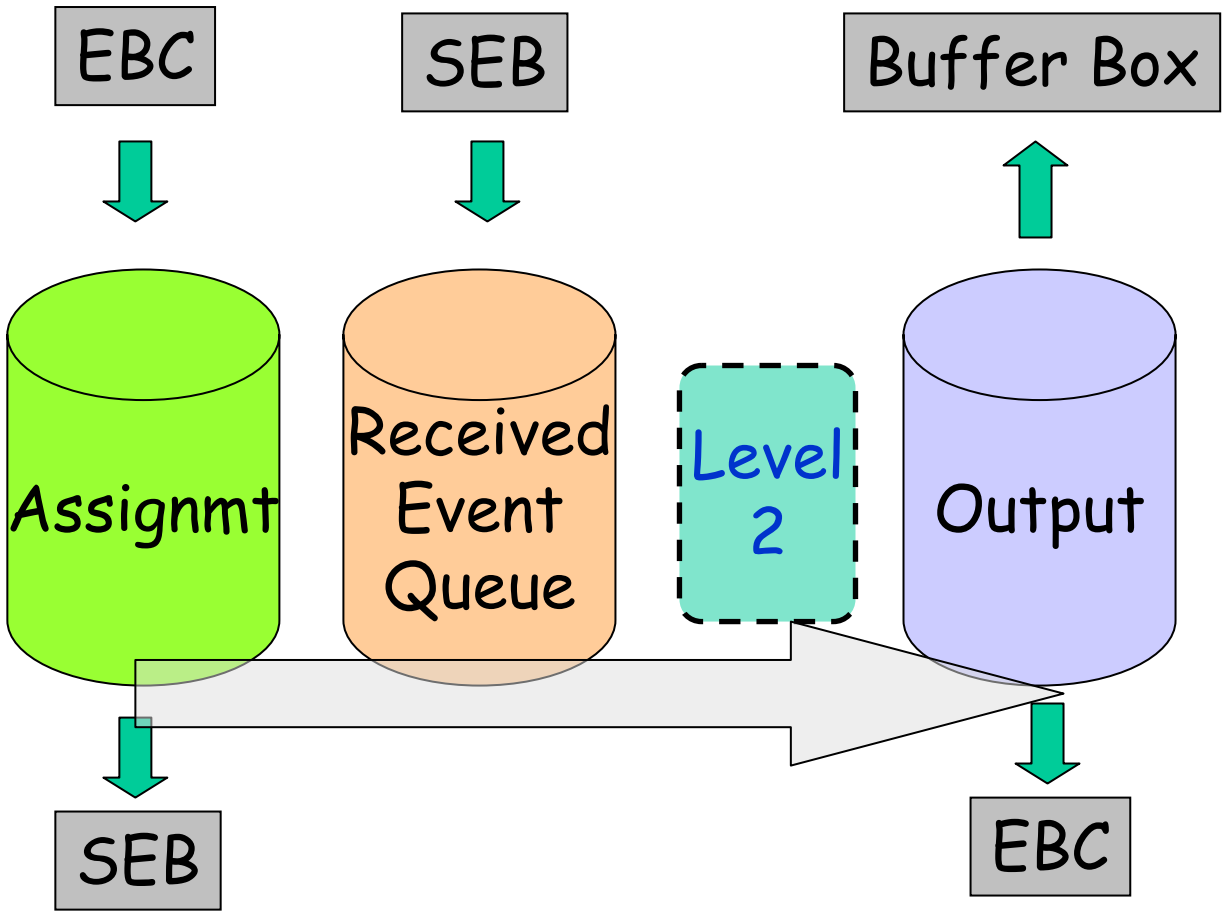
- Messaging system
  - Each message has a well defined response/action associated with it
  - Data sent as messages
  - Control also implemented as message system
  - Example: ATP sends a DataRequest Message to SEB, SEB responds with Data Message for requested event number
- All network communication done with UDP
- Broadcast messages sent via Multicast protocol

Sub Event Buffer (SEB)



