Causes of Data Center Failures: Can DCIM Prevent Them?

WHITE PAPER

from

GreenField Software Private Limited

March 2016

Data Center failures are not uncommon. A partial failure is sometimes euphemistically termed an outage. A few high profile failures (or outages) are listed below:

Amazon Data Center in North Virginia: Initially triggered by a cable fault in the utility power distribution system, the impacted zone was transferred to DG Power. Unfortunately, the primary DG overheated and the back-up DG failed to come up due to a wrong configuration. The affected zone was left without primary, secondary and even power from the back-up DG system. Pundits' take on this: Amazon made a mistake in the testing process.

Multiple Outages at eBay: As of September 14, 2014, eBay encountered twelve reported outages, for multiple reasons. The 14th September event apparently had to do with "unexpected power issues with storage arrays linked to some of our databases".

Unfortunately, data center failures have serious consequences including financial losses that could run into millions of dollars. IT failures at the Royal Bank of Scotland, apparently caused due to poor testing, change management and running legacy equipment, resulted in the Bank being issued a £42m fine by UK's regulatory body, the Financial Control Authority. This was in addition to the Bank setting aside £175m to reimburse customers who lost money as a result of being unable to access their accounts in the summer of 2012.

A data center failure is an unplanned outage that leads to severe disruption of critical applications or services beyond a defined period depending on the monthly uptime SLA required for that data center or the application or service. Even disruption of few critical applications or services maybe considered an outage, although it may not be a complete failure of the entire data center. For a 99% uptime Data Center, the outage tolerance can be as high as 7 hours in the month. That reduces to 4 minutes for 99.99% uptime, but would be just 130 seconds for a 99.995% uptime. Hopefully, the services or applications should still be available from a DR Center, but the primary would have to be docked as an instance of a Data Center failure.

Here's a simple 3x3 formula to overcome the three root causes of data center failures: human error, lack of integration between facility and IT systems/teams, and poor capacity planning.

1. **Avoiding Human Errors**
   a. Automate Standard Operating Procedures with real-time monitoring
   b. Disallow ad hoc actions (read Change Management)
   c. Apply common sense
2. **Integrating facility and IT systems/teams**
   a. Maintain an updated Power Chain
   b. Educate the two teams on the criticality of collaboration
   c. Implement "single pane of glass" view
3. **Improving Capacity Planning**
   a. Avoid single point of failures
   b. Maintain allowance for your weakest link
   c. Forecast resources before adding equipment

This Paper attempts to show how Data Center Infrastructure Management (DCIM) software have evolved over the last few years to avoid the common pitfalls in data center operations, and thus prevent outages, saving organizations from serious financial and reputation losses.

Avoiding Human Errors

Automate Standard Operating Procedures (SOPs) with real-time monitoring: While to err is human, error tolerance levels in data center operations have to be nearly as low as surgeon's errors on an operating table or a pilot's errors while flying a commercial plane. So what does the medical industry or the aviation industry do to minimize such errors? They have a detailed SOP and surgeons and pilots are thoroughly trained to abide by them. Moreover, redundancies do not stop at system levels. A commercial plane has multiple engines, but also flies with two pilots on board. A surgeon has a trained assistant by her side. To reduce human errors, modern aircrafts can run on auto-pilot mode. Surgeons rely on medical diagnostic devices monitoring a patient's parameters real-time.

A Data Center SOP would have multiple sections, one of which would be Risk Management for preventing failures. In this section, it should document first which data points from which devices should be monitored – and at what frequency intervals. Next, what are the actions to be taken if a device fails. If the SOP manual exists, and it is being followed, the facility staff can take remedial actions as per SLA norms. Now imagine there is no SOP and an entire bank of UPS fails at midnight. Either all hell would break loose, or a smart guy would reckon moving the entire IT load to the back-up bank or maybe even shift the load to the Disaster Recovery Center. In this situation, while a disaster is averted because of the quick thinking and action of an individual, the state of affairs is unacceptable.

