Report on the Failure of Information Systems Computer Room in W92 on December 20th, 2002

Table of Contents

Executive Summary

On December 20th, 2002 at 6:03AM a Cambridge Electric Power surge damaged the Uninterruptable Power System in the W92 IS Computer Room. This resulted in the loss of all services from that facility from approximately 6:55AM until around 10AM when most critical services were restored. Most services were back by 11AM and complete restoration was accomplished around 12:30PM.

The timing and nature of the failure resulted in staff not being notified until about 7:30AM. However at 7:30AM we only knew that "something" was wrong, but not exactly what. Attempts to analyze the problem from home combined with rush hour traffic exacerbated the time required to restore service.

This document describes what happened, what went wrong, and what steps we can and will take to minimize the likelihood and impact of a similar failure in the future. Some remedial steps are simple and don't require a significant resource commitment. These we are already working on implementing.

Other steps will require significant resource an expanse. We will have to make trade-offs between implementing these steps or accept the resultant risk of failure.

What Happened?

W92 is connected directly to Cambridge Electric's (CELCO) power grid and is not connected to the MIT Cogeneration facility. Power to the computer facility is protected via an MGE Uninterruptable Power Supply (UPS).

At approximately 4:30AM on the 20th the CELCO grid experienced a power "glitch" which was repeated at approximately 4:45AM and a much smaller glitch occurred at 5:45AM. The UPS system in W92 rode out these glitches without problem. However at 6:03AM the power surged. According to eye witnesses elsewhere in Cambridge the power failed and returned several times in a very short interval.

On this power surge the UPS suffered a catastrophic failure. It did successfully switch over to internal battery power, but was unable to return to the AC Mains after the surge. The UPS batteries provided power until approximately 6:55AM at which time the computer room suffered a complete loss of power.

Services Effected

All services provided by the W92 computer room were affected. This included:

1. External Internet Connectivity

All external Internet Connectivity was lost. This included connection to Internet2 as well as Genuity, our commodity Internet provider.

2. E-Mail service:

   One of set of MIT E-mail routers was taken out of service, however the redundant set in building 24 remained operational. However most MIT Post Office servers are currently located in W92 and they were unavailable.

3. Web Services:

   web.mit.edu services and other web-based services such as search and webmail,were unavailable.

4. Athena services, including approximately 2/3 of the Andrew File System (AFS), the primary storage for the Athena Computing System.

Services Unaffected

Kerberos, Domain Name Lookup Service (DNS), and Zephyr (instant messaging) continued to operate because the servers that provide these services are redundant and located in other buildings besides W92. However enough other services were non-operable, that it probably didn't matter that these services were available.

Staff Response

Network Operations and Athena Operations maintain an on-call staff member. Both also have automated monitoring systems that automatically call on-call and other staff when critical failures are detected.

Unfortunately the UPS system in W92 was never integrated into this system. Therefore when the UPS switched to battery power, no alert was sent out. When power completely failed, the system that does the paging was among those that failed. Therefore no alert was sent at that time.

Jeff Schiller was notified directly by phone call at 7:30AM by Jim Bruce that there was a massive failure of the MIT Campus Network. It was not immediately evident that power was out in W92. Instead a failure of one of MITnet's primary routers was suspected. As a result his first response was to setup and login remotely to see if he could figure out what was going on. He directly dialed up the phone line connected to the console of our external routers. However this phone line was busy. This could have been due to other staff members also attempting to diagnose the problem.

Ron Hoffmann was notified of the problem. Discussions between Jeff and Ron

led to the hypothesis that there might not be power in W92. Facilities was called around 8AM. However facilities was unaware of any problem on campus or in Cambridge that could account for the apparent loss of power in W92.

Both Ron and Jeff at this point drove in to W92. Ron arrived around 8:30AM.

