2016 Operations Stay Treat

Improving System Reliability

A Case Study of Accelerator UPSs

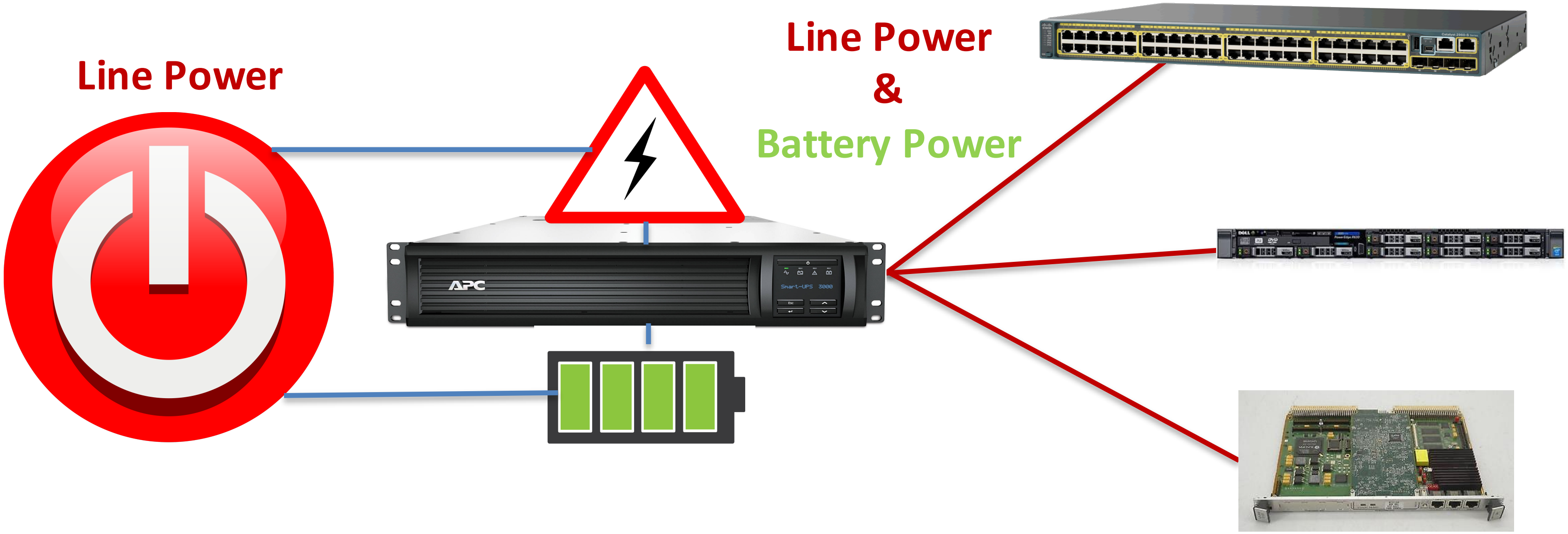
Anthony Cuffe

Jefferson Lab

Summary

- Typical UPS Usage
- Backstory
- Initial Status and Analysis
- Reliability Study
- Maximizing Availability, Reliability and Manageability
- Automation, Monitoring and Logging Tools
- Results
- Key Points
- Q & A

Typical UPS Usage in A.C.E.



Protection from Power Loss and Poor Line Quality for:

- Critical Workstations and Servers
- Network Infrastructure
- RF, Vacuum and other critical IOCs

Backstory *What just happened?*

- ⦿ Historically our group maintained around 10-15 rack-mounted, enterprise grade UPSs.

  … This was manageable without much effort.

- ⦿ After a re-org, our group became the proud owners of 100+ UPSs.

  … And most of them weren't shiny and new!

- ⦿ After staging through denial, anger, bargaining, and depression, we accepted it.

  … This is how we got over most our grief.

Initial Status *The Reality of Square One*

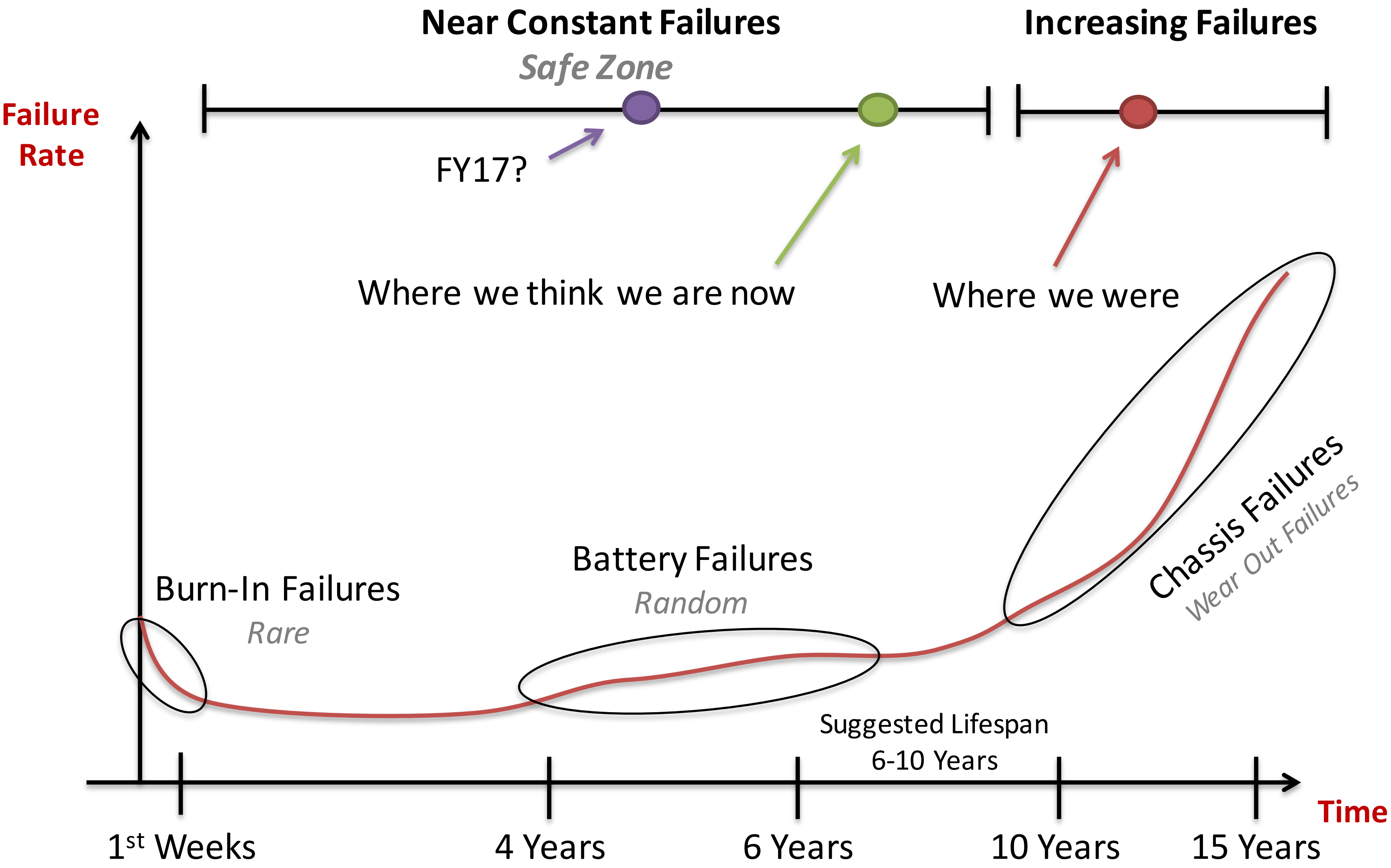
- Responsible for 100+ UPSs *(and growing!)*
- ~10 were outright failed
- ~10 were in a faulted state
- 40+ were found to be 8-16 years old *(5 year lifetime)*
- Configuration was manual, onerous and inconsistent
- Deployment took a very long time *(1++ hour)*
- Monitoring was impossible
  - *Sheer volume of error messages (X,000+/per day)*
  - *System age*
- Our attitudes towards the system were awful
  - *Dread*
  - *Apathy*
  - *A few resumes were tuned up*