> **How DCIM Helps**. We recommend going beyond a SOP manual and introduce automated processes. That would mean including SOP as part of one of the Data Center Management tools being used by most of the Data Center staff. The best candidate is the Data Center Infrastructure Management (DCIM) software.
>
> A policy-driven DCIM can list the conditions of device failures and what actions to be taken and by whom, besides (and beyond) sending passive alerts only. The next generation of DCIM Software would include control functions, and large parts of failure management would be software defined.

Disallow Ad Hoc Actions (read Change Management): Ad hoc actions, especially by rookies, result in errors that sometimes cause adverse situations in the data center. This is usually linked to managing routine changes, which are quite common in data center operations.

Not well trained and lacking close supervision, the rookie sets out doing a seemingly routine change like placing newly-arrived servers on racks. Placing them on ad hoc basis on empty U-spaces can lead to power trips.

> **How DCIM Helps**. DCIM, that not only detects overloading but also defines in advance where such servers can be placed, would stop such ad hoc actions. A combination of policy and workflow-based approval systems prevents ad hoc actions in Move-Add-Change operations.

Apply Common Sense: There can be no substitutes to applying common sense and maintaining discipline. If anyone brings in food and drinks into the data center, that has to be dealt with firmly. Common sense tells us not to try too many changes all at once. God forbid, if something goes wrong it would take that much longer to identify the root cause.

> **How DCIM Helps**. Workflow-based approval system in DCIM can detect and avoid if multiple changes are being attempted during a same time period.

Integrating Facility & IT Systems/Teams

Maintain an Updated Power Chain: It has become a fairly routine occurrence that an upgrade or scheduled maintenance activity has led to power or network outage in some part of the data center.

> **How DCIM Helps.** This can be prevented if there is an up-to-date Power Chain on the DCIM. When a scheduled activity on any part of the Chain is being undertaken, all downstream owners must be notified for making contingency plans. Again, this is part of Change Management, discussed earlier.

Educate the Facility & IT Teams on Criticality of Collaboration: Collaboration and working as a team have to be built as necessary competencies among the Data Center staff. As maintaining a certain PUE level maybe a SLA requirement, there could be a reluctance by the Facilities team to increase cooling when it may be genuinely required, if just asked by the IT team.

> **How DCIM Helps**. Visualization of the data center with danger-level hot spots, would obviously alert the Facilities team to take immediate corrective actions even if PUE is compromised. Proactive alerts with escalation across the two teams have helped collaboration wherever DCIM has been implemented.

Implement Single Pane of Glass View: Despite extreme redundancies, we have witnessed high profile data center failures. Part of the reason is the lack of visibility about the health of equipment in the other half. Imagine this real life scenario: Main source of power fails and IT systems are automatically now on UPS power. But power back-ups malfunction, as it happened in Amazon's Virginia Data Center. We now have a problem!

> **How DCIM Helps.** DCIM can be configured to broadcast alerts to System, Network and Application administrators when there is outage of main power. If DG or alternate power does not activate within five minutes, this too can be broadcasted through the Data Center organization for immediate remedial actions to be taken per the Automated SOP guidelines.

Improving Capacity Planning

Avoid Single Point of Failures: This usually occurs during operation stage. While the initial design has redundancies built at each layer, a common occurrence in the later operational phase, when more devices are being installed, is to overlook a small component (example, a Layer 3 Switch) through which the newly procured devices are being connected. While all critical downstream components are redundant, the failure of a single Layer 3 Switch can have disastrous consequences.

> **How DCIM Helps**. A visual Power Chain in DCIM can detect such a single point of failure.

Maintain Allowance for your weakest link: This could be the Layer 3 Switch we just talked about. Here comes the issue of Processes and People. Process must define that in this case, (if for whatever reason a built-in redundancy cannot be configured), auto-monitoring for the Layer 3 Switch has to be at 5 seconds (or less) intervals and couple of spares (frequently tested as per SOP norms that they are working) kept securely and near at hand to do an immediate replacement. On the People front, we have to ensure that we have two competent persons, trained in doing this replacement, available on-site 24x7. Hence, this off-line redundancy is not just at spare level but also on technical resource front, since at device level the redundancy is missing in the power chain.

