Disaster Recovery and Data Centre Operational Continuity

J. Caballero Bejar, C. Caramarcu, J. De Stefano Jr., M. Ernst, J. Fetzko, C. Gamboa, C. Hollowell, J. Hover, S. Kandasamy, M. Karasawa, Z. Liu, S. Misawa, W. Strecker-Kellogg, O. Rind, J. Smith, T. Wlodek, A. Wong, D. Yu, A. Zaytsev, X. Zhao

Brookhaven National Laboratory, Department of Physics, P.O. Box 5000, Upton, NY 11973
E-mail: tony@bnl.gov

Abstract. The RHIC and ATLAS Computing Facility (RACF) at Brookhaven Lab is a dedicated data center serving the needs of the RHIC and US ATLAS community. Since it began operations in the mid-1990's, it has operated continuously with few unplanned downtimes. In the past 15 months, Brookhaven Lab has been affected by two hurricanes and a record-breaking snowstorm. In this presentation, we discuss lessons learned regarding (natural or man-made) disaster preparedness, operational continuity, remote access and safety protocols, including overall operational procedures developed as a result of these recent events.

1. Introduction

The RHIC and ATLAS Computing Facility (RACF) at Brookhaven National Lab (BNL) operates a large-scale multi-purpose computing facility 24x7, serving a geographically diverse, worldwide scientific community that participates in various projects in which BNL is involved. The major components of the RACF are the 42,000-computing core Linux Farm, the 15 PB disk storage system, the 35 PB robotic tape storage facility, and the high-availability general computing infrastructure. The systems are connected together by a high-speed, 100-Gbps capable backbone with over 3500 active ports.

The RACF has operated continuously since the mid 1990's, but in the past 15 months, the Long Island area has been affected by two hurricanes and a blizzard, paralyzing the area roads, damaging the area's power infrastructure and hindering physical access to the BNL. These natural events underscored the need for the RACF's on-going efforts to sharpen (and accelerate in some instances) its contingency plans to minimize disruptions to facility operations and to establish procedures that facilitate staff remote and physical access to data center systems that host essential RACF services. This paper outlines the steps taken or being taken in the areas of data center infrastructure, survivability of critical services and disaster management.

2. Data Center Infrastructure

The RACF is spread out in three distinct physical areas: BCF, Sigma-7 and CDCE, totaling 13,000 ft^2 (1200 m^2). The RACF has occupied the BCF since 1996, while the Sigma-7 and CDCE spaces have been exclusively used by the RACF since 2008. Figure 1 shows the redundant nature of the power infrastructure at the facility level. Individual floor PDUs provide appropriate redundant power to almost every rack in the RACF. The single central source of cooling supports individual CRAC units, which provide cooling redundancy at the rack level. In 2013, the RACF initiated a project to hook up selected CRAC units in the Sigma-7 space to the diesel generator for power redundancy, and another project to hook them up to domestic water as back-up for the chilled water tower.

