RHIC relativistic heavy ion collider

RHIC & ATLAS Computing Facility
at Brookhaven National Laboratory

ATLAS EXPERIMENT

RACF Site Report

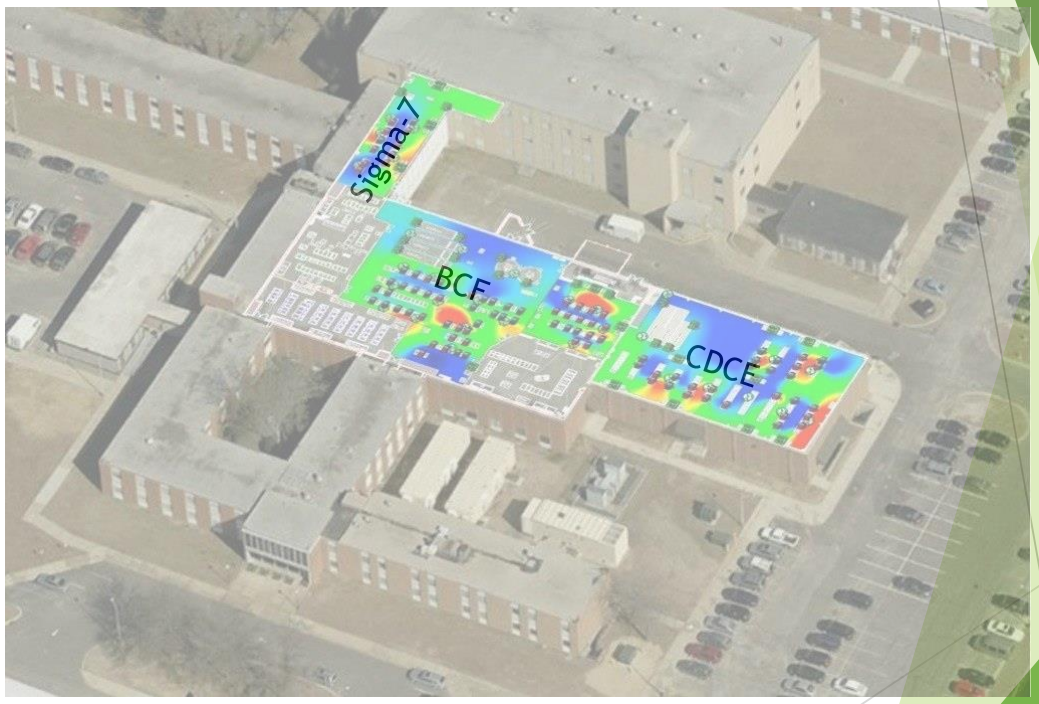
William Strecker-Kellogg

HEPiX—October 2016

BROOKHAVEN NATIONAL LABORATORY

RHIC/ATLAS Computing Facility Overview

- Part of BNL physics department, started to provide computing services for RHIC experiments
  - STAR and PHENIX have ~15kCPU and 10's of PB of data each
- USATLAS Tier 1 Facility
  - Largest ATLAS Tier 1 in the world
  - Around 20kCPU of compute
  - Provides 10's of PB of storage on dCache, HPSS, and CEPH



Evolving Computing @ BNL

- Computational Science Initiative (CSI)
  - Lab-management supported group with goal of centralizing scientific computing at BNL
  - Leverages expertise in RACF and other groups to create synergy and enable more advanced computing
  - Multiple components
    - Center for Data-Driven-Discovery (C3D)
    - Scientific Data Computing Center (SDCC)
      - Comprised of RACF
    - Computational Science Lab (CSL)