> **How DCIM Helps**. DCIM can be custom-configured for different devices with different polling frequencies.

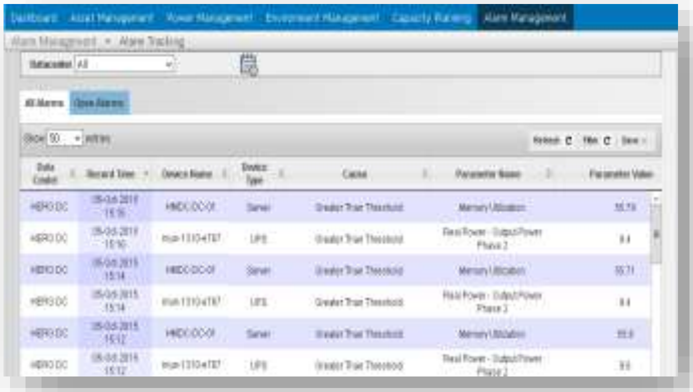
Forecast resources before adding equipment: Do I have available power, space and cooling to add more IT devices? If not, are there some stranded capacities that can be released? Failure to answer the first can lead to a power trip. Failure to answer the second can lead to wasteful capital expenditure or unnecessarily increasing power costs.

> **How DCIM Helps.** DCIM Software's Capacity Planning capabilities can help the Data Center Manager with inventory management of power, space and cooling – what's immediately available, releasing stranded capacities and requisition for additional resources right up to actual provisioning.

How GFS Crane® DCIM Prevents Data Center Outages

GFS Crane is a complete DCIM where the functions of power and environment monitoring, asset management and capacity planning all converge towards providing a 99.95% uptime data center. The key elements of GFS Crane DCIM that prevent a data center outage are:

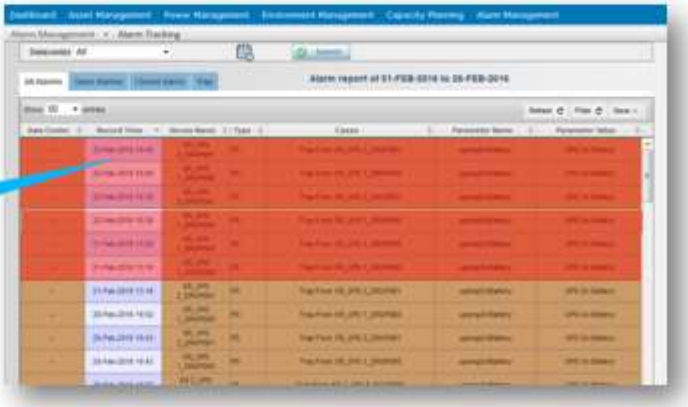
- ✓ **Policy-driven**: defines data center's standard operating procedures; includes which input-output data points from which devices and at what frequency intervals they be monitored



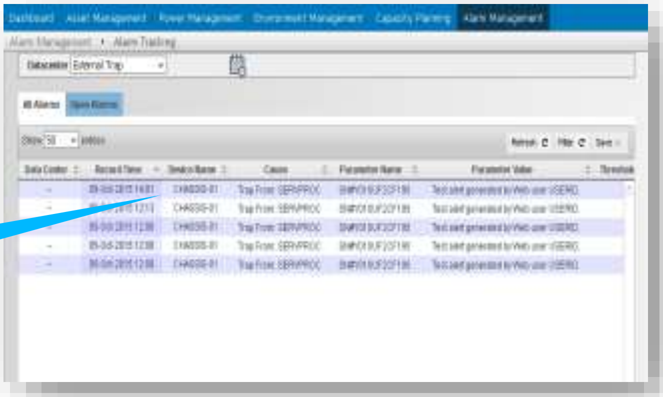
Data Center	Record Time	Device Name	Device Type	Cause	Parameter Name	Parameter Value
HERO DC	05-06-2015 15:16	HMDC-DC-01	Server	Greater Than Threshold	Memory Utilization	55.79
HERO DC	05-06-2015 15:16	mum-1310-4TB7	UPS	Greater Than Threshold	Real Power - Output Power Phase 2	8.4
HERO DC	05-06-2015 15:14	HMDC-DC-01	Server	Greater Than Threshold	Memory Utilization	55.71
HERO DC	05-06-2015 15:14	mum-1310-4TB7	UPS	Greater Than Threshold	Real Power - Output Power Phase 2	8.4
HERO DC	05-06-2015 15:12	HMDC-DC-01	Server	Greater Than Threshold	Memory Utilization	55.8
HERO DC	05-06-2015 15:12	mum-1310-4TB7	UPS	Greater Than Threshold	Real Power - Output Power Phase 2	8.8