Jefferson Lab

Initial Analysis *Is the system worth improving?*

- Is the system's functionality still necessary?
  - *Servers, network gear and critical IOCs need to survive power interruptions.*
  - *UPSs provide the only low-cost solution to fulfill this need.*
- Is the benefit worth the cost and staff time?
  - *~ $50K to make the overall system reliable and ~$10-20K/year to maintain*
  - *The recovery costs and reduced beam availability associated with frequent power-outages greatly exceeds the cost of maintaining UPSs for critical systems.*
- Is it possible to improve the cost/benefit ratio?
  - *We felt we could decrease the time spent maintaining these systems by applying practices common to our other systems.*
  - *The benefit could be improved with better reliability and selective application.*
  - *Costs could be spread over years with a well maintained system.*
- Do you have management buy in?
  - *We had to convince management that the benefit would be worth the time, effort and procurement dollars.*
  - *It is impossible to improve or maintain a system without resources!*

Reliability Study *UPS Failure Rate*



* Curve estimated based on experience, dependent on operating environment

Improving Availability

Availability Depends on Reliability and Manageability

$$AVAILABILITY \approx \frac{MTBF}{(MTBF + MTTR)}$$

- **Maximize Reliability**
  - *Make **MTBF** as big as possible*
  - *Mitigate unreliability where **MTBF** cannot be reduced*

- **Maximize Manageability**
  - *Make **MTTR** as small as possible*
  - *Proactive replacement/maintenance where **MTTR** cannot be reduced*

MTTR - ***Mean Time to Repair.***

MTBF - ***Mean Time Between Failures.***

Maximize Reliability *Increase MTBF*

- Get the System to a Manageable State (stability point).
  - *Inventory audit – You have to know what is out there.*
  - *Replace failed units.*
  - *Repair/replace units in error/faulted state.*

- Reduce Complexity
  - *Get rid of unnecessary units.*
  - *Remove unnecessary components (PDUs, Transfer Switches …)*
  - *Reduce unwanted loads on the system.*

- Implement Redundancy
  - *Design/Deploy systems with redundant power supplies.*
  - *Implement redundant systems so failures are actually tolerable.*

- Monitor for Faults, Failures and Predicted Failures
  - *Email Alerts – from units themselves and active polling*
  - *Alerts for low Runtimes, Faults and Failures*
  - *Performance Graphing for historical perspective and statistics*
  - *Failure Prediction*
  - *Centralized Logging for history and analytics*

Maximize Maintainability *Decrease MTTR*

- *Corrective* rather than *Preventative* maintenance
  - *Predicted failures are replaced before problems arise*
  - *Actual failures are replaced quickly*
  - *Batteries are replaced when there are faults not on a schedule*
  - *Configuration changes and updates are done globally*
- On the shelf spares
- Implement redundancy wherever possible
  - *MTBF and MTTR are zero if system is redundant!*
  - *Critical systems always have dual power supplies and UPSs*
- Automate as much as possible
  - *Research available tools (open source and vendor provided!)*
  - *Automation produces consistent and complete configuration*
  - *Custom scripts can automate tasks and deployment (~5 min)*
  - *"If you do something more than once, you should write a script to do it!"*

Vendor Tools *Automation and Monitoring*

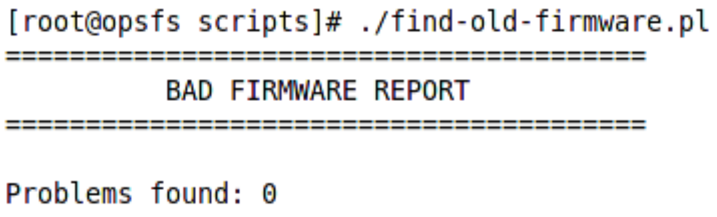
APC UPS Network Management Tools

- *Monitoring, Control and Remote Access*
  - *Web browser*
  - *Command line interface*
  - *SNMP, SSH, FTP …*
- Event and Data Logging
- *Scheduled Self-Tests*
- *Email Notification*
  - *Faults, Failures & Warnings*
  - *Self-Test Results*
  - *Configuration Changes*