Evolving Computing @ BNL

- Institutional Cluster
  - 108 2 Dual-Core-GPUs + 2 head nodes
  - 4xEDR 2-level fat-tree Infiniband
- Expanding to 200 clients soon
  - Ordering ~90 more nodes
- GPFS storage cluster
  - 4xFDR Infiniband—24 GB/s→42GB/s
- SLURM batch system
  - Node-level scheduling
  - Shares determined by CSI
- New Knight's Landing Cluster
  - For LQCD, shared with CSI
    - 144 Nodes, KNL CPUs
    - Dual-rail OmniPath interconnect
    - Unique in the world (as of Oct 5)
  - OmniPath←→Infiniband gateway for GPFS connectivity
- See Alex's CHEP talk for details about our benchmarking and testing of OmniPath

Blue Gene Room

- RACF/CSI are new tenants
- 1000 ft^2 of usable space
  - Name derived from hosting IBM's Blue Gene systems (for HPC applications) in recent past
- Upgraded in early 2016 to support 400 kW of computing equipment
  - Sufficient to handle 20 high-power density enclosures (20 kW/rack)
  - Will house IC, KNL and other HPC-like equipment

SLURM Batch System

Deployment

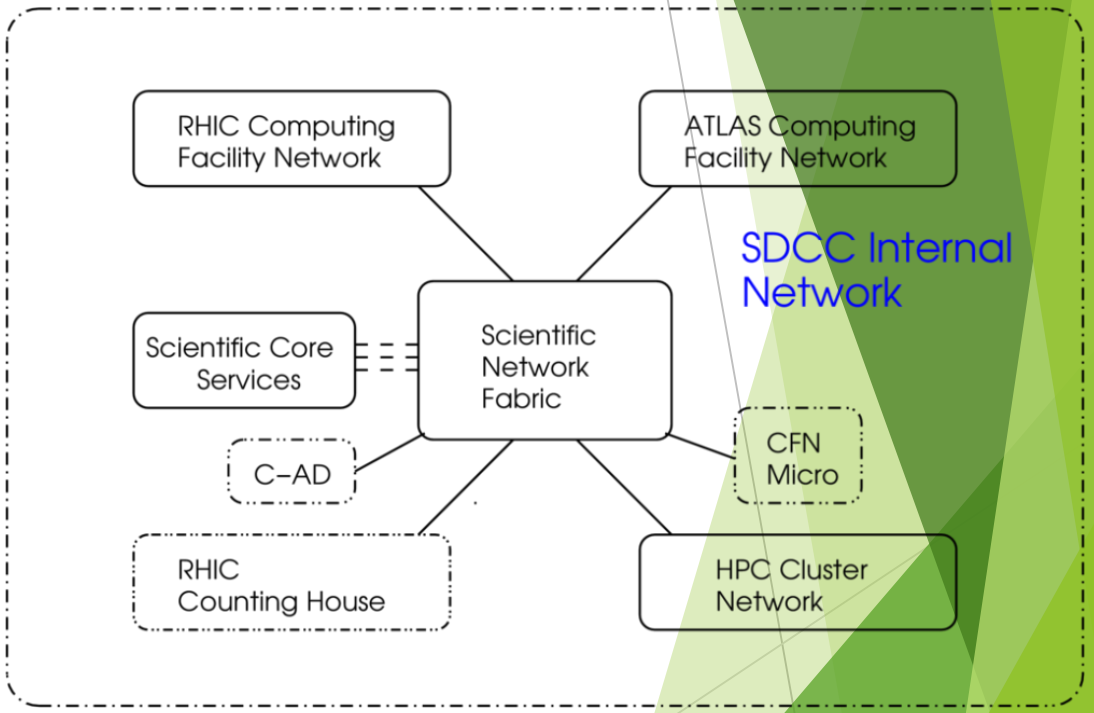
- Evaluation done over summer
- Fully puppet-driven deployment
- Scales well, integrates well
  - Issues with simultaneously monitoring usage and fair-share (decaying usage)
  - Able to schedule at node, cpu, core, thread, & GPU/KNL granularities and more

Integration

- Work done to tie-in to existing HTC resources
- Utilize HTCondor's grid universe & job-router
- Able to send eligible condor jobs transparently into SLURM's backfill-queue
  - Josh Barnett, my SULI intern
  - See Chris Hollowell's CHEP talk