- ✓ **Real-time alarms:** sends active alerts sent to concerned individuals with recommendations of what needs to be done – based on SOP guidelines

Real time alarms on Critical Device Health

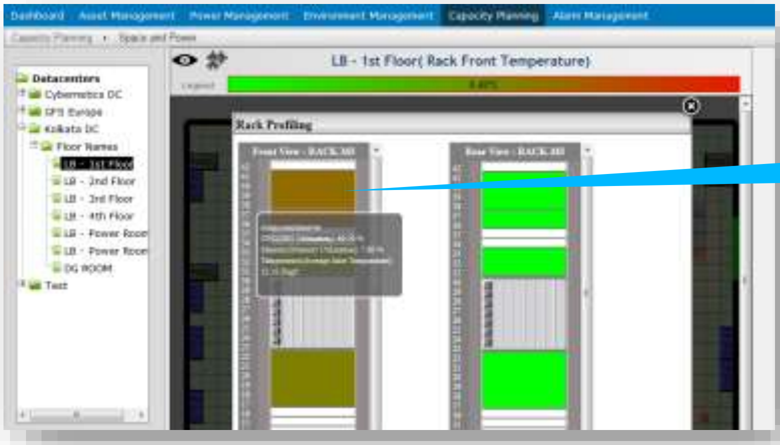


Alarm Showing input power failure to UPS



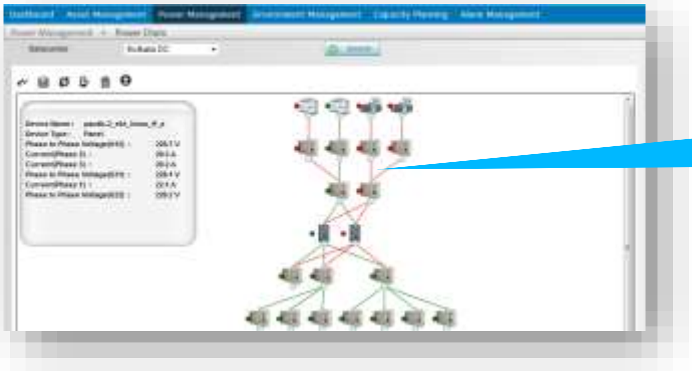
Alarm Delivery through SNMP trap

- ✓ **Visual Data Center Layout**: spots location of alarm, allows isolation of affected devices and prevent cascading impacts downstream.



Identifying racks with hot-spots which can lead to IT

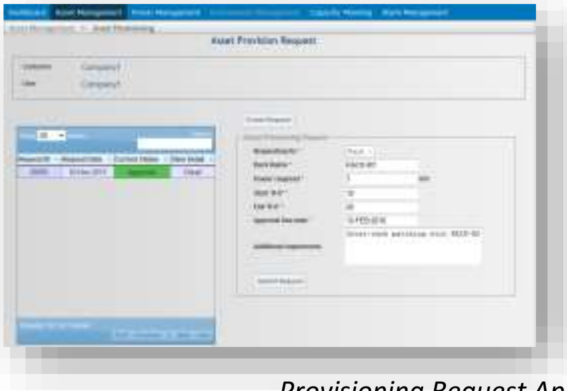
- ✓ **Visual Power Chain:** allows spotting any single point of failure. If unavoidable, polling for that device can be configured to be more frequent