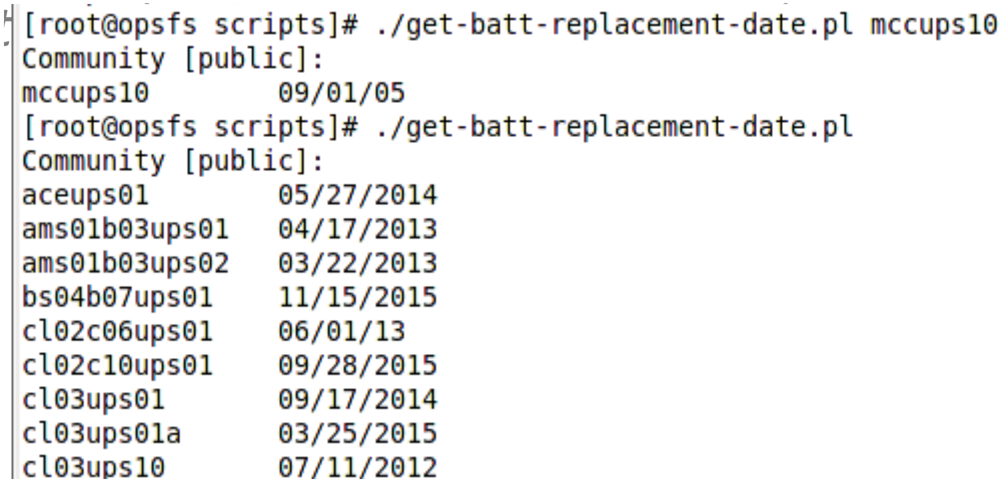
Custom Tools *Automation*

Scripting Framework

- *Global management tasks via custom scripts*
- *Scripts leverage vendor tools*
- *Simple text file database of UPSs*
- *Centralized configuration management*
- *Configuration consistency checks*
- *Information gathering tools*



```
[root@opsfs scripts]# ./find-old-firmware.pl
========================================
         BAD FIRMWARE REPORT
========================================

Problems found: 0
```



```
[root@opsfs scripts]# ./get-batt-replacement-date.pl mccups10
Community [public]:
mccups10        09/01/05
[root@opsfs scripts]# ./get-batt-replacement-date.pl
Community [public]:
aceups01        05/27/2014
ams01b03ups01   04/17/2013
ams01b03ups02   03/22/2013
bs04b07ups01    11/15/2015
cl02c06ups01    06/01/13
cl02c10ups01    09/28/2015
cl03ups01       09/17/2014
cl03ups01a      03/25/2015
cl03ups10       07/11/2012
```

Jefferson Lab

Custom Tools *Automation*

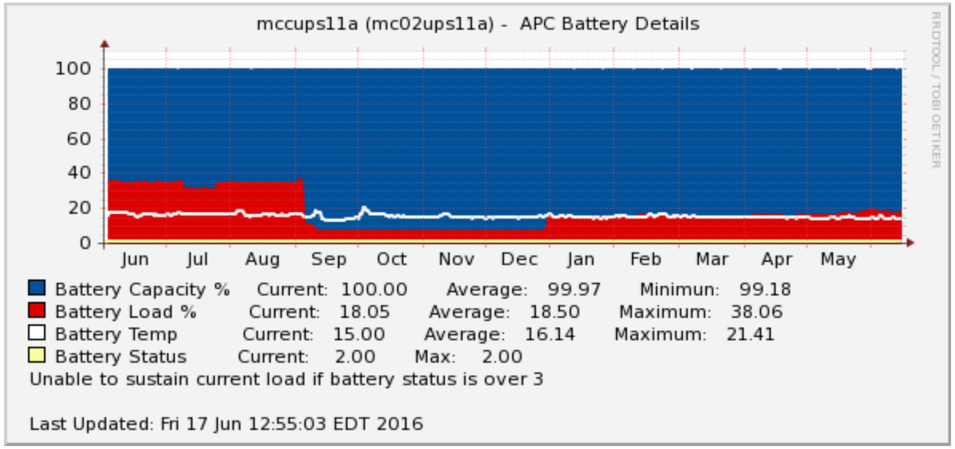
- **Rapid Deployment System (~5 mins)**
  - *Automated deployment with custom scripts*
  - *Process leverages vendor provided tools*
  - *Easy and well documented procedure*
  - *Consistent and complete configuration*
  - *Rapid and automated reconfiguration*

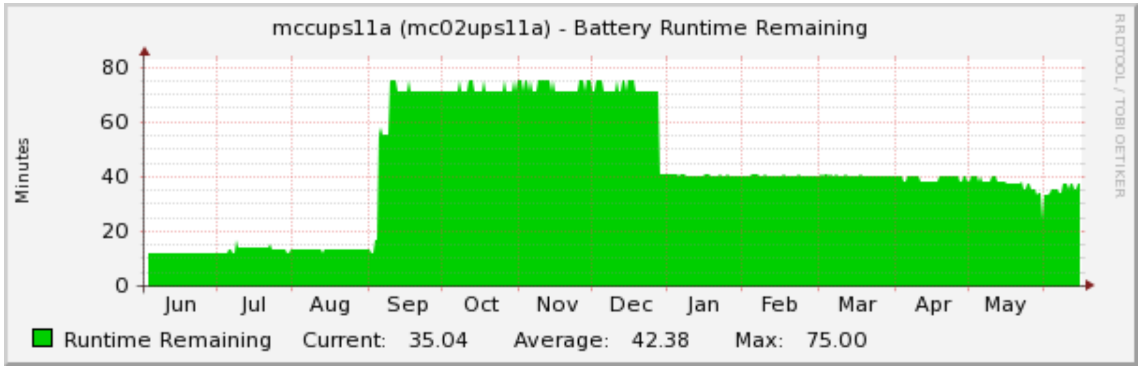
```
[upsadm@node scripts]#  ./deploy_new_ups.pl mccups13
Username [apc]:
Password [apc]:
Community [public]:

Ready to run on the following UPSes.
mccups13
Continue (y/N):
```