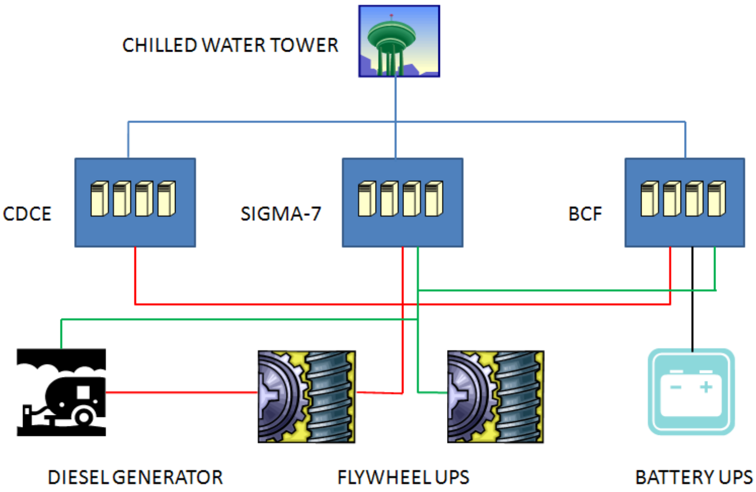


Figure 1: General layout of the data center physical infrastructure

To estimate the usefulness of this power redundancy scheme, we collected data on power fluctuations or losses since 2003, when the UPS monitoring software was initially installed at the RACF. We have an average of 4 events per year (see Figure 2), mostly due to weather events (thunderstorms and lightening) in late spring and summer. The spikes in 2008 and 2009 were a result of power interruptions due to the renovation of the Sigma-7 area and construction of the CDCE area.

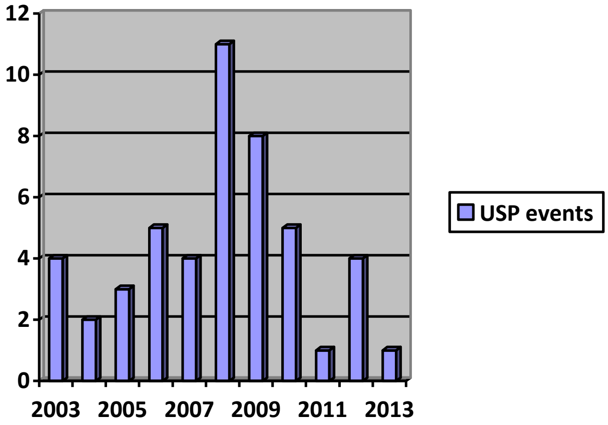


Figure 2: Number of times per year when UPS power is activated

3. Survivability of Critical Services

This section focuses on two issues: BNL-internal developments to insure the survivability of the RACF's critical services, and ESNET efforts to provide an alternate path from BNL to the WAN.

The RACF enhances the resiliency of critical services with security measures, redundancy of physical hardware (cold and hot spares) and usage of virtualized hosts to achieve high availability. In addition to BNL-wide mechanisms for timely desktop patching, intrusion detection, phishing and firewalls, the RACF selectively employs several other tools to improve the user experience, minimize downtime and increase the reliability and resilience of its hardware and services: a) a Puppet [1]-based configuration management to quickly build and rebuild hosts, b) automatic termination of mis-behaving jobs with systemic memory leaks, infinite loops, excessive local disk usage, etc, and c) rebootless patching of Linux-based systems via KSplice [2].

ESNET provides connectivity from the ATLAS detector to BNL via its Manhattan hub. Access to the LIMAN (Long Island Metropolitan Area Network) has been enhanced by the recent completion of the 100 Gb-capable alternate path between BNL and Manhattan (see Figure 3). The redundant path increases the resiliency of the connectivity between BNL and the external world.