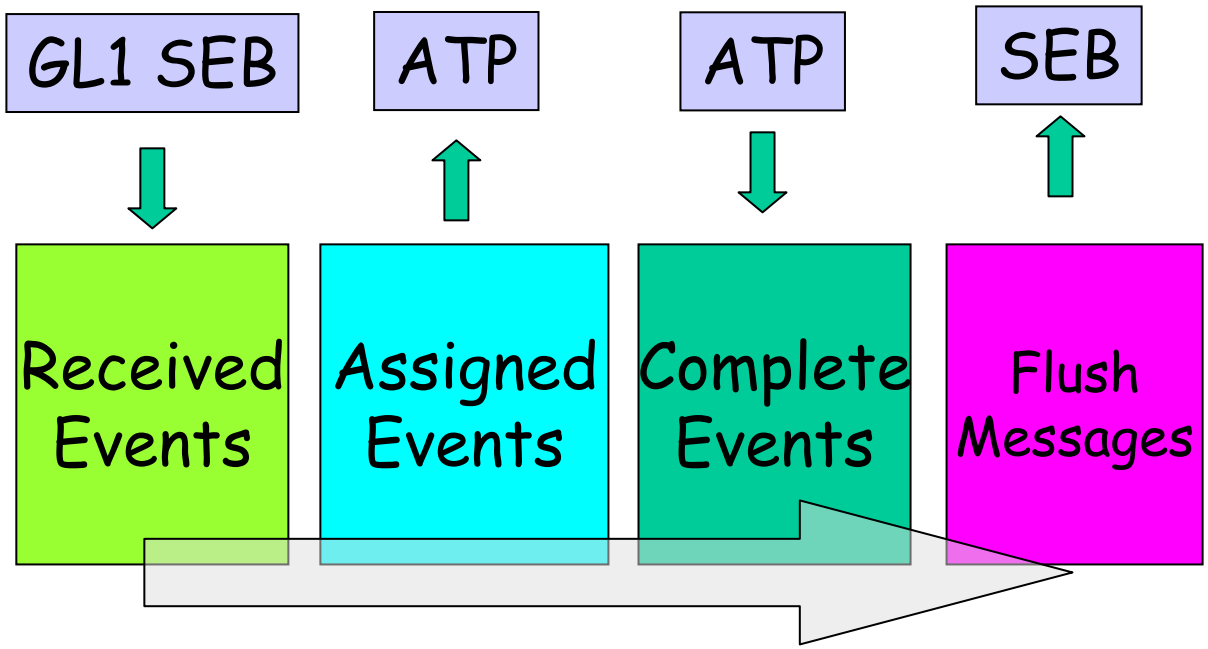
- Four threads: 1 Receive, 2 Transmit, 1 Flush
- Events received from JSEB
- ATPs send requests for events
- Event sent to ATP
- After event is acknowledged by EBC, it is scheduled for flushing

Assembly Trigger Processor (ATP)



- Threads: 1 Receive, 2 Processing, 1 Output
- Receives assignments from EBC
- Sends requests to SEBs
- Receives events from SEB
- Optionally executes Level-2 on assembled event
- Writes to BB
- Notifies EBC of completion

Event Builder Controller (EBC)



- Receives notification of events from GL1
- Assigns incoming events to ATPs
- Uses a round-robin load-leveling scheme
- Receives notification of event completion from ATPs
- Sends signal to SEBs to flush data from memory