Monitoring Tools *Graphing with Cacti*

Cacti – Open Source Polling and Graphing Framework

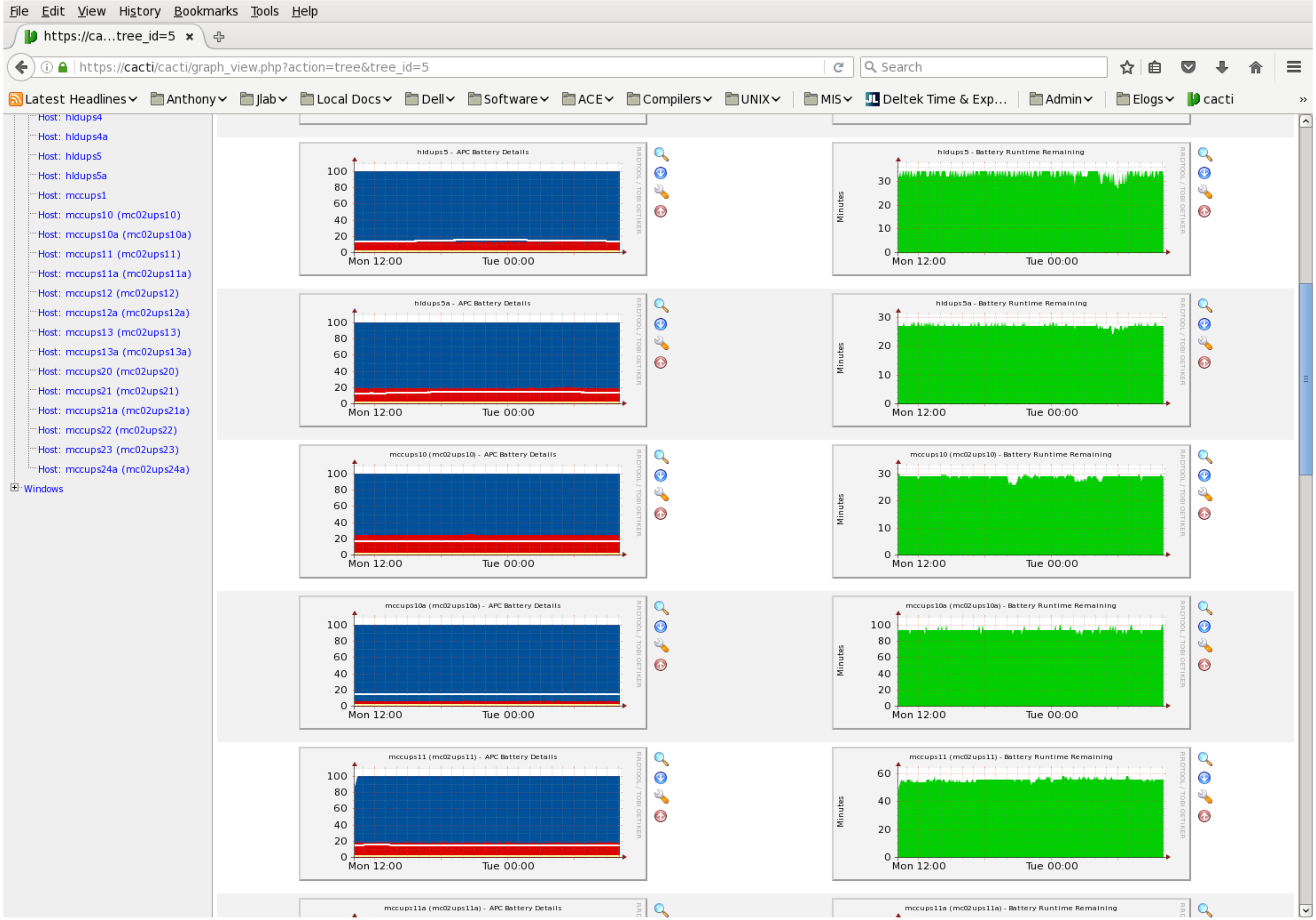
- *Performance Monitoring via SNMP*
- *Adaptable to Graph Virtually Anything*
- *Historical Statistics*
- *Problem Identification*
- *Failure Prediction*
- *Global Overview*
- *Free*





Monitoring Tools *Cacti*

Custom Global Overviews



- Host: hldups4
- Host: hldups4a
- Host: hldups5
- Host: hldups5a
- Host: mccups1
- Host: mccups10 (mc02ups10)
- Host: mccups10a (mc02ups10a)
- Host: mccups11 (mc02ups11)
- Host: mccups11a (mc02ups11a)
- Host: mccups12 (mc02ups12)
- Host: mccups12a (mc02ups12a)
- Host: mccups13 (mc02ups13)
- Host: mccups13a (mc02ups13a)
- Host: mccups20 (mc02ups20)
- Host: mccups21 (mc02ups21)
- Host: mccups21a (mc02ups21a)
- Host: mccups22 (mc02ups22)
- Host: mccups23 (mc02ups23)
- Host: mccups24a (mc02ups24a)