Network Re-architecture

- HPC-Core→"Scientific Network Fabric"
  - Developed as central hub for scientific data-intensive workflows (3Tbps currently)
  - Plan is to move all lab scientific computing and data gathering to spokes off of Scientific Network Hub
    - High N←→N bandwidth (100's Gbps)
  - Detectors / Clusters / Processing-Farms
    - All able to exist in separate security zones
  - "Outside" of BNL campus
    - ATLAS moving first due to IPv6 requirements
    - Working on Cyber-Security clearance



HPSS Tape Systems

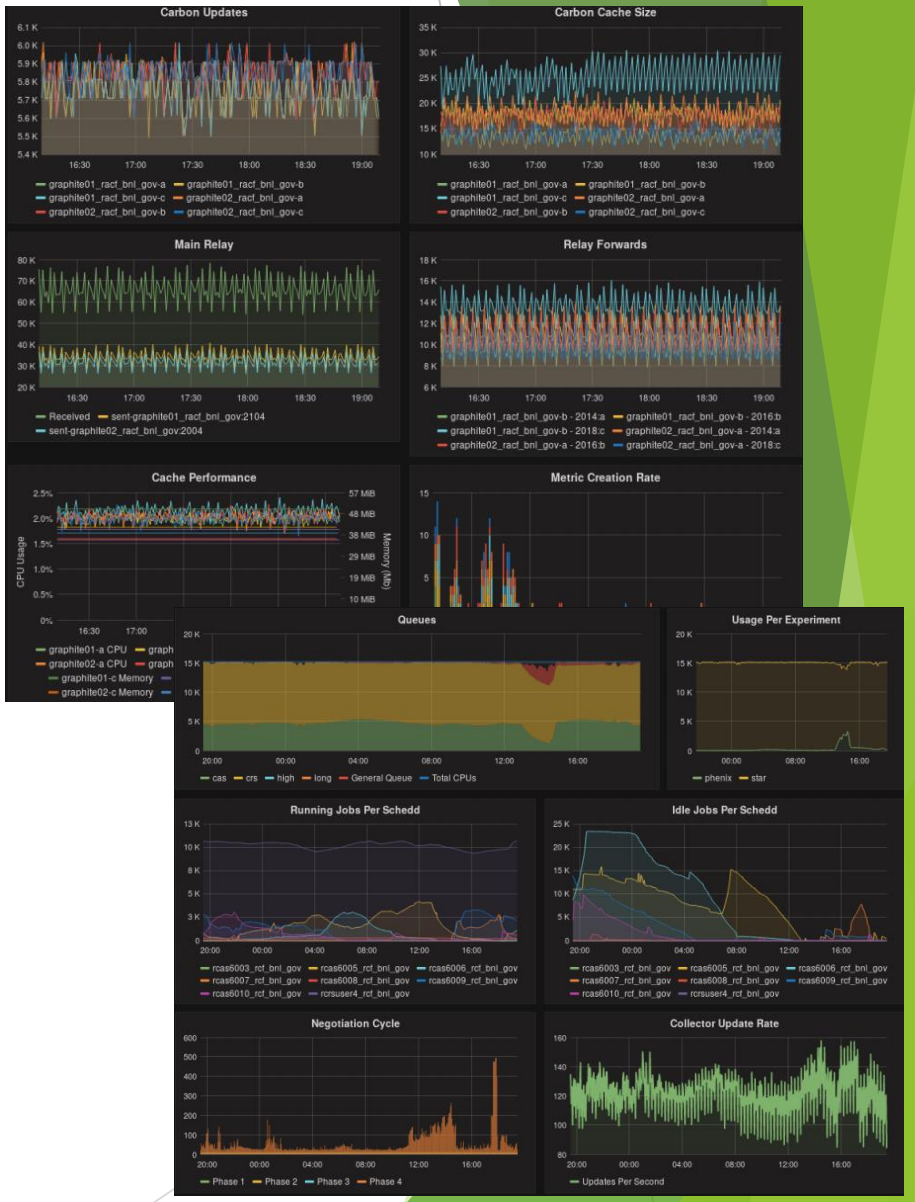
- Planned update to 7.4.3 for Nov 2016
  - RHEL7 client support
- ~90PB of data on tapes
  - 65,000 tapes in 8 SL8500 libraries
  - 1.6PB disk cache
- Core & movers on RHEL6