The Data Center Operations Services Team (DOST) in W91 noticed problems starting at 7:00AM, which is consistent with when the UPS failed. DOST paged the Network Operations on-call person at 7AM, 7:15AM and finally at 7:38AM. The Network Operations voice-mailbox was incorrectly configured to only page on messages marked "URGENT" by the caller via the voicemail menus. Anne Salemme, Network Operations on-call was first paged at 7:38AM. We presume that this was the first page that was explicitly marked urgent. At that point both Ron and Jeff were already working the issue. At approximately 8:00AM Ron Hoffmann paged Cana McCoy, who was the Athena Server Operations on-call staff member that morning. By shortly after 8:30 she was on route to w92 and had paged Jonathon Weiss.

It should be noted that the UPS system is a large high power device that works directly 480V AC Power. Operation and debugging needs to be performed by staff familiar with its operation. Not all W92 operations staff are trained and familiar enough with its operation diagnose its failure and take steps to avoid a complete failure.[1]

Our initial goal was to place the UPS in "bypass" mode, which would route power around it and permit us to restore service, albeit without the protection of the UPS.

However the documented procedure for performing this operation failed. Eventually we used a more invasive procedure to manually bypass the failed UPS. This procedure was successful and we restored electric power to the computer room at around 9:40AM.

Once we were convinced we had power successfully restored, we began a phased turn-on of network and server equipment. Additional Athena Server Operations staff members were paged to remotely assist with this process.

Foremost in our mind was concern that we not do anything that would make the situation worse. For example if we powered up the equipment only to have power lost a second time, risks more damage to data and equipment.

External connectivity was restored around 10AM and E-Mail service shortly afterward.

---

[1] At this point we believe losing power to the computer room was inevitable. Although we eventually bypassed the UPS, we did so with power to the computer room already turned off. We believe that if we attempted to bypass the UPS while it was still running on its internal batteries, that the computer room power would have failed or at least "glitched" at that point, crashing all systems.

Some AFS servers were restored shortly after 10AM, but others took longer, as did the Web servers and several other services. The delays were due both to some failed network equipment and the requirement that software consistency checks (fsck) needed to run prior to full service restoration.

All AFS servers were operational by 10:30AM, and by 11AM most remaining services were operational as well. The last service (stellar) did not come back on line until 12:30PM, because it failed its software consistency checks and needed programmer intervention. Various developer and testing machines were brought up later in the afternoon.

Aftermath

In addition to taking the steps outlined above, we also notified the UPS vendor's service department.

Shortly after the UPS was bypassed, an MGE service technician arrived, having traveled in from Nashua, New Hampshire. This was a "not supposed to happen" failure and the technician took quite a few hours to diagnose the problem. His efforts were also constrained because he did not want to perform tests that might result in us losing bypass power!

Around 1PM the technician determined that the "charger board" had failed and that he could get a replacement at their Stow, MA parts depot. Around 4PM the technician returned from Stow, part in hand... or so he thought.

What the technician had was a box labeled to contain the part we required, but in fact it contained a different part, which would not help us. He informed us that the nearest alternative part was in California and that he would have it flown in for next day delivery. Tom Coppeto, who was on hand at this point, made it clear to the technician that this was not an acceptable answer. This resulted in several calls with MGE the upshot of which was that they dispatched a part via truck from New Jersey and the on-site technician went back to Stow to search for a part (maybe it was in the box that had the label that corresponded to the part he was holding!).

He returned around 11PM with the correct part and installed it. He then waited around for the truck from New Jersey, which arrived at 3AM. By 3AM our UPS batteries, which had been depleted, were recharged and we were once again fully protected.

The technician noted that our unit could use a more through test and one was scheduled for the following Monday (which happened without incident).

We also learned that the other UPS that MGE had on the same CELCO circuit as W92 also failed in a very similar "should never happen" fashion. No doubt they

are looking into what happened![2]

What worked well

We suffered minimal damage to equipment. Only one switch board failed, and one RAID unit experienced a minor failure that resulted in a small, easily corrected inconsistency. No end-user data was lost, no stored e-mail was lost. Once power was restored service restoration went well, with no major surprises.

Could this failure have been prevented?

In analyzing this failure, We looked at whether or not a different staff response or a different configuration of equipment might have avoided this failure.