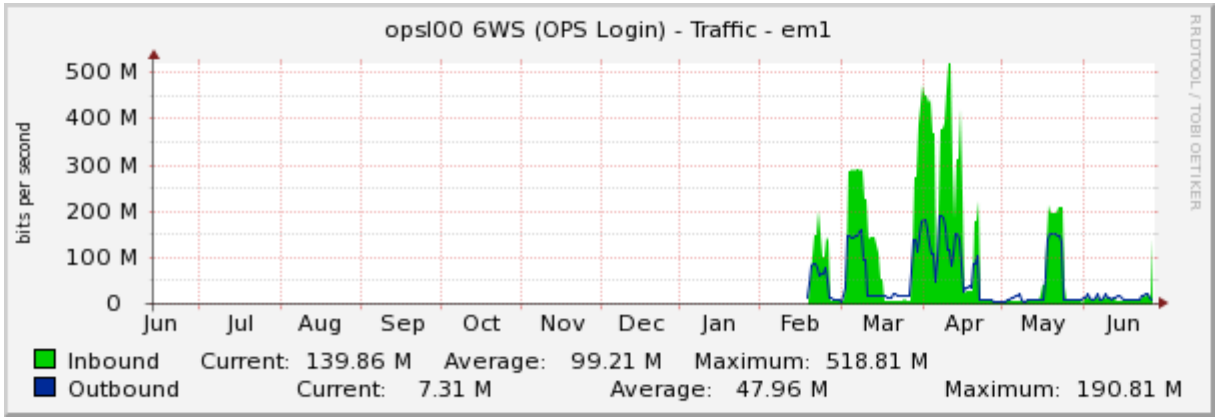
Windows

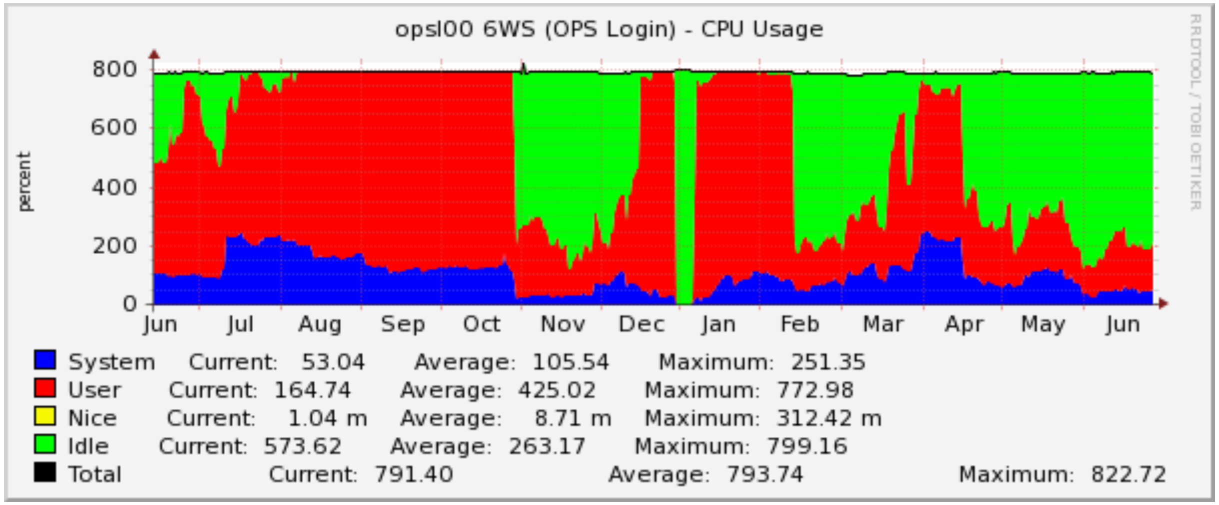
hldups5 - APC Battery Details
100
80
60
40
20
0
Mon 12:00
Tue 00:00
hldups5 - Battery Runtime Remaining
Minutes
30
20
10
0
Mon 12:00
Tue 00:00
hldups5a - APC Battery Details
100
80
60
40
20
0
Mon 12:00
Tue 00:00
hldups5a - Battery Runtime Remaining
Minutes
30
20
10
0
Mon 12:00
Tue 00:00
mccups10 (mc02ups10) - APC Battery Details
100
80
60
40
20
0
Mon 12:00
Tue 00:00
mccups10 (mc02ups10) - Battery Runtime Remaining
Minutes
30
20
10
0
Mon 12:00
Tue 00:00
mccups10a (mc02ups10a) - APC Battery Details
100
80
60
40
20
0
Mon 12:00
Tue 00:00
mccups10a (mc02ups10a) - Battery Runtime Remaining
Minutes
100
80
60
40
20
0
Mon 12:00
Tue 00:00
mccups11 (mc02ups11) - APC Battery Details
100
80
60
40
20
0
Mon 12:00
Tue 00:00
mccups11 (mc02ups11) - Battery Runtime Remaining
Minutes
60
40
20
0
Mon 12:00
Tue 00:00
mccups11a (mc02ups11a) - APC Battery Details
mccups11a (mc02ups11a) - Battery Runtime Remaining
RRDTOOL / TOBI OETIKER

Monitoring Tools *Cacti*

Adaptable to Graph Virtually Anything

… even mis-behaving users

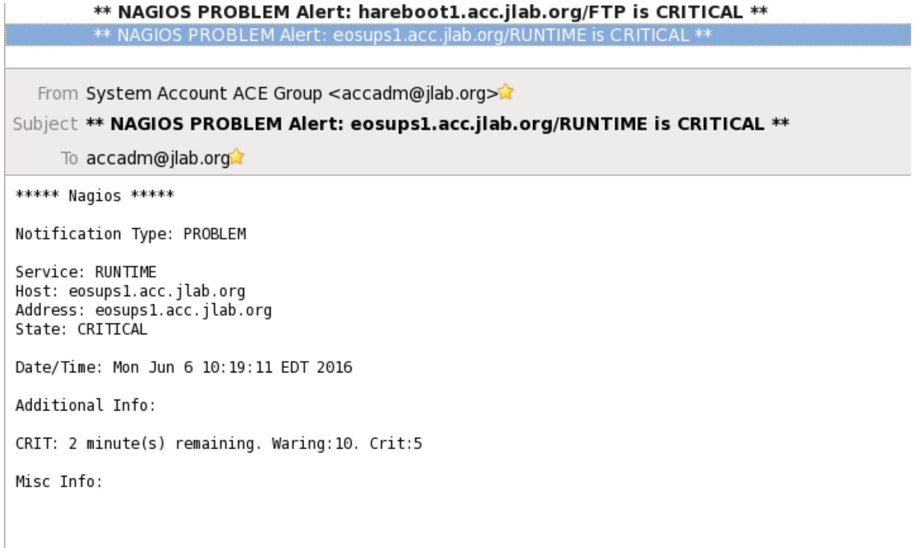




Monitoring Tools *Status Polling with Nagios*

Nagios – Open Source Monitoring Framework

- *System and Service Status Polling via SSH, SNMP, …*
  - *Runtime Warnings*
  - *Fault and Failure Detection*
  - *Active and Negative Service Monitoring (want SSH but not FTP running)*
- *Easily adapted to almost any need and for multiple groups*
  - *Already monitoring services for others groups like AHLA*
  - *Group based service/system notification*
- *Email/SMS Alerts*
- *Web based H.U.D*
- *Monitor for Monitoring*
- *Free*