Puppet / Certificates

- CRL files downloaded at first run, 5 year “expiration”
  - Clients older than 5 years will break once CRL expires even if their cert is OK
- Our puppet infrastructure is now >5 years old
  - First mass-issued certs expired Oct 1
    - Certs are pre-generated on the master for easy rebuild of farm nodes
  - CA good for 20 years, cert generation still defaults to 5 years
    - CLI option on-generation or config-file entry fixes this
- 3.8 Rumored to go EOL by end of this year
  - Tested “future-parser”
  - Planned migration is still a long way off
    - Is anyone on 4.x yet!?

Monitoring & Data Analytics

- Evaluated several different time-series-data warehouses
  - OpenTSDB, Prometheus, InfluxDB, Graphite
- Graphite
  - Open source, clusterable, well supported, simple architecture
  - Requires hardware with good IOp/s capacity
- Grafana
  - Frontend to most every popular data warehouse
  - Powerful query-builder and templating engine
  - Supports users with roles and multiple orgs
- See my talk later this week



Linux Farm

- New RHIC Machines
  - 28 Dell R720xd systems
  - 2 Broadwell E5-2680v4 2.4 GHz CPUs (56 logical cores total)
  - 128 GB DDR4 2400 MHz RAM
  - 2 120 GB 6Gbps SDD SATA drives
    - Fixes GPFS root-drive-latency issue
  - 12 3.5" 4 TB 7200 RPM 6Gbps SATA drives
  - Hardware RAID5 (PERC H730P)
- New ATLAS Machines
  - 100 Dell R430 systems
  - 2 Broadwell E5-2690v4 2.6 GHz CPUs (56 logical cores total)
  - 128 GB DDR4 2400 MHz RAM
  - 4 3.5" 2 TB 7200 RPM 6Gbps SATA drives
- SL7—evaluated and functioning
  - No upgrade plans yet

HTCondor

- Version 8.4.x still latest stable release (as of Oct 6)
- Major issue with Schedd halting for up to an hour
  - No disk/network IO of any kind
  - No syscalls
  - GDB shows a mess (STL)
- Schedd spending up to an hour recomputing internal array using autocluster→jobid and the reverse mapping
  - Ticket #5648
- Jobs would die after their shadows couldn't talk to their schedd that went dark
- After day of debugging, code fix was implemented
  - Built & tested at BNL, max computation time reduced 1h to 2s
  - Thank you to Todd & TJ
- Fix released in 8.4.7