Let's first consider what happens when the UPS switches to battery. When this happens we have between 30 and 40 minutes to get power input to the UPS restored or the UPS into bypass mode (assuming that we do have commercial power), or the computer room will drop power. This really isn't enough time to do much unless staff is on-hand in W92 at the time of failure. At 6AM this is not likely.

Even if staff was on hand, this may not have been enough time to analyze the situation and make the necessary decisions, especially since the documented procedure for putting the UPS into bypass mode probably would not have worked. We would likely have decided to take a "safe" approach and perform an orderly shutdown of the computer room. The effect on service would be the same, a multi-hour service outage.

Frankly, we do not have the protection in our power environment that would prevent this sort of failure. Even if we doubled the amount of UPS battery, giving us twice the time would not have helped in our early-morning no-staff-on-hand situation. Remember, this was a "not supposed to happen" failure of the UPS.

Could Redundancy Help?

One obvious question comes up. Could redundancy have helped us in this situation? If we had distributed services to other locations besides W92 could we have minimized the impact of losing W92?

We believe the answer to this question is "Yes, it would have helped, but it would not have helped enough to make this a non-event." We suspect that having half of the MIT community being unable to read e-mail is almost as bad as the whole community. Either situation is disruptive to the Institute Community at large and doesn't reduce the pressure on staff to restore the

[2] This UPS is at the Millennium Pharmaceuticals plant.

services that are non-functional.

In fact having all of our critical services co-located has the property that we could restore them all fairly quickly. Imagine a software failure that effects all services (which can and does happen), having services geographically dispersed adds travel time many times over as staff shuttle from one location to another in an attempt to diagnose and/or repair the problem.

Life is full of trade-offs. We made the trade off to keep critical services "at hand" to minimize downtime in the event of a common mode software failure at the expense of a major failure if a common mode infrastructure problem resulted.

Having said the above, there are some things which we need to think about. For example moving one of our mail routing servers out of W92 (say to W91) would permit people to send e-mail that they composed. We should not re-home the W91 network to one of the routers in W92, we should leave it homed to the NW12 router so that if W92 is off-line, W91 is still connected to the network. We may move some of the Post Office servers to our building 24 location. But, as stated above, it isn't clear that this will help the situation we had. Instead it will help the situation where something physically destroys the equipment in W92.

What if the failure lasted longer?

What if the failure lasted longer? Although all of MIT's external connectivity is via a router located in W92, the physical fiber comes to campus at different points. We have the ability to re-route this fiber so that it terminates in E19 instead of W92. We could restore both commodity Internet service and Internet2 (NOX) service in this fashion. As we make design changes in our fiber network, we will likely want to ensure that we maintain this ability to re-route the fiber in the future. We should also maintain spare routing equipment in E19 to facilitate this.

Although it is also possible to physically move the external router from W92, it is not a light weight device and moving it would be a cumbersome effort, particularly under pressure to restore service (i.e., move it fast, but don't damage it!).

Web service could also be moved by physically moving one or more reasonably sized servers. However web service depends on AFS service, which is harder to relocate.

E-mail is a problematic service to deal with. It does not lend itself well to redundancy; you can only have one IMAP or POP mailbox location. Although it would be possible to setup "skeleton" IMAP servers (i.e., servers without people's currently stored mail) this would have a negative impact on clients that keep a local copy of the mail. Such a situation could cause the local clients to mistakenly believe that the missing mail was in fact intentionally deleted, and the local copy

would be deleted. Worse still, when regular service is restored, the clients may then delete the mail on the "real" server, mistakenly believing that it had been intentionally deleted by the end-user.

So a medium term failure (about a day) would best be dealt with by leaving e-mail service unavailable, or alternatively cobbling together a webmail only solution so people could read new incoming mail while not having access to their old mail. Obviously having this cobbled together solution ready to go would be a useful thing to do. The catch is making sure that it is tested and known to work.

A truly long term outage, assuming the mail servers were intact, would involve providing a secondary source of power to the machine room until the primary power could be repaired. If the room is physically unusable, then the equipment could be relocated elsewhere on campus.

Frankly a long term outage of W92, say lasting more then a day or so, is really the topic for a disaster recovery plan, which we should put in place.