**** NAGIOS PROBLEM Alert: hareboot1.acc.jlab.org/FTP is CRITICAL ****

** NAGIOS PROBLEM Alert: eosups1.acc.jlab.org/RUNTIME is CRITICAL **

From System Account ACE Group <accadm@jlab.org>

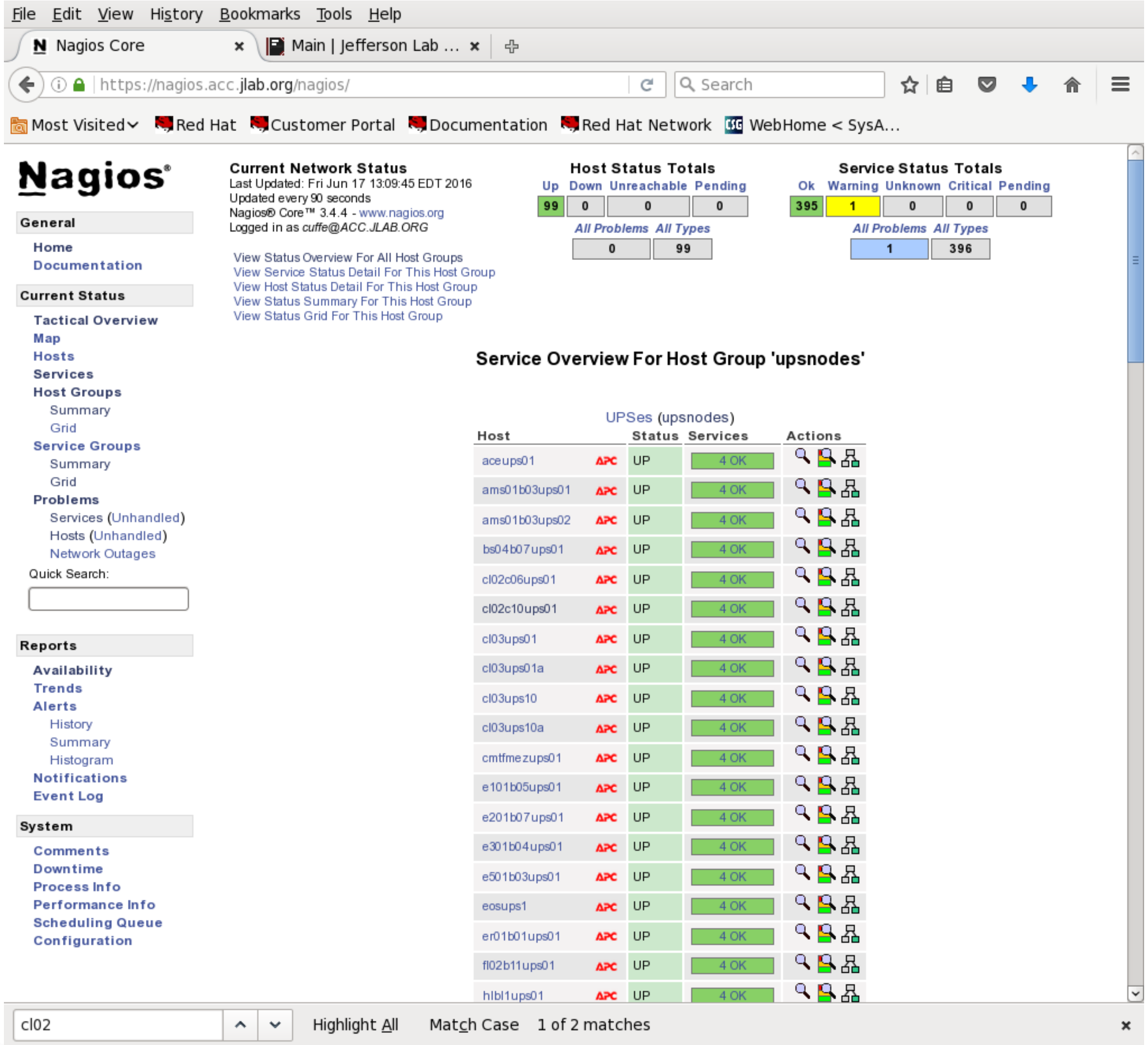
Subject **** NAGIOS PROBLEM Alert: eosups1.acc.jlab.org/RUNTIME is CRITICAL ****

To accadm@jlab.org

```
***** Nagios *****

Notification Type: PROBLEM

Service: RUNTIME
Host: eosups1.acc.jlab.org
Address: eosups1.acc.jlab.org
State: CRITICAL

Date/Time: Mon Jun 6 10:19:11 EDT 2016

Additional Info:

CRIT: 2 minute(s) remaining. Waring:10. Crit:5

Misc Info:
```

Monitoring Tools *Status Polling with Nagios*

Heads-Up-Display



Current Network Status
Last Updated: Fri Jun 17 13:09:45 EDT 2016
Updated every 90 seconds
Nagios® Core™ 3.4.4 - www.nagios.org
Logged in as *cuffe@ACC.JLAB.ORG*

View Status Overview For All Host Groups
View Service Status Detail For This Host Group
View Host Status Detail For This Host Group
View Status Summary For This Host Group
View Status Grid For This Host Group

Host Status Totals

Up	Down	Unreachable	Pending
99	0	0	0

All Problems	*All Types*
0	99

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
395	1	0	0	0

All Problems	*All Types*
1	396

Service Overview For Host Group 'upsnodes'

UPSes (upsnodes)

