Computing and Mathematical Sciences Department

California Institute of Technology

Continuity of Operations Plan

2011-2012 Scholastic Year

Background

The Computing and Mathematical Sciences Department (hereafter "CMS") was formed during the 2010-2011 academic year as a merger between the Computer Science Department, the Applied Computational Mathematics Department, and the Control and Dynamical Systems Department. The department consists of over 25 full-time tenured or tenure-track faculty members, approximately 200 graduate students, 15 administrative employees, several research fellows and postdoctoral scholars. The department also maintains academic support for the entire Caltech undergraduate community. Academic activities include computationally intensive simulations and modeling, theoretical computer science and quantum information research, computer graphics research, data mining, efficient computing, asynchronous chip design, networking, and undergraduate pedagogy.

Most of the department's personnel are housed in the Walter and Lenore Annenberg Center for Information Science and Technology, which finished construction in 2009. The building was awarded LEED Gold certification, and contains one of the few seismically-rated server rooms on campus meeting Zone 4 criteria.

Crisis Team

Typically, any event in this document will require contacting at a minimum the Systems Administration Team.

David Leblanc, x2402, emergency contact: pager_leblancd@caltech.edu

Patrick Cahalan, x3290, emergency contact: pager_psc@caltech.edu, 626.497.9254

In addition, in most cases the department head and the department administrator should also be contacted to facilitate the information flow to the primary internal customers (the faculty and administrative staff).

Mathieu Desbrun, x6230, emergency contact 626.399.1971

Jeri Chittum, x6251, emergency contact 626.859.3695

In some specific scenarios, other contacts are necessary. They are listed where appropriate.

Scope of Planning

What Is Not Covered?

Certain types of crises exceed the design parameters of the CMS cluster and support equipment. Currently, certain types of potential exception scenarios are regarded as either statistically unlikely or of sufficient additional consequence that capital expenditures to mitigate those risks are regarded as inadvisable, given available resources. For example, an earthquake of sufficient magnitude to catastrophically damage the Annenberg building's central server room and the enclosures therein would also cause enough damage across the entire physical campus to render the Institute, as a whole, unable to provide sufficient chilled water or power to keep the central server room in operation, among other more human and regrettable consequences.

Thus, maintaining this level of uptime from a systems engineering standpoint would represent expenditures that will not provide practical benefit; in the event that the crisis actually occurs, restoration of services at the Institute level would be required before the CMS cluster could resume operations. Similarly, there currently is no plan to restore operations in the event of a high altitude detonation of a nuclear device, nor formal plan to maintain continuous operation in the event the Annenberg building is destroyed by accident or design. Finally, due to manpower restrictions, it is impractical to consider malicious insider activity at the level of the systems administration team. While malicious activity can be mitigated at the user level, and even to some extent at the level of escalated privileges, it is essentially impossible to plan for a nefarious, directed agent at the level of the support staff.

What Is Covered?

There are five main exception scenarios that are regarded both as sufficiently likely, and effectively mitigated, and thus they are covered in this plan. The first is a mid-intensity earthquake that damages the Institute (but not to the extent that chilled water, network connectivity, and power are no longer capable of being provided for the building), which causes one or more of the safety systems to be triggered, shutting down the cluster and requiring manual intervention to restore normal operations. The second is a high-intensity earthquake that damages the Institute to the extent that some chilled water and/or power are available, but only enough to provide limited operations, while the network remains available, if limited. The third is a facility fire in the Annenberg building. The fourth is a major hardware equipment failure which renders part or all of the running services inoperable. The last is a security intrusion where some or all of the user community has data that is destroyed or manipulated by an outside, intelligent agent. Hacker activity on college campuses is very common, particularly at Caltech where the campus (by longstanding demand) has a mandated open-Internet policy. Since most CMS users work both remotely and in the facility, and since many off-site users and collaborators require access for various research purposes, the likelihood of compromised user credentials is very high. While two-factor authentication systems could reduce this likelihood, the maintenance of physical hardware tokens is substantial and loss of those tokens by traveling members of the community could cripple their ability to work when offsite.

System Dependencies

Each service in the department has dependencies, some of which are outside the scope of the technical staff's responsibilities and areas of authority. The dependency tree in Figure 1 illustrates:

For the server room (1T1 Annenberg) to be at optimal capacity, the dedicated transformer must be operational, and the chilled water supply must be uninterrupted. The dedicated transformer distributes power to three power distribution units, which in turn distribute power to the 20 racks inside the server room. A failure of the main transformer or the campus chilled water supply requires services to be offline until repairs have been completed by the Physical Plant department.

Both the Boot Order and Shutdown Order are posted inside the 1T1 server room. They are not attached herein as they are subject to change with more regularity than this document.

Backup Regimen

All necessary information to restore any service to working order, from data stores to configuration information, is stored on the central file server (the Network Appliance). This machine performs snapshots on the following basis: every hour, keeping the last 6 hourly snapshots; every day, keeping the last two daily snapshots; and every week, keeping the last two weekly snapshots. In addition, the central file server is backed up to tape with a level zero backup once a month, and incremental backups every week. One copy of a monthly backup is stored offsite at the Citibank branch on the corner of Lake Avenue and Del Mar every four months.

Incident Checklists

Incident Type: Earthquake. Probability: nearly statistically certain within 30