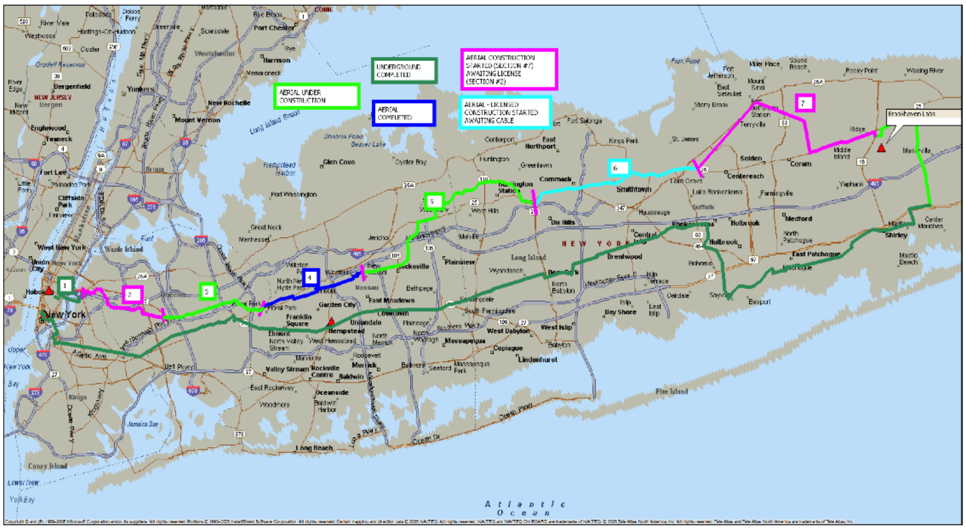

AERIAL UNDER CONSTRUCTION
UNDERGROUND COMPLETED
AERIAL COMPLETED
AERIAL CONSTRUCTION STARTED (SECTION #7) AWAITING LICENSE (SECTION #2)
AERIAL - LICENSED CONSTRUCTION STARTED AWAITING CABLE
Brookhaven Labs
New York
N E W Y O R K
Long Island Sound
Atlantic Ocean
1
2
3
4
5
6
7

Figure 3: Existing (bottom green line) and recently completed alternate LIMAN path (top multi-colored line) connecting BNL (small red triangle near right upper corner) to Manhattan (left side of figure)

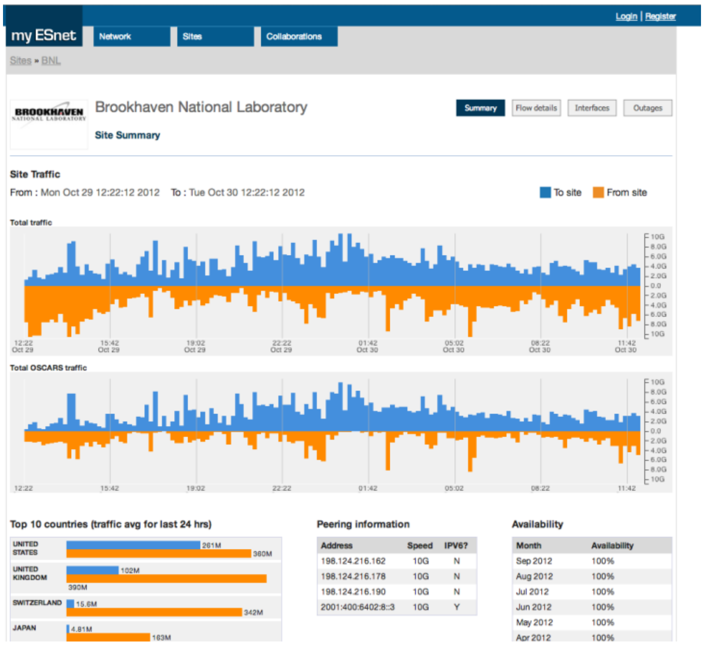
4. Disaster Management

BNL's location on Long Island at the north-eastern edge of the U.S. mainland is vulnerable to Atlantic hurricanes in late summer/fall and to nor'easters (hurricane-like events characterized by high winds and heavy snow or rain) in winter. On average, 1-2 such weather events affect BNL per year with varying intensity. Man-made disasters such as widespread power blackouts and localized accidental power or cooling outages occur less frequently.

The 2012 Hurricane Sandy devastated the electrical power grid in the New York and New Jersey areas, and the 2013 Superstorm Nemo (see Figure 4) paralyzed local roads for days. In both cases, the RACF decided to operate throughout the weather events, and BNL never lost vital (power and cooling) infrastructure support (see Figure 5).

The decision was based on our previous operational experiences. The effort required to power down (2-3 hours) was disproportional to the effort required to restore services (9-10 hours), and procedures developed by the staff allow the RACF to remotely power it down in case vital infrastructure is lost. We outline the procedures, lessons learned and still-to-do work below.