Level-2: Design

primitives perform calculations, record self, available to all algorithms

algorithms high level decisions based on primitive calculations

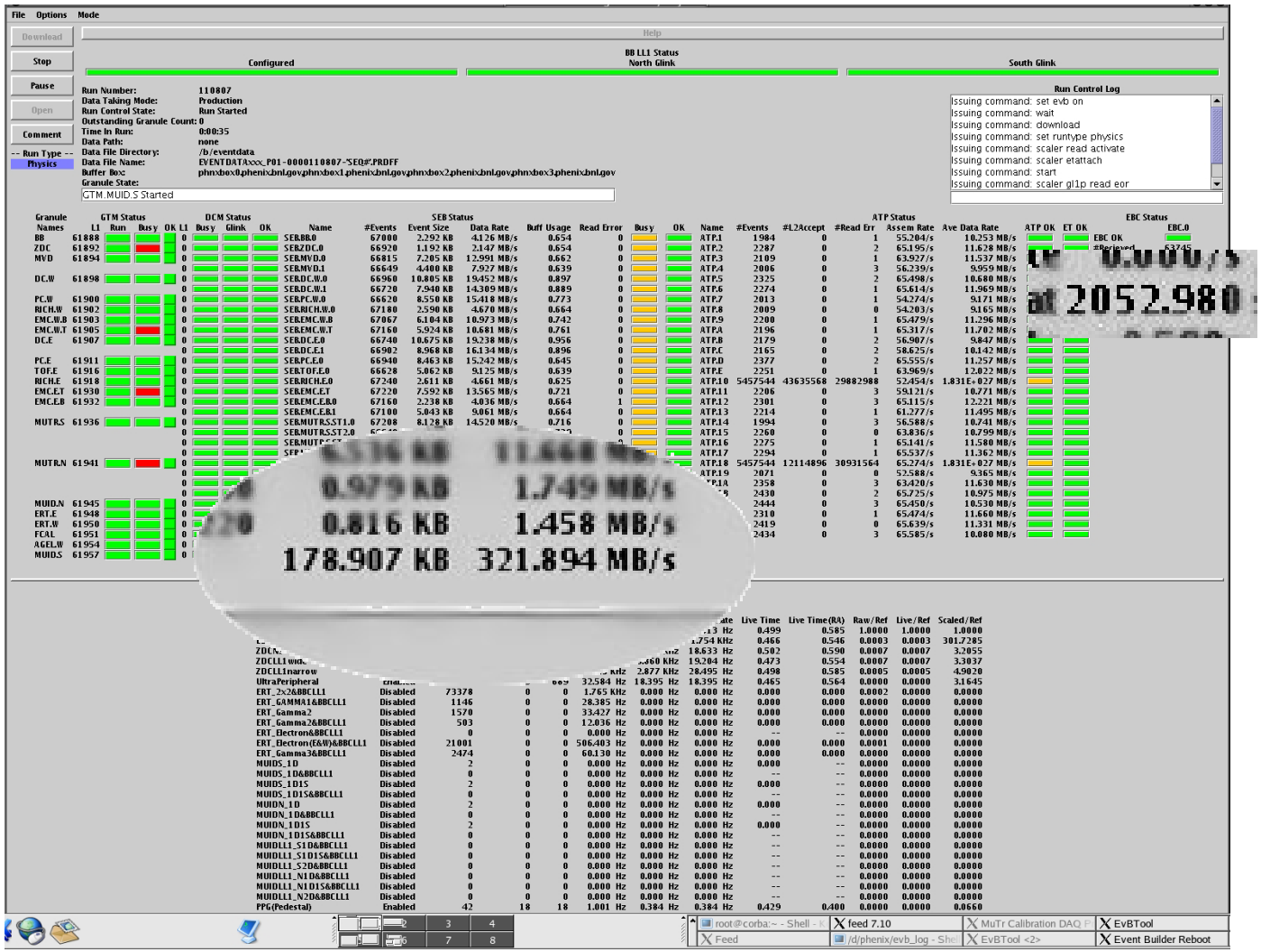
accessors access raw data in PHENIX *Packet* form

dbprimitives calibration constants, geometry info

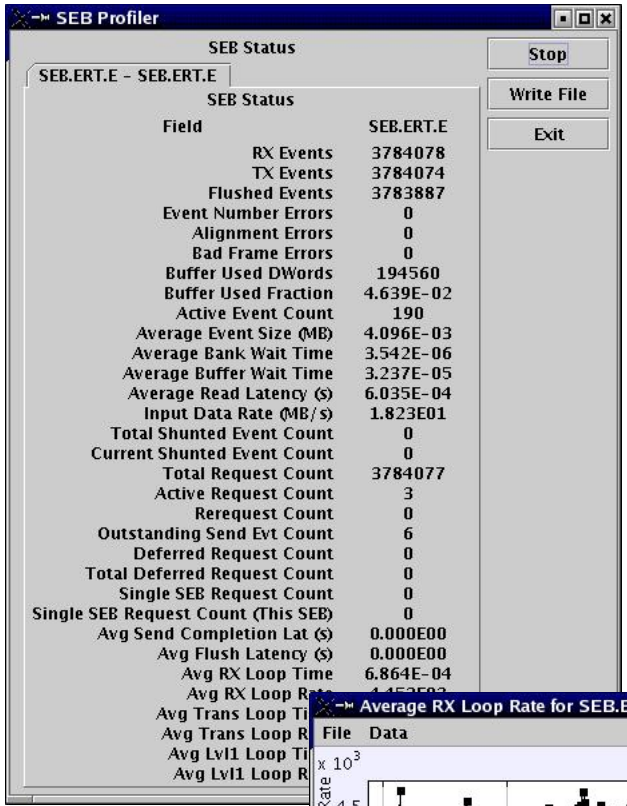
classes (Component Base Classes, Lvl2Array,etc)