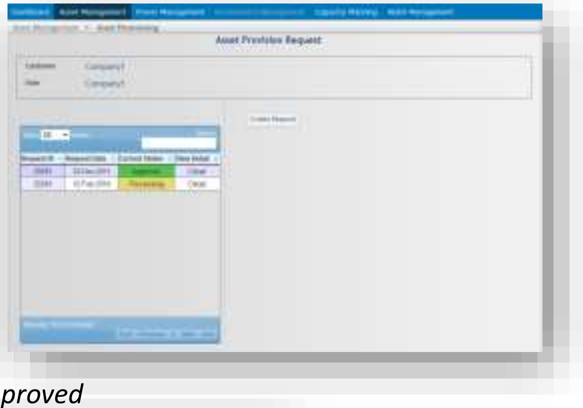
Critical Device Relationship Mapping to identify impact of a failure on downstream devices

- ✓ **Workflow-based approval system**: ensures disciplined change management

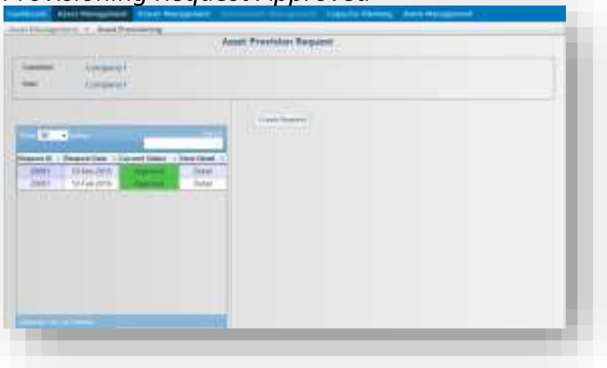
Provisioning Request Raised



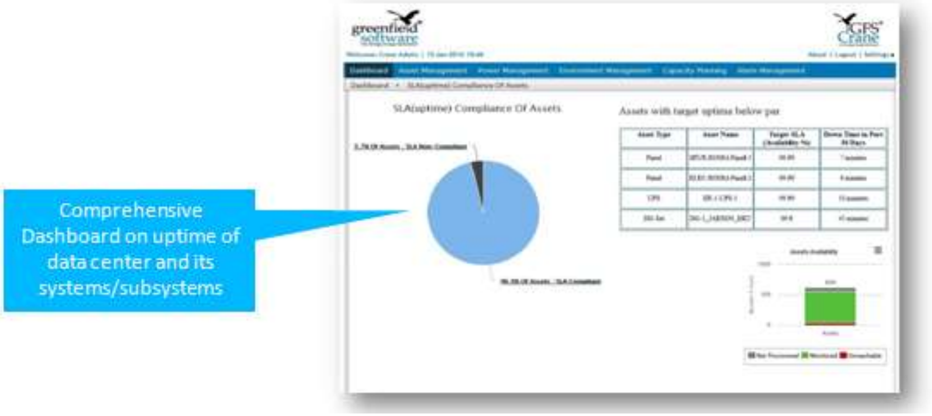
Request Processed through Workflow Engine