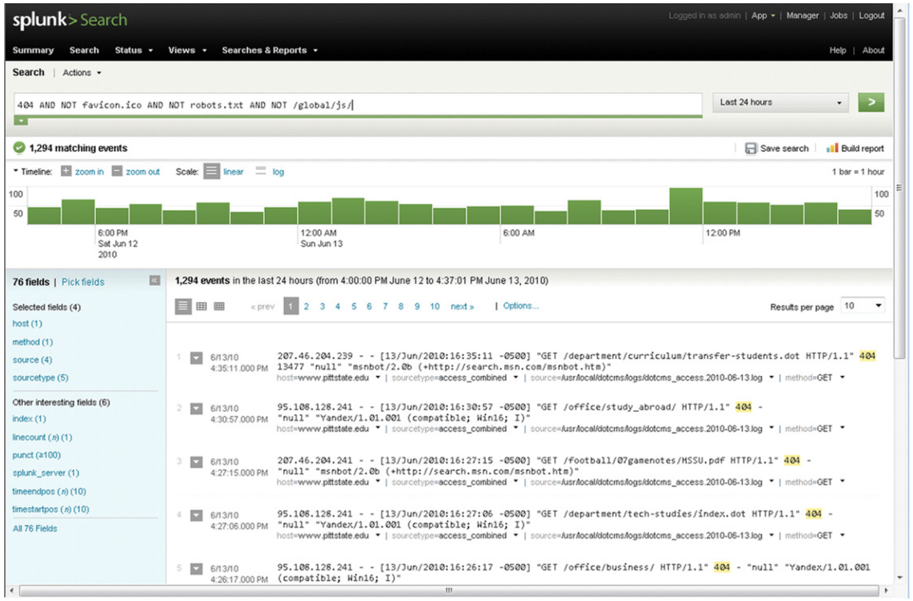
Host		Status	Services	Actions
aceups01	APC	UP	4 OK	
ams01b03ups01	APC	UP	4 OK	
ams01b03ups02	APC	UP	4 OK	
bs04b07ups01	APC	UP	4 OK	
cl02c06ups01	APC	UP	4 OK	
cl02c10ups01	APC	UP	4 OK	
cl03ups01	APC	UP	4 OK	
cl03ups01a	APC	UP	4 OK	
cl03ups10	APC	UP	4 OK	
cl03ups10a	APC	UP	4 OK	
cmtfmezups01	APC	UP	4 OK	
e101b05ups01	APC	UP	4 OK	
e201b07ups01	APC	UP	4 OK	
e301b04ups01	APC	UP	4 OK	
e501b03ups01	APC	UP	4 OK	
eosups1	APC	UP	4 OK	
er01b01ups01	APC	UP	4 OK	
fl02b11ups01	APC	UP	4 OK	
hlbl1ups01	APC	UP	4 OK	

Logging Tools *Centralized Logging w/Analytics*

- **Swatch** (*Syslog Watch*)
  - *Filter and Summary Tool for centralized logging – filter out normal, noisy logs*
  - *Daily Reporting via Email*
  - *Extensible Filtering – easily add custom filters*
  - *Open Source*
- **Splunk**

Splunk is a fully featured, powerful platform that collects and indexes any machine data from virtually any source in real time. We feed it our centralized logs.

- *Index any Text Data*
- *Interactive Search Results*
- *Monitoring and Alerting*
- *Reporting and Analysis*
- *Event Correlation*
- *Custom Dashboards*
- *Add-in apps*
- *Already extended to other groups*
- *Not Free but DOE contract in works!*



```
404 AND NOT favicon.ico AND NOT robots.txt AND NOT /global/js/
```