- Logical place to run Level-2: ATP (contains whole event)
- Primitives/Algorithms saved in ATP data stream
- Database primitives can read from db or ASCII files
- Can be run during event building or as later pass over data stream

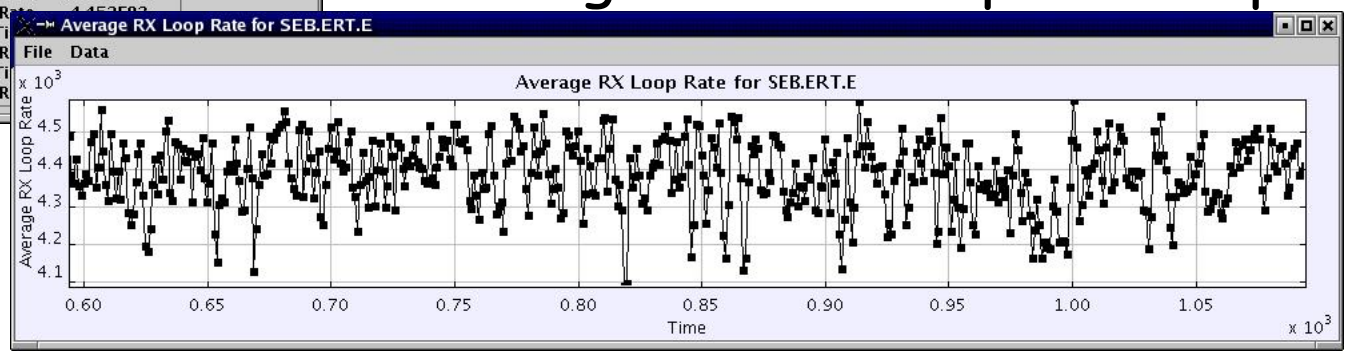
Run Control


File Options Mode
Download
Stop
Pause
Open
Comment
-- Run Type --
Physics
Help
Configured
BB LL1 Status
North Glink
South Glink
Run Number: 110807
Data Taking Mode: Production
Run Control State: Run Started
Outstanding Granule Count: 0
Time In Run: 0:00:35
Data Path: none
Data File Directory: /b/eventdata
Data File Name: EVENTDATAxxx_P01-0000110807-'SEQ#.PRDFF
Buffer Box: phnxbox0.phenix.bnl.gov,phnxbox1.phenix.bnl.gov,phnxbox2.phenix.bnl.gov,phnxbox3.phenix.bnl.gov
Granule State:
GTM.MUID.S Started
Run Control Log
Issuing command: set evb on
Issuing command: wait
Issuing command: download
Issuing command: set runtype physics
Issuing command: scaler read activate
Issuing command: scaler etattach
Issuing command: start
Issuing command: scaler gl1p read eor
GTM Status
DCM Status
SEB Status
ATP Status
EBC Status
0.979 KB 1.749 MB/s
0.816 KB 1.458 MB/s
178.907 KB 321.894 MB/s
root@corba:~ - Shell
feed 7.10
MuTr Calibration DAQ
EvBTool
Feed
/d/phenix/evb_log - She
EvBTool <2>
Event Builder Reboot

Performance Monitor



- Each component keeps track of various statistics
- Data served via CORBA calls
- Java client displays stats in "real time"
- Strip charts to study data as function of time
- Histograms to compare components