Provisioning Request Approved

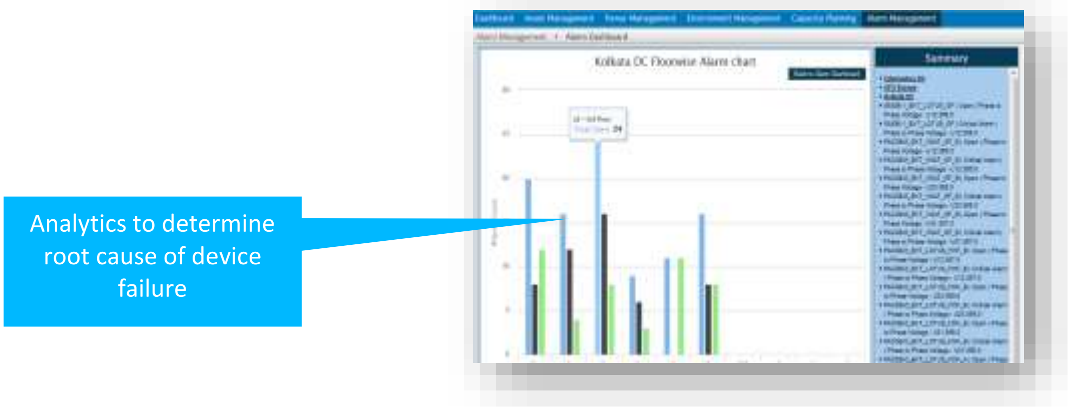


✓ **Analysis on device performance**: generates device Uptime reports; root cause analysis of failures and aging analysis of equipment helps to either improve performance or prevent unexpected failures


greenfield software
GPS Crane
About | Logout | Settings
SLA(uptime) Compliance Of Assets
Assets with target uptime below par
Assets Availability

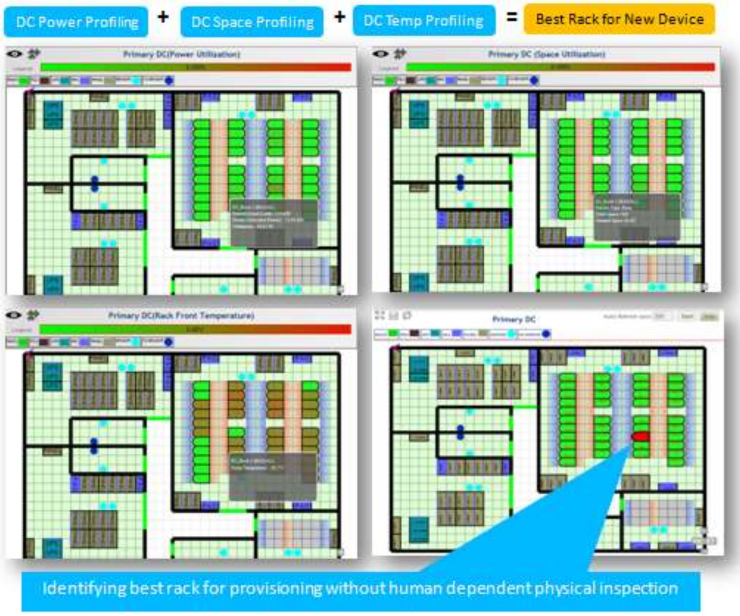
Asset Type	Asset Name	Target SLA (Availability %)	Down Time in Past 30 Days
Panel	[illegible]	[illegible]	[illegible]
Panel	[illegible]	[illegible]	[illegible]
UPS	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

Comprehensive Dashboard on uptime of data center and its systems/subsystems



Analytics to determine root cause of device failure

✓ **Capacity Planning & Provisioning**: locates rack space options based on available power and current heat loads. Recall, ad hoc placements can cause power trips on racks.


DC Power Profiling
+
DC Space Profiling
+
DC Temp Profiling
=
Best Rack for New Device
Primary DC(Power Utilization)
Primary DC (Space Utilization)
Primary DC(Rack Front Temperature)
Primary DC

Identifying best rack for provisioning without human dependent physical inspection

Summary

In a world of always-on service delivery, data center failures are unthinkable. The financial implication and loss of reputation alone make it imperative that we put in place systems that prevent such failures. Data Center Infrastructure Management (DCIM) software helps to avert the three principal causes of data center failures – human errors, lack of integration between facilities and IT systems or teams, and lack of capacity planning. If DCIM delivers on the High Availability promise alone, it would have paid back and more the investment made on its deployment in less than two years. If we account for other gains such as improvements in energy and operational efficiencies, the payback period for DCIM investment can be as low as fifteen months.

GreenField Software Private Limited is an Indian venture pioneering intelligent infrastructure management solutions. The product portfolio includes GFS Crane®, a complete DCIM suite, with installations in enterprise data centers of Financial Services, Telecom, Power Utilities, Media, Oil & Gas, Discrete Manufacturing and Higher Education.

For more details:

Email: sales@greenfieldsoft.com

Visit: www.greenfieldsoft.com