# Event Building at Multi-kHz Rates:

## Lessons Learned from the PHENIX Event Builder

Real Time 2005
Stockholm, Sweden
5-10 June 2005

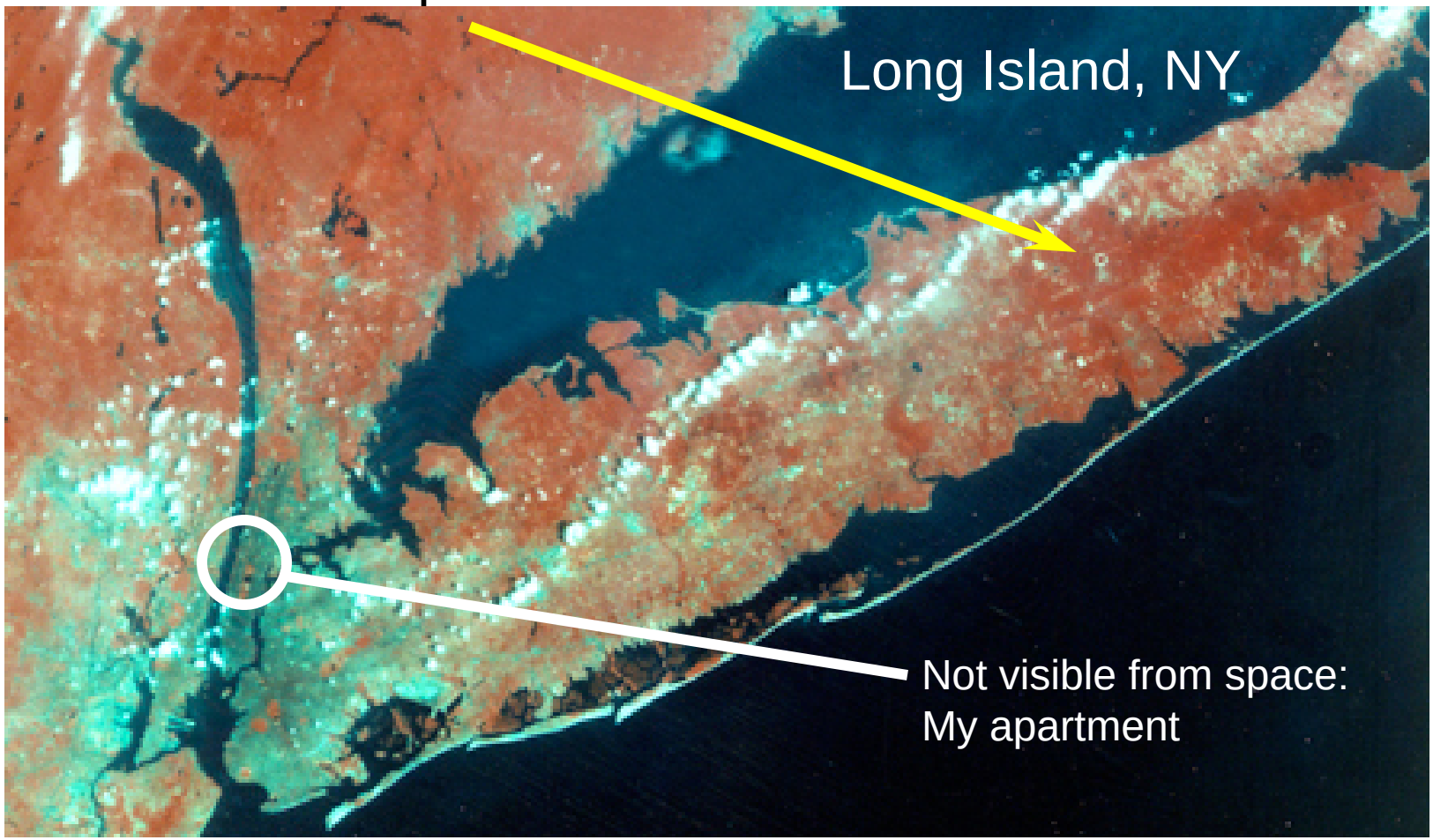
David L. Winter
*for the PHENIX Collaboration*
COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