Figure 4: Stranded cars clog up area roads following Superstorm Nemo



Brookhaven National Laboratory

Site Summary

Site Traffic

From : Mon Oct 29 12:22:12 2012 To : Tue Oct 30 12:22:12 2012

Peering information

Address	Speed	IPV6?
198.124.216.162	10G	N
198.124.216.178	10G	N
198.124.216.190	10G	N
2001:400:6402:8::3	10G	Y

Availability

Month	Availability
Sep 2012	100%
Aug 2012	100%
Jul 2012	100%
Jun 2012	100%
May 2012	100%
Apr 2012	100%

Figure 5: WAN traffic at BNL during Hurricane Sandy

4.1. Lessons Learned

The first consideration in disaster management has been whether a pre-emptive shut down of the facility is necessary. The RACF philosophy has evolved towards keeping the facility running mainly because improvements in infrastructure have made the RACF services more resilient, but there is no default course of action. Decisions are made by RACF management in consultation with BNL site management, taking into consideration the severity of the event.

The ability to remotely access the RACF is essential for staff to keep the RACF operational. All staff have high-availability access to administrative gateways via VPN, and some staff also have BNL-provided cell phones with tethering capabilities for alternate internet access through the telecommunication providers' network. This helps bypass service interruptions from internet providers, as it happened during Hurricane Sandy.

The RACF has built an automatic monitoring and alert system over the years. The system monitors power, cooling and facility services. Alerts are sent via email and text messages to cell phones. The environmental monitoring system is currently being improved by migrating from wired one-dimensional temperature probes to a system with wireless 3-D probes integrated with a real-time heat map by Synapsense [3]. See Figure 6. For certain systems (such as the Linux Farm), cooling or power failure triggers an automatic shutdown (after a grace period) to decrease the chance of an electrical fire. The shutdown can be surgically targeted to each of the three physical areas (BCF, Sigma-7 and CDCE.

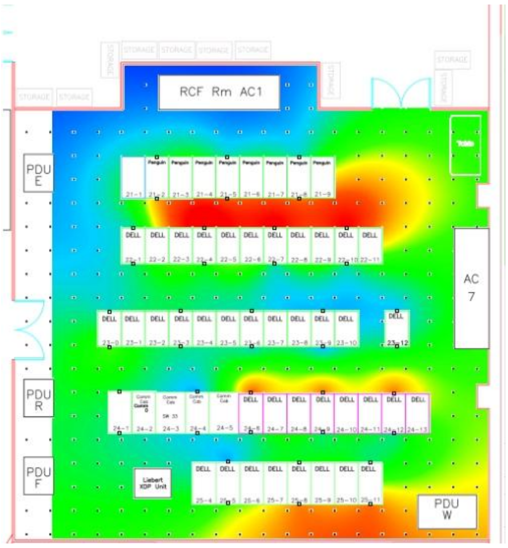


Figure 6: Synapsense real-time heat map of parts of the BCF space.

4.2. *Future Plans and Conclusions*

BNL has long-term plans to build an auxiliary chilled water tower and to provide generator back-up power to the main chilled water tower. This would address the single point of failure as shown in Figure 1. Additionally, the RCF will connect additional CRAC units to the diesel generator in 2014 and beyond, budget permitting.

Although tools to assist in disaster recovery and insure operational continuity have been gradually added to the RACF in recent years, only ad-hoc emergency-handling procedures existed prior to 2011. The RACF is developing a more formal process to address operational continuity, and the facility is reviewing the criticality of services, the role of essential and supporting staff and a systematic process to organize and define future emergency handling procedures.

Operational continuity is often an overlooked subject matter in data centers, but recent events have shown it deserves more attention. It has spurred BNL to invest in a more resilient data center infrastructure, motivated the RACF to update operational procedures and integrate existing, separate software packages into a coherent system. Our extended loss of physical access to the data center has sharpened the RACF's focus on providing and maintaining robust remote access and communication tools to the staff. As a result, operational continuity has improved efficiency of facility services and increased productivity of facility users.

References

[1] Puppet: www.puppetlabs.com
[2] Ksplice: www.ksplice.com (KSplice was recently acquired by Oracle)
[3] Synapsense: www.synapsense.com