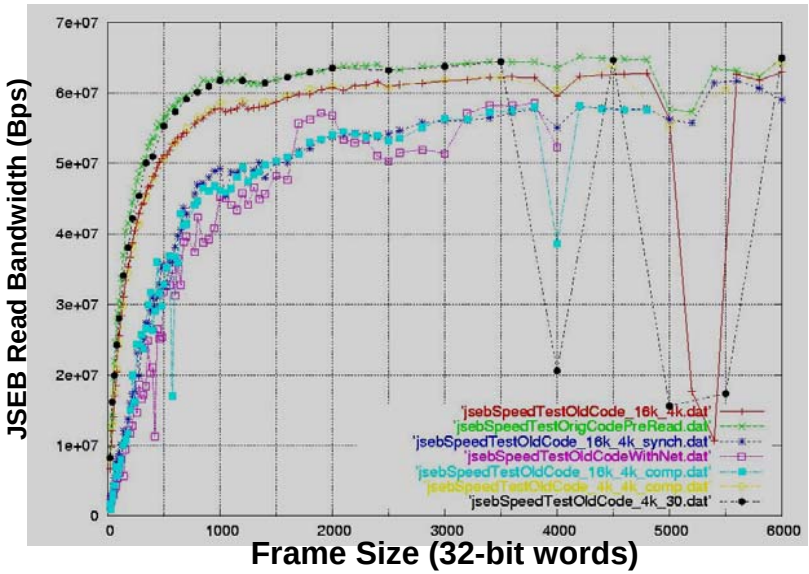


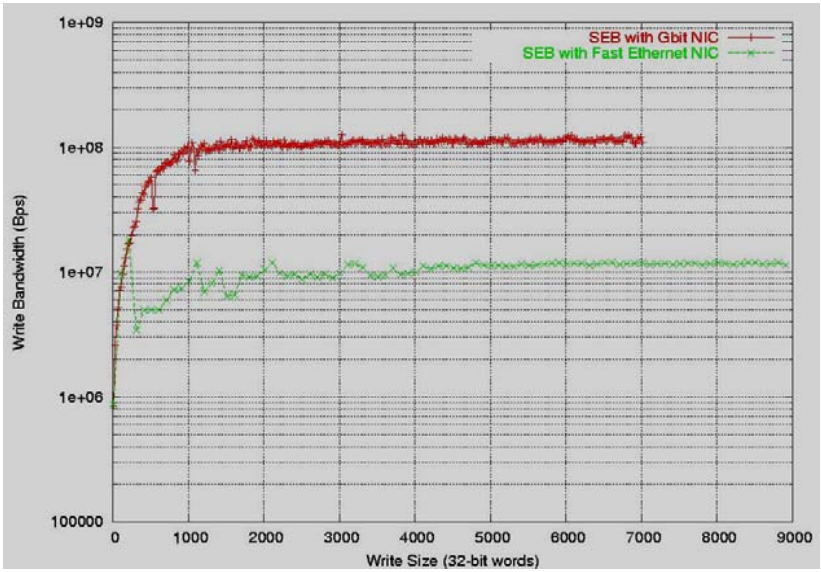
Future

- Expansion of capacity
  - EvB Components – more servers -> more data faster!
  - Switch – room for five more boards -> additional 80 ports
- Linux port
  - Made realistic by removal of ATM
  - Common platform for Level-2, offline, etc.
  - Easier to develop & maintain
  - Flexibility & tunability
  - Take advantage of P4 hyperthreading

Future (2)

- In a position to really study the performance of various pieces of the EvB
- Attention to the details of architecture





Summary

- High interaction rates at RHIC and PHENIX interest in rare physics processes place stringent constraints on DAQ design
- The DAQ is designed to be fast and allow various subsystems to be run independently
- The Event Builder assembles event fragments and sends whole events to storage
- EvB is a natural stage to perform Level-2 decisions (though it's not the only place)
- Critical part of making Run4 the success it's been!
  - Achieved archiving rates up to 350 MB/s
  - Event rate throughput as high as 2 kHz (>3 kHz for pp?)

The People

- Event Builder
  - Brian Cole, Jiangyong Jia, Dave Winter, Jiamin Jin
- Level-2
  - Tony Frawley, Brian Cole, Jiangyong Jia, Jiamin Jin, Hugo Pereira da Costa, Xiaochun He, Gobinda Mishra, Chris Cleven, Hai Qu, Jamie Nagle
- Emeritus: Peter Steinberg, Sean Kelly, Sotiria Batsouli (and my apologies to any others I've missed…)