# Outline

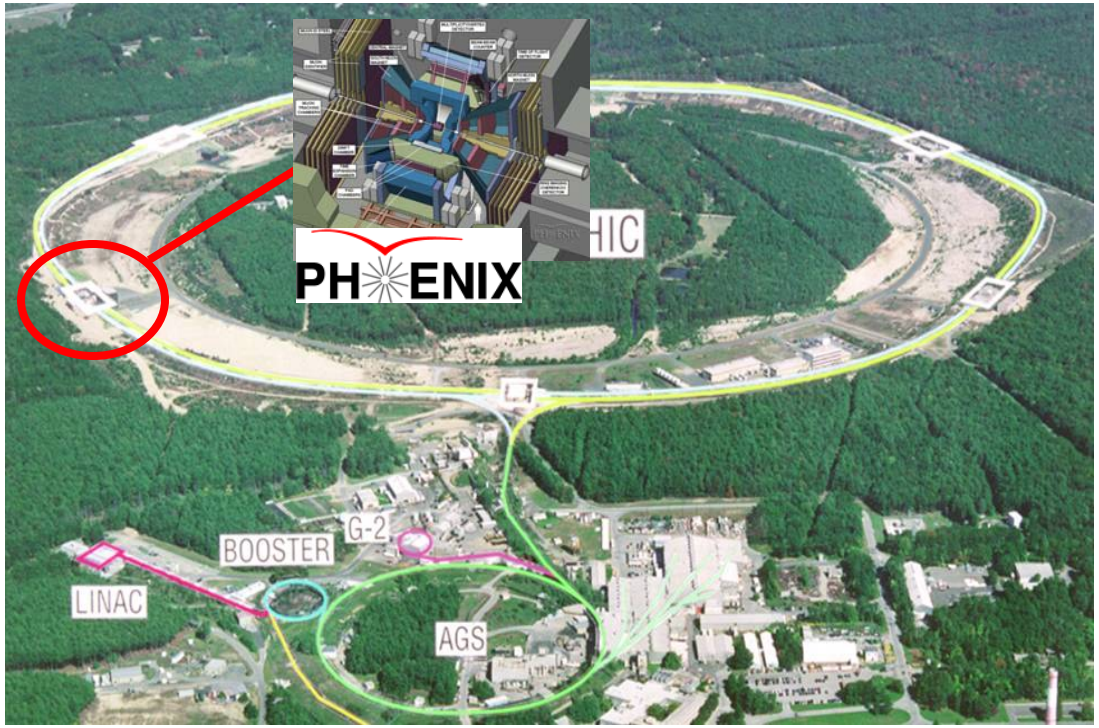
- RHIC and PHENIX
- DAQ & Event Builder
  - DAQ and EvB overview
  - Design and History of the Event Builder
  - How to reach 5 kHz (or die trying…)
- Summary

# RHIC

- RHIC == Relativistic Heavy-Ion Collider
- At Brookhaven National Laboratory, 70 miles from NYC
- Is visible from space



# Taking Data at RHIC

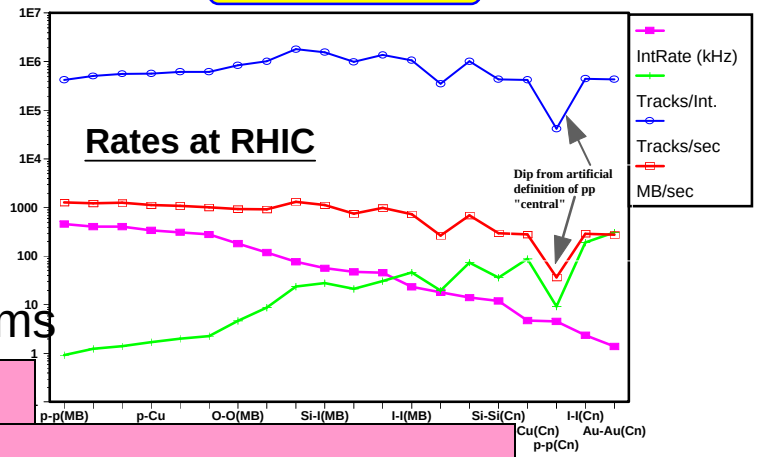



- Two independent rings (3.83 km circ)
- Center of Mass Energy:

$$\sqrt{s_{NN}} \approx \frac{Z}{A}(500\,\mathrm{GeV})$$

  - ➔ 500 GeV for p-p
  - ➔ 200 GeV for Au-Au (per N-N collision)
- Luminosity
  - Au-Au: 2 x $10^{26}$ $cm^{-2}$ $s^{-1}$

**PHENIX DAQ**




- PHENIX faces unparalleled challenges
  - Needs to handle wide range of collision systems
  - Designed for rare probes
    - High statistics, deadtime-less DAQ
  - Large, complicated detector
    - 14 subsystems
    - ~ 200 kB/Event (AuAu)

Run-5

~170(90) kB/event CuCu(pp)
5 kHz rate
⇨0.5-0.9 GB/s !!