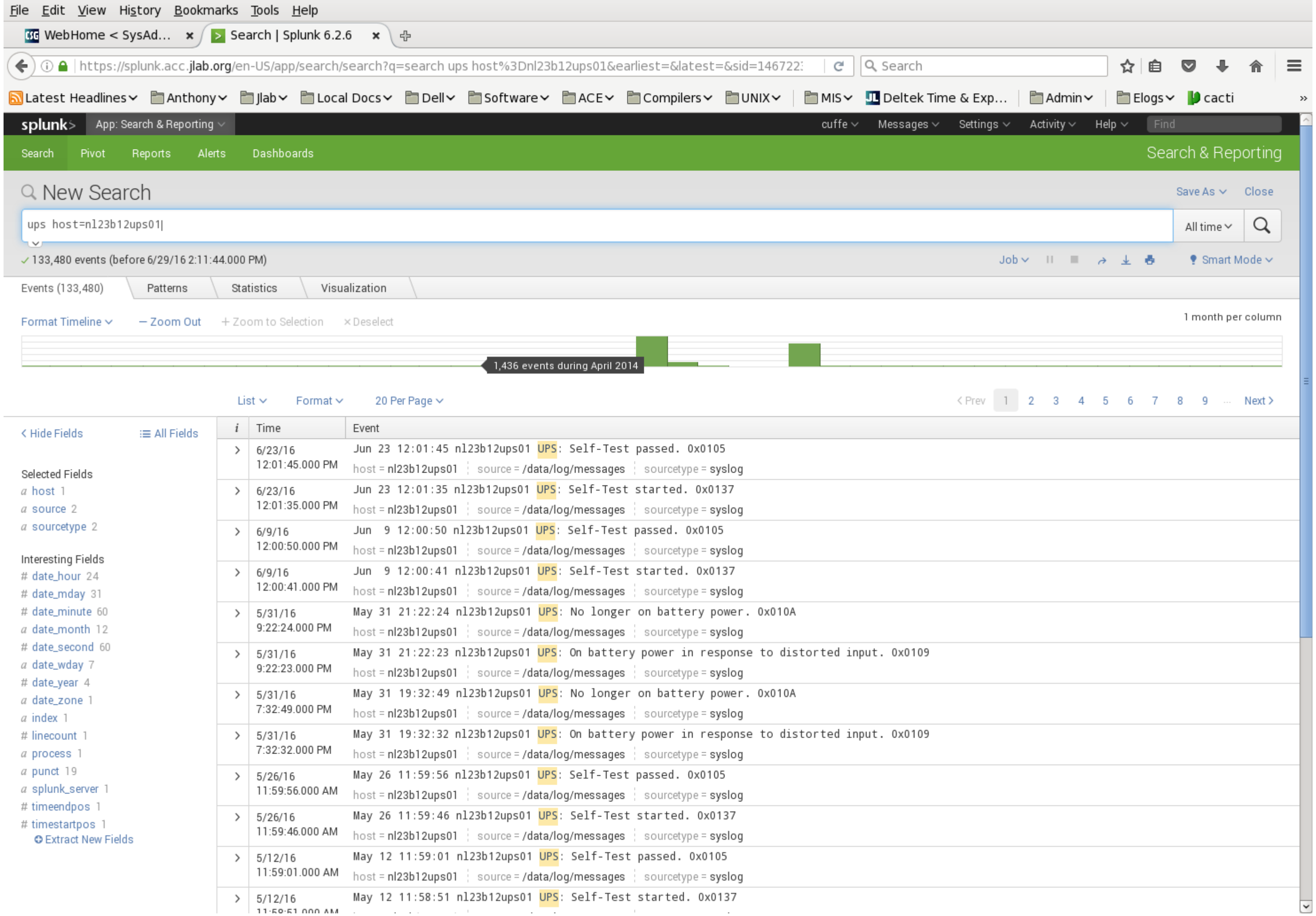
1,294 matching events

1,294 events in the last 24 hours (from 4:00:00 PM June 12 to 4:37:01 PM June 13, 2010)

```
1 6/13/10 4:35:11.000 PM
207.46.204.239 - - [13/Jun/2010:16:35:11 -0500] "GET /department/curriculum/transfer-students.dot HTTP/1.1" 404
13477 "null" "msnbot/2.0b (+http://search.msn.com/msnbot.htm)"
host=www.pittstate.edu | sourcetype=access_combined | source=/usr/local/dotcms/logs/dotcms_access.2010-06-13.log | method=GET

2 6/13/10 4:30:57.000 PM
95.108.128.241 - - [13/Jun/2010:16:30:57 -0500] "GET /office/study_abroad/ HTTP/1.1" 404 -
"null" "Yandex/1.01.001 (compatible; Win16; I)"
host=www.pittstate.edu | sourcetype=access_combined | source=/usr/local/dotcms/logs/dotcms_access.2010-06-13.log | method=GET

3 6/13/10 4:27:15.000 PM
207.46.204.241 - - [13/Jun/2010:16:27:15 -0500] "GET /football/07gamenotes/MSSU.pdf HTTP/1.1" 404 -
"null" "msnbot/2.0b (+http://search.msn.com/msnbot.htm)"
host=www.pittstate.edu | sourcetype=access_combined | source=/usr/local/dotcms/logs/dotcms_access.2010-06-13.log | method=GET

4 6/13/10 4:27:06.000 PM
95.108.128.241 - - [13/Jun/2010:16:27:06 -0500] "GET /department/tech-studies/index.dot HTTP/1.1" 404 -
"null" "Yandex/1.01.001 (compatible; Win16; I)"
host=www.pittstate.edu | sourcetype=access_combined | source=/usr/local/dotcms/logs/dotcms_access.2010-06-13.log | method=GET

5 6/13/10 4:26:17.000 PM
95.108.128.241 - - [13/Jun/2010:16:26:17 -0500] "GET /office/business/ HTTP/1.1" 404 - "null" "Yandex/1.01.001
(compatible; Win16; I)"
```

Logging Tools *Splunk*

Deep-Dive Your Data



```
ups host=nl23b12ups01
```

✓ 133,480 events (before 6/29/16 2:11:44.000 PM)

Selected Fields

- a host 1
- a source 2
- a sourcetype 2

Interesting Fields

- # date_hour 24
- # date_mday 31
- # date_minute 60
- a date_month 12
- # date_second 60
- a date_wday 7
- # date_year 4
- a date_zone 1
- a index 1
- # linecount 1
- a process 1
- a punct 19
- a splunk_server 1
- # timeendpos 1
- # timestartpos 1

Extract New Fields

i	Time	Event
›	6/23/16 12:01:45.000 PM	Jun 23 12:01:45 nl23b12ups01 UPS: Self-Test passed. 0x0105 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	6/23/16 12:01:35.000 PM	Jun 23 12:01:35 nl23b12ups01 UPS: Self-Test started. 0x0137 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	6/9/16 12:00:50.000 PM	Jun 9 12:00:50 nl23b12ups01 UPS: Self-Test passed. 0x0105 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	6/9/16 12:00:41.000 PM	Jun 9 12:00:41 nl23b12ups01 UPS: Self-Test started. 0x0137 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/31/16 9:22:24.000 PM	May 31 21:22:24 nl23b12ups01 UPS: No longer on battery power. 0x010A host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/31/16 9:22:23.000 PM	May 31 21:22:23 nl23b12ups01 UPS: On battery power in response to distorted input. 0x0109 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/31/16 7:32:49.000 PM	May 31 19:32:49 nl23b12ups01 UPS: No longer on battery power. 0x010A host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/31/16 7:32:32.000 PM	May 31 19:32:32 nl23b12ups01 UPS: On battery power in response to distorted input. 0x0109 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/26/16 11:59:56.000 AM	May 26 11:59:56 nl23b12ups01 UPS: Self-Test passed. 0x0105 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/26/16 11:59:46.000 AM	May 26 11:59:46 nl23b12ups01 UPS: Self-Test started. 0x0137 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/12/16 11:59:01.000 AM	May 12 11:59:01 nl23b12ups01 UPS: Self-Test passed. 0x0105 host = nl23b12ups01 source = /data/log/messages sourcetype = syslog
›	5/12/16 11:58:51.000 AM	May 12 11:58:51 nl23b12ups01 UPS: Self-Test started. 0x0137

Jefferson Lab

Results *Higher Availability with Lower TCO?*

- **Before (10–15 units):**
  - *Configuration was manual, onerous and inconsistent*
  - *Deployment took a long time (+1 hour)*
  - *Attitudes towards system were apathetic or dreadful*
  - *Monitoring was nonexistent or a deluge.*
  - *Failures were commonplace and faults were ignored.*

- **After (100+ units):**
  - *Configuration is automatic, fast and consistent.*
  - *Deployment is very quick (~5 minutes).*
  - *Replacement is now simply a routine task .*
  - *Monitoring is proactive, accurate and very predictive.*
  - *Failures are rare and faults are dealt with quickly.*

Key Points to Take Away

- **Buy in by management a must.**
  - *You cannot maintain a system without resources!!*
- **Redundancy wherever possible**
  - *MTTR and MTBF are virtually zero with redundancy!*
- **Automate as much as possible**
  - *Ensures standardization.*
  - *Reduces staff time.*
  - *Improves the culture.*
- **Monitor as much as possible**
  - *Failure notification and prediction is critical!*
  - *Fixing things before they break means zero downtime!*
  - *What you don't know usually hits you hardest!*
  - *Pretty graphs and real data translate into better budgets!*
- **Find or create tools to make things faster and easier**
  - *Research vendor and open source solutions*
  - *Simple solutions like shell scripts can be very powerful*
  - *Leverage the expertise of other groups*

Q & A