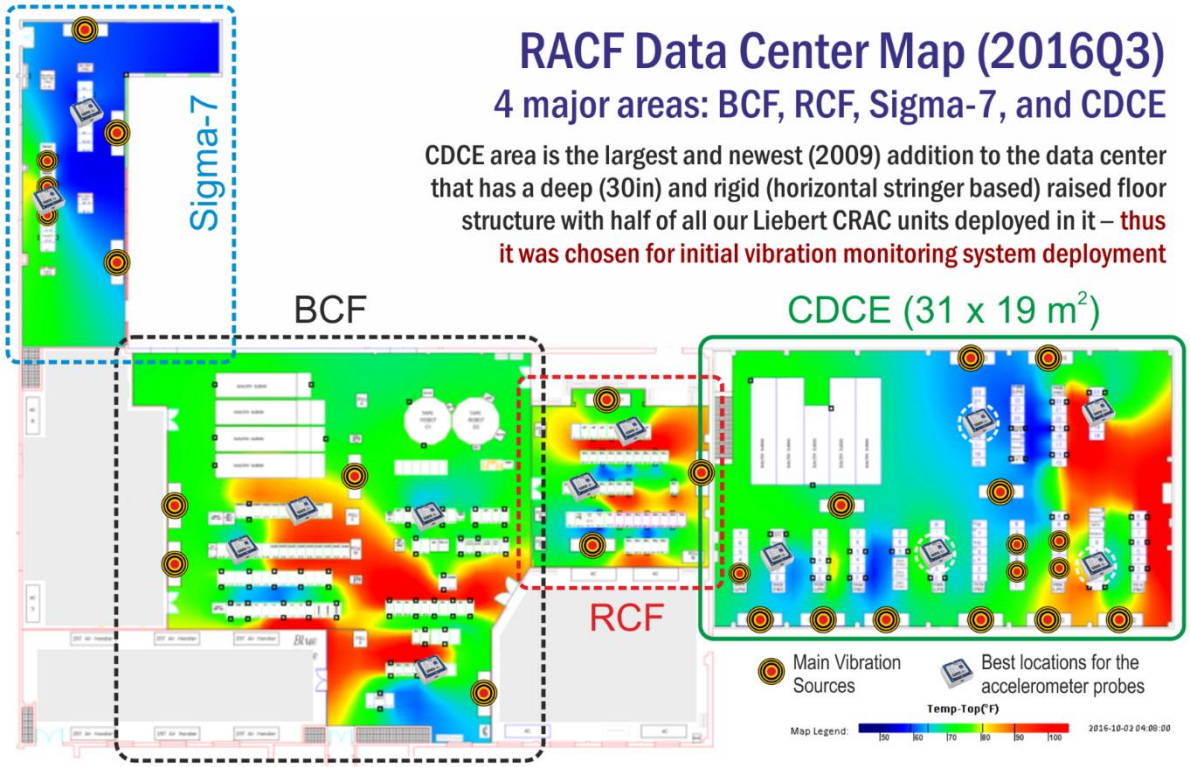
Ceph

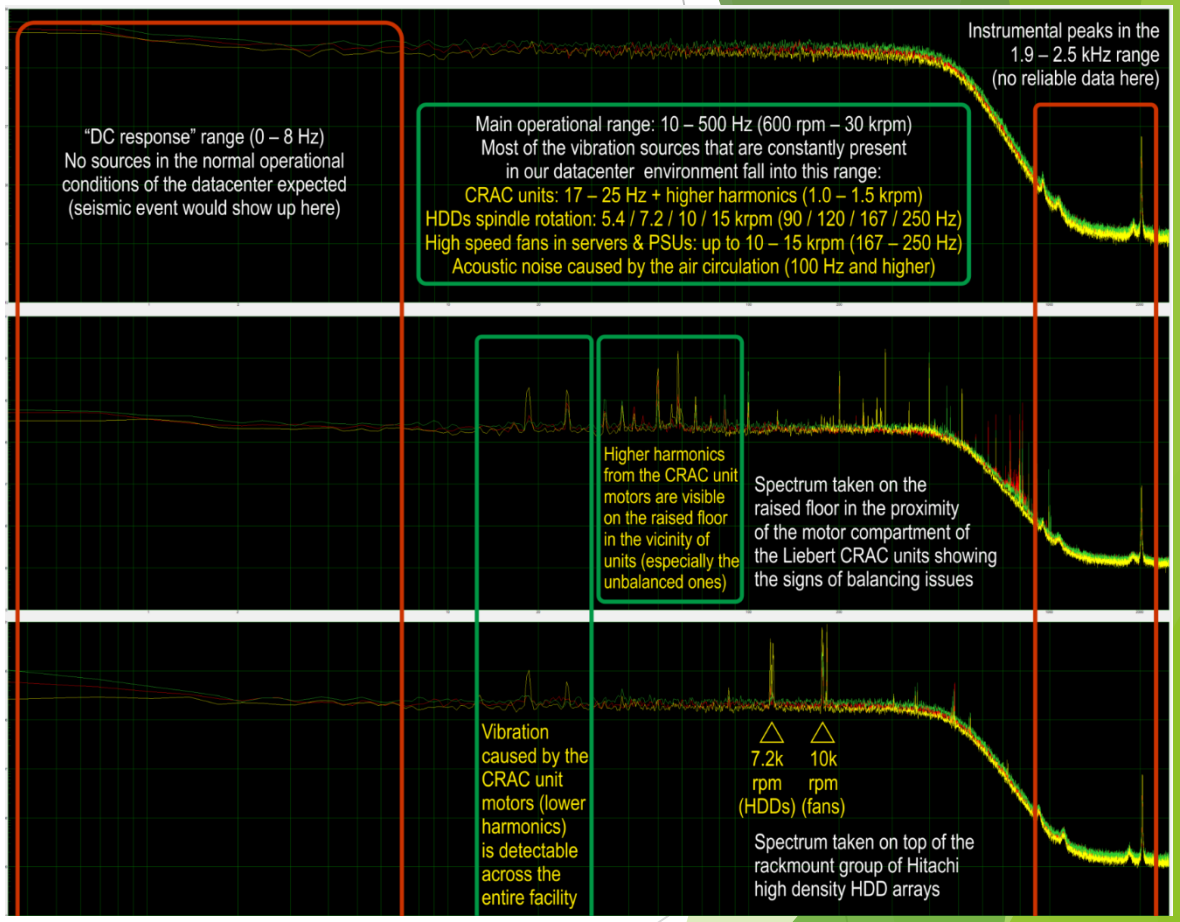
- Two Ceph clusters of 1.2 PB/0.4 PB (raw/usable) and 1.8 PB/0.6 PB capacity – “old” and “new” cluster.
  - Run hardware retired from BNL ATLAS dCache (3.7k 1 & 2 TB drives in 13 racks)
  - Added distributed (6 TB) and centralized (10 TB) NVMe devices to the new Ceph cluster to boost write performance
    - 10k+ ATLAS Event Service clients connecting from the WAN—ongoing, no results yet
- Significant software / storage layer reconfiguration in June, to cope with the ATLAS ES workloads coming to BNL from NERSC machines that exceeded the design specifications of both clusters by one order of magnitude!
  - The “new” cluster capability scaled from 2k→24k concurrent connections plus and 1.1 GB/s of aggregated WAN traffic to the RadosGW/S3 subsystem was demonstrated with the ANL→BNL S3 tests in July
  - Plan to double capacity (to 2 PB usable) early in 2017 and further increase the I/O performance by using the cache tiering mechanism and low latency NVMe PCIe SSD devices mentioned above (Intel P3700 series).
- See Alex’s talk “Ceph Based Storage Systems at the RACF” for more details

Vibration Monitoring

- RACF cooling done with Liebert CRAC units pressurizing a raised floor
  - CRAC units can have failing bearings that require periodic maintenance
- Raised floor allows long distance propagation of mechanical vibration from CRAC units, disk arrays, and even cooling fans in compute nodes
  - Vibration may not be noticed immediately in the absence of constant monitoring
- 2016Q2: design and deploy a high sensitivity vibration monitoring system to provide us with means to understand the interplay between various sources of vibration and monitor for excessive levels
- 2016Q2-3: evaluated several platforms and chose Lansmont tri-axial MEMS accelerometers (and the SaverXware software that comes with it)
  - Aug-Sep 2016: first order provided 3 such devices
- Raw data storage and the software for analyzing the vibration sources is still under development with expectations to have it in place by 2017Q1
  - By adding control and data-readout server infrastructure we can scale the system horizontally up to 16 probes in total.

Vibration Monitoring





Questions?

Thank You!

Information here compiled with help from the entire RACF staff, special thanks to Chris Hollowell, Tony Wong, & Alexander Zaytsev