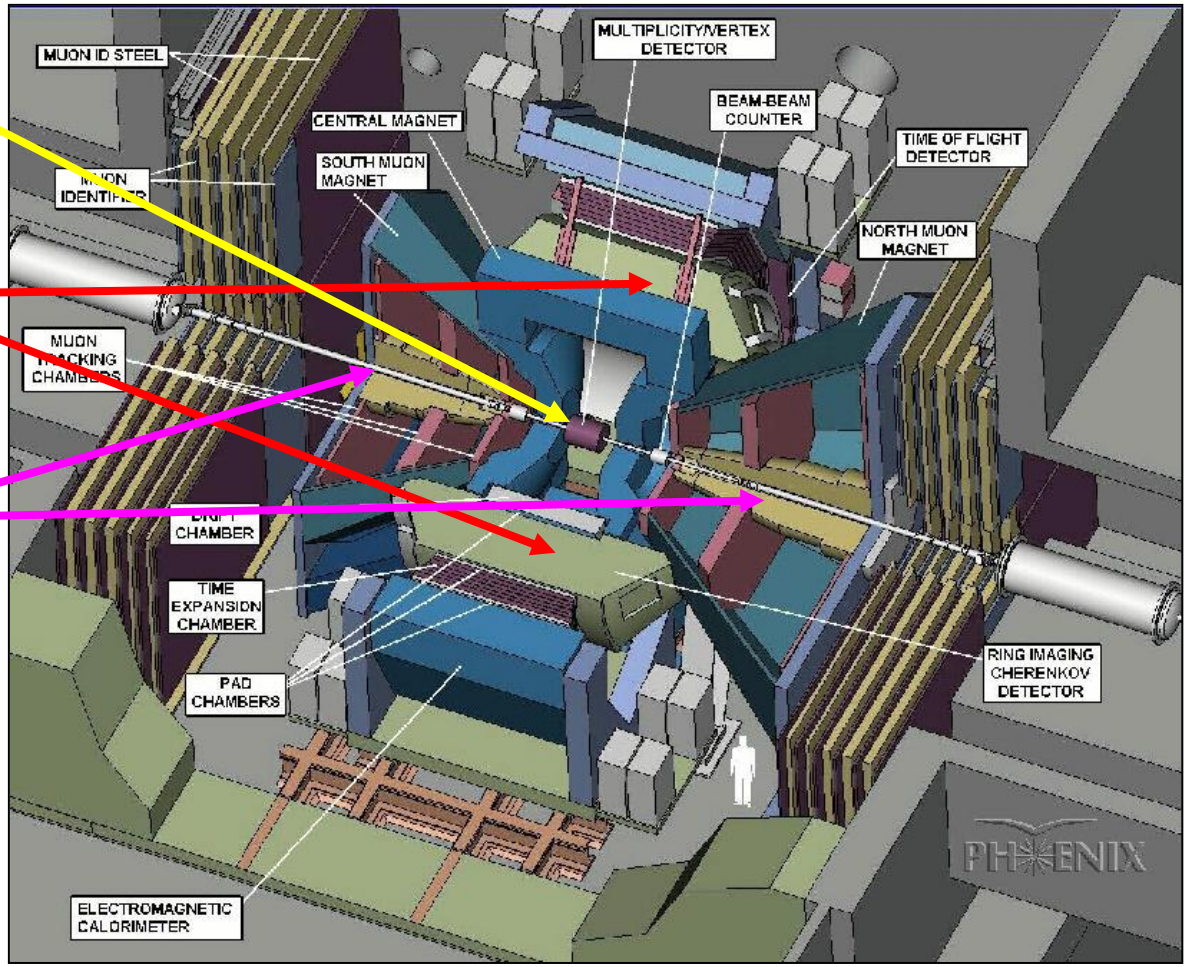
# PHENIX

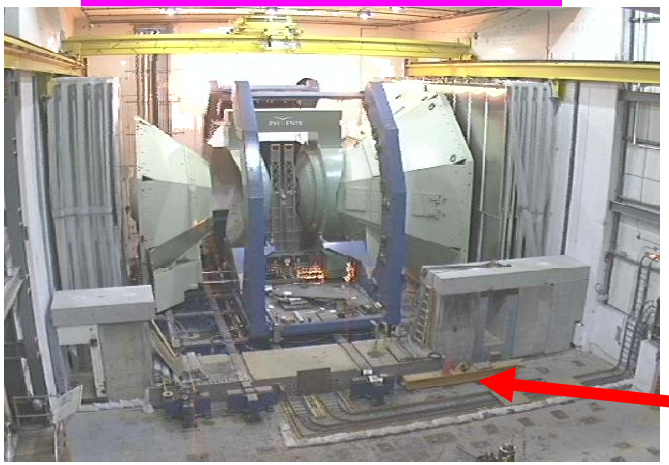
- **P**ioneering **H**igh-**E**nergy **N**uclear **I**nteraction e**X**periment

**Event characterization detectors in center**

**Two central arms for measuring hadrons, photons and electrons**

**Two forward arms for measuring muons**



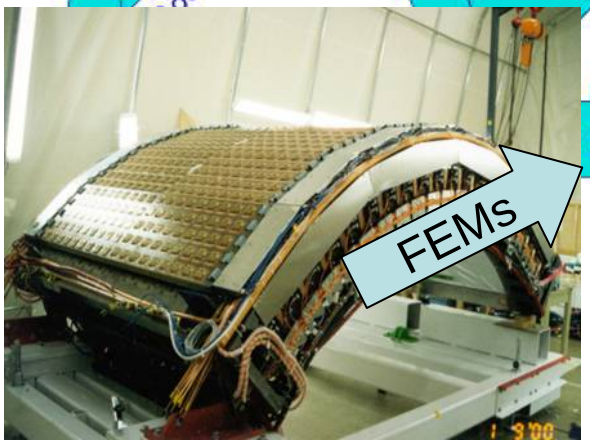



PHENIX, circa Jan 1999

# PHENIX DAQ Overview

A **granule** is a detector subsystem, including readout electronics

A **partition** is one or more granules that receives the same triggers & busies



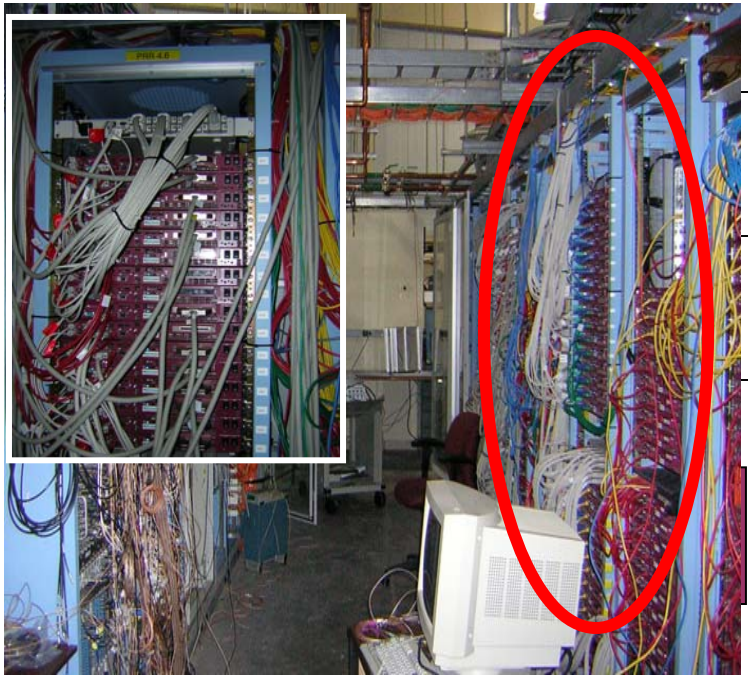


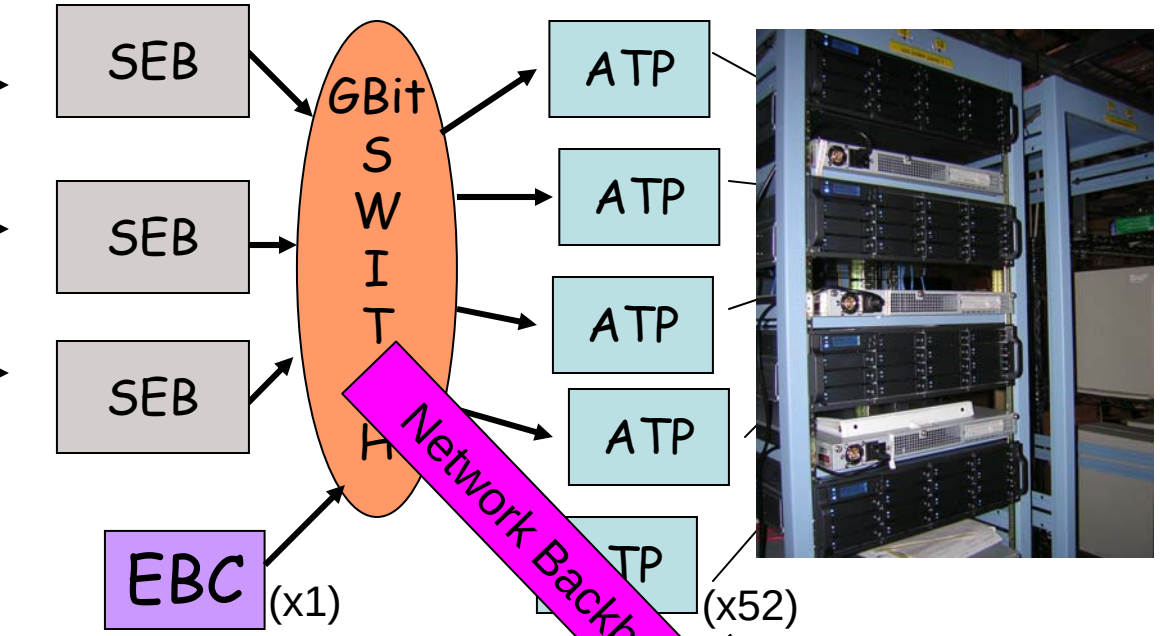


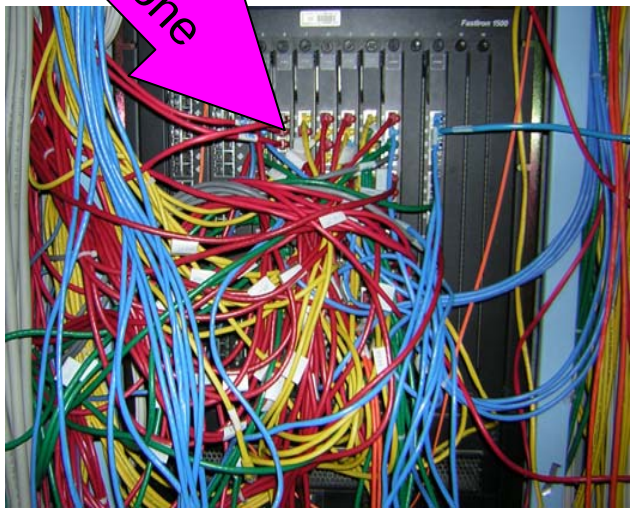
- 14 Granules
- 38 DCM Groups / partitioners
- 1-15kB / Group
- 6 data loggers
  - 32 TB local disk space

1 Data Collection Module Group => 1 Partitioner Module

# PHENIX EvB at-a-glance





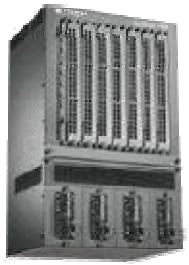



- Suite of multithreaded client/server apps
- **SubEvent Buffer** (SEB): receives data from Data Collection Modules
- **Event Builder Controller** (EBC): receives event notification, assigns events, acknowledges event completion
- **Assembly Trigger Processor** (ATP): polls SEBs for subevents, optional stage for Level-2, writes assembled events to local storage
- All connected by gigabit backbone

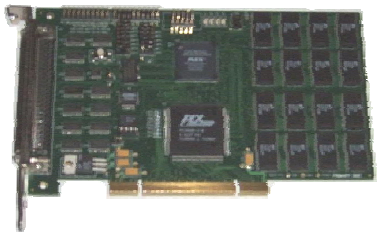
# Key Features of EvB

- Software and Design
  - Pull architecture with automatic ATP load-balancing
  - Extensive use of C++, STL, and Boost template library
  - Slow control via CORBA (IONA's Orbix)
- 105 1U Rack-mounted dual CPU x86 servers
  - 1.0 GHz PIII & 2.4 GHz P4 Xeon
  - Gigabit NIC on PCIX (Intel PRO/1000 MT Server)

- Foundry FastIron 1500 Gigabit Switch
  - 480 Gbps total switching capacity
  - 15 Slots, 12 in use (includes 128 Gigabit ports)



- JSEB: custom-designed PCI card



  - Interface between EvB and incoming data stream
  - Dual 1 MB memory banks (allows simultaneous r/w)
  - Programmable FPGA (Altera 10K)
  - Latest firmware enables continuous DMA – up to 100 MB/s I/O

# PHENIX EvB's is an evolutionary tale

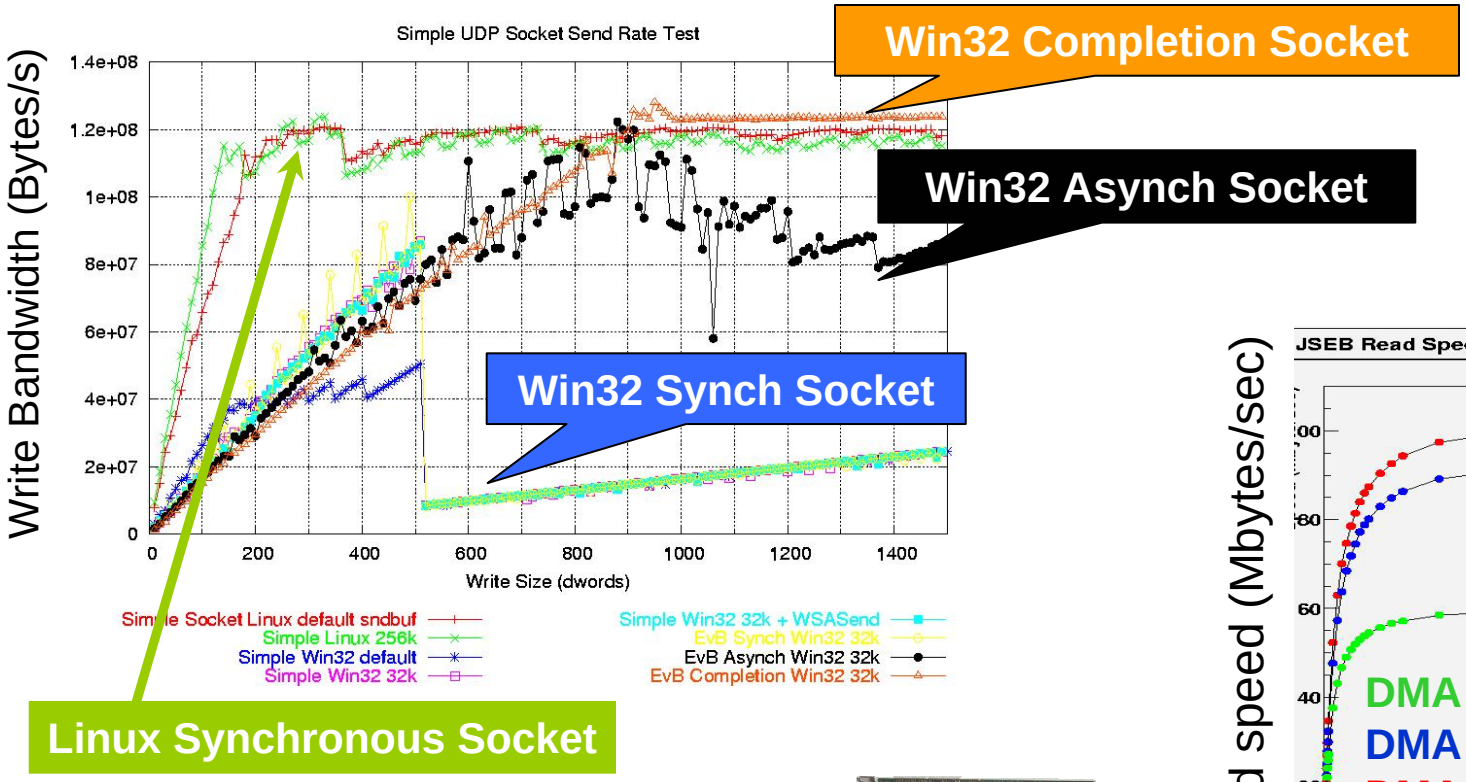
- Circa 2000: Started as 24x4 system at 20 Hz
  - TCP data transfer DCMs => SEBs
  - EvB using Win32 + ATM
    - EvB development started in 1997: Linux new and ATM poorly supported in it, GbitE but a gleam in the netops' eye.
- Run3 (2002-3): GbitE switch deployed for data transfer (control remained in ATM domain)
- Run4 (2003-4):
  - Introduced LZO compression in data stream (at ATP)
  - Completed conversion to GbitE (for data AND control)
    - Multicast used for event requests and event flushing
    - UDP used for data communication and remaining control tasks
- Run5 (2004-5): continued growth to 40x40 system at >5 kHz
  - Completed port from Win32 to Linux (FNAL SL3.0)
  - All communication switched to TCP
- Overarching driver: need to keep EvB in operation while evolving its design to exploit advancing technology
  - Now have 5 years of development history under our belts