Other Concerns

This failure resolved around the UPS system. However we have another risk to operation, namely the diesel generator that provides backup power to the computer room via the UPS.

The UPS is designed to deal with three distinct situations:

1. Provide smooth power across momentary interruptions in power and power surges.
2. Provide power for the time necessary for the diesel generator to start up in the even of a complete commercial power failure.
3. If the diesel fails to start, the UPS will drop power when its batteries are depleted. This will show the equipment a sudden loss of power, but protect the equipment from "dirty" power associated with surges.

It is this third situation that we need to be concerned with. If the diesel generator fails to start, we only have between 30 and 40 minutes to get it working, or we will drop power to the computer room.[3]

Experience has shown that Facilities does not respond in this kind of time window. Keep in mind that we are likely to need Facilities support at a time when the campus is without power and the few (sometimes only one) electrician on duty may have other priorities. In general,we are always a secondary priority behind animal quarters.

[3] The Diesel Generator was not a factor in this particular failure. As far as the diesel system was concerned, commercial power was on and being provided to the UPS.

We need to investigate whether it is appropriate to provide training on the generator to local staff.

Learnings/Remedial Steps

All failures of this type are learning experiences. The list below articulates some of the things we learned and some remedial steps that we can take. Some of these are "low hanging fruit", steps that will not require additional resources and can be implemented quickly. We are already working on implementing them. Others will require more resources, these will require understanding trade-offs:

- Provide automated monitoring of the UPS so staff knows when a failure has occurred. This one is already implemented.
- Provide additional UPS power for the paging system so out-call paging works even when the room is "dark." This one is being implemented.
- Consider a second UPS and/or additional batteries to provide extended battery runtime and protection against a UPS failure. This will requires additional resources as well as potentially a redesign of the computer room to accommodate it.
- Obtain spares kit for the existing UPS. This step is being pursued.
- Move some services outside of W92 where appropriate.
- Change the way the MIT homepage is maintained so that it is stored on multiple servers in multiple buildings[4] so that web service can be maintained across the failure of any single server or computer room (this will require process changes in how the MIT home page is updated, but should be possible given the replication abilities of AFS).
- Investigate staff training on the generator.[5]
- Maintain the ability to re-route fiber to move external connection to other locations as a backup (this may require resources to maintain spare router capacity).

Conclusions

Was the failure on December 20th avoidable? Almost certainly. However at what price? We made explicit and implicit trade-offs when we constructed the room and when we made our staffing decisions.

---

[4] IS has facilities in W92, W91, building 24 and E40 that can be used to house redundant data.

[5] This is a pro-active step that would not have helped this particular failure, but might prove beneficial should the generator fail to automatically start during a real power failure, thus avoiding a similar computer room "blackout."

One of these implicit staffing trade-offs is that we rely on a small but capable staff. This means that we don't have in-place plans to deal with every eventuality. Instead we react as necessary as situations develop. This strategy means that we can deploy new technology quickly; we don't have to alter our plans to deal with the new service.[6]

An explicit trade-off was installing only one UPS. We built the W92 computer room against strong push-back from the construction team that was concerned with keeping down the costs of the W92 renovation. Not putting W92 on the MIT power grid, but leaving it on the much more failure prone CELCO grid was another trade-off.

Given these trade-offs, we should expect occasional failures of this nature.

Acknowledgments

Both the Network Operations and Athena Server Operations teams responded to this event and restored service as quickly as possible.

Ron Hoffmann, Jonathan Weiss, Cana McCoy, Mark Silis, Tom Coppeto and Jeffrey Schiller spent a significant amount of time and effort diagnosing the problem, bringing up services and dealing with getting the UPS repaired.

Theresa Regan also deserves recognition for helping us get things running and keeping the rest of IS informed as well as IS's customers. We are also sure that others in IS, worked long and hard to help mitigate the effects of this service outage while we worked to restore service.

Jeffrey Schiller authored this report with contributions from Jonathan Weiss, Jag Patel, and James Kretchmar.

[6] We have also witnessed other organizations fail utterly when such plans fail in practice and the staff are not capable of altering them on the fly.