# Run5 Goal: 5 kHz “or bust”

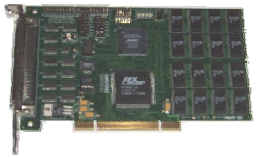
- Strategic identification of bottlenecks continued from Run4
  - Requires targeted testing, ability to profile in “real time”
- Major decision to port 90k+ lines of code to Linux
  - Con: Darling of the physics community *does* have its limitations
  - Pro: Superior performance far beyond what we expected
- Jumbo frames: used in Run4, but dropped for Run5
- Level-2 triggers to be used as filter on *archived data*
  - Last used in ATP in Run2
  - Ability to log data at >300 MB/s “obsoletes” Level-2 in ATP

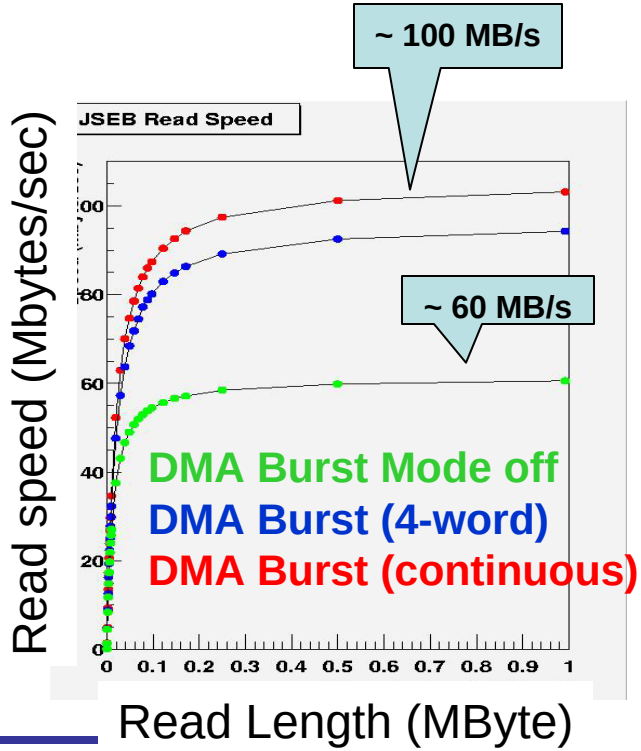
# Performance bottlenecks

- What is our primary limitation? Our ability to:
  - Move data across the network
  - Move data across the bus (PCI or memory)

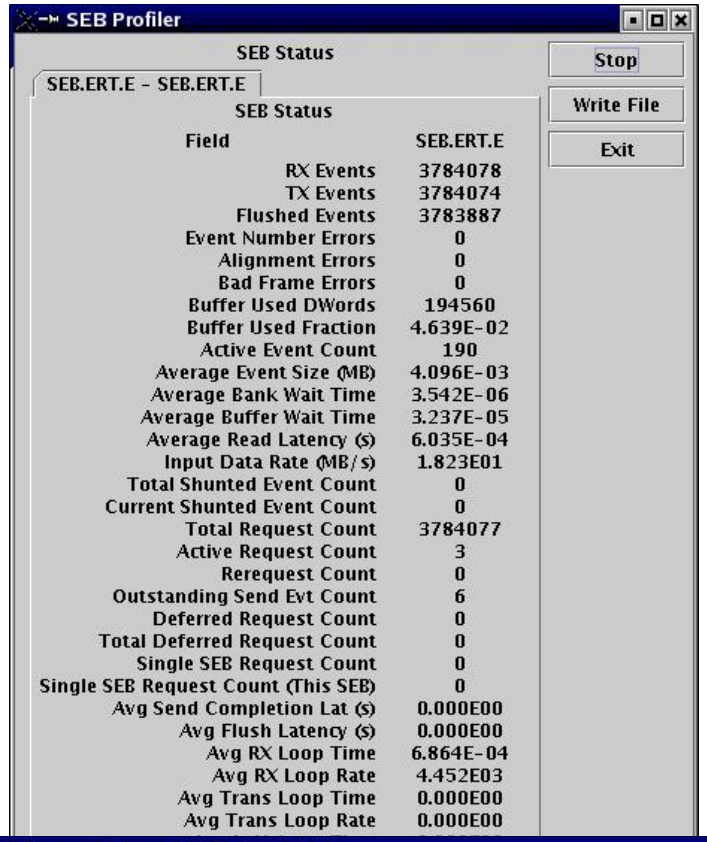



What about the JSEB?

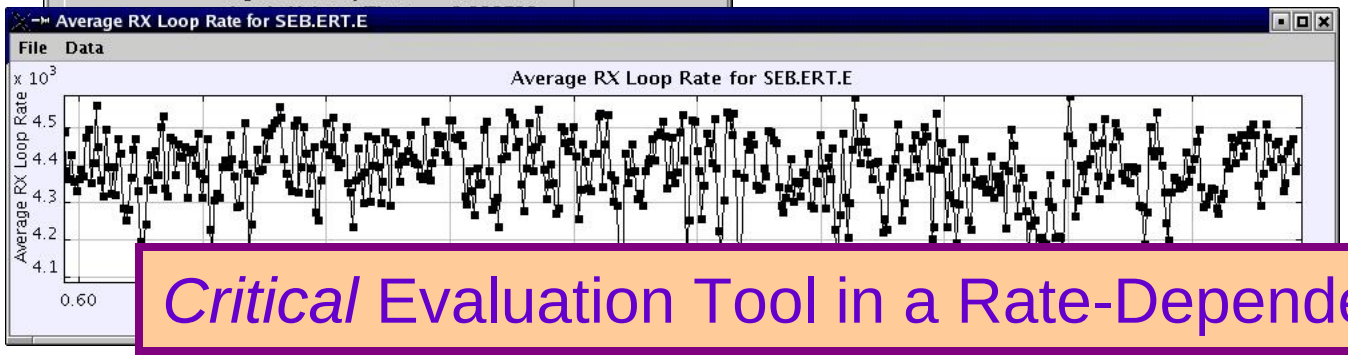


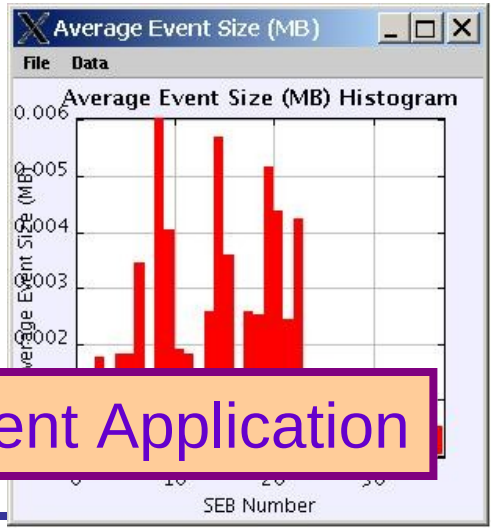


# Monitoring Performance




- Each component keeps track of various statistics
- Data served via CORBA calls
- Java client displays stats in “real time”
- Strip charts display data as function of time
- Histograms display data as function of component



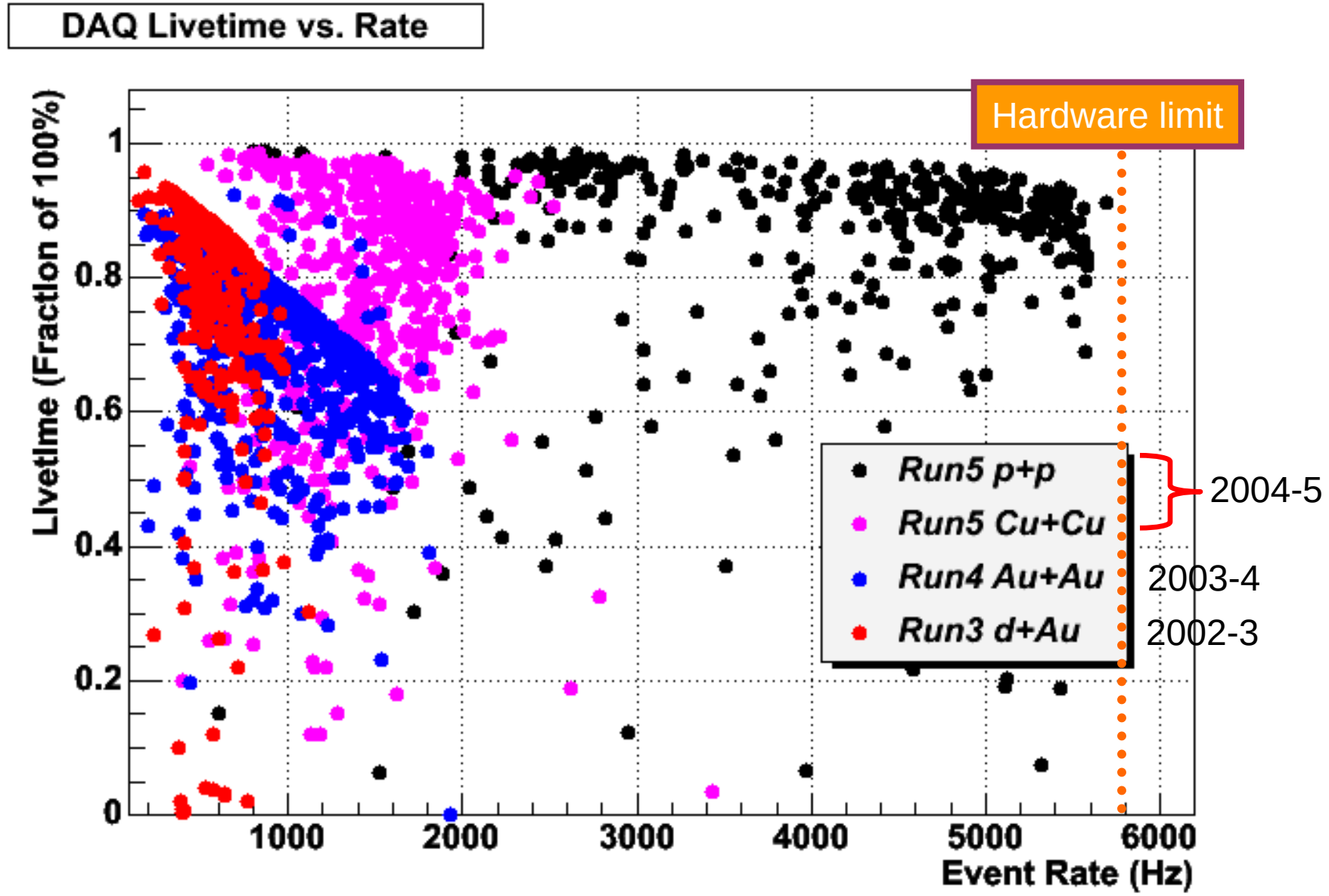




*Critical* Evaluation Tool in a Rate-Dependent Application

# No lack of technical issues…

- [Vanilla] Linux: easier said than done
  - No asynchronous [socket] I/O (later this became a feature)
  - No native system level timers (available to users)
  - No OS Event architecture (as available in Win32)
  - Not all of TCP/IP stack necessarily implemented
    - e.g. TCP high watermark doesn't work – but also silently fails!
  - Proprietary Jungo (www.jungo.com) driver (for JSEB) exposed insidious memory mapping bug in 2.4 kernel
    - Learned the true meaning of kernel "panic" as beam time approached…
- Jumbo Frames: Incompatible with NFS drivers (BUT, unnecessary with Linux performance)
- Very little sensitivity to parameters of Intel PRO/1000 driver
  - Interrupt coalescence, Tx/Rx descriptors, etc.
- Spin Locks vs. Mutexes: choose carefully!

# Figure of Merit: Livetime vs. Rate



# Summary

- RHIC and PHENIX face unparalleled requirements of rate and data volume driving the design and implementation of the DAQ and the Event Builder
- The PHENIX Event Builder is a fully object oriented system that is critical to the success of our DAQ
  - All control and real-time monitoring achieved via CORBA
  - Introduces nearly zero overhead
  - Steady-state operations at >~ 5 kHz and >~ 400 MB/s in Run5
  - Can do Level-2 triggering
    - Ability to archive 100's MB/s makes this unnecessary (for now)
    - Future performance limitations may require it
  - Significant evolution over the past 5 years, successfully integrating new technologies as they have emerged
- The world's fastest and highest data volume DAQ, most likely will continue to be until the LHC era

Thanks to all past and present PHENIX EvB developers:
B. Cole, P. Steinberg, S. Kelly, S. Batsouli, J. Jia, F. Matathias
Additional thanks to W. A. Zajc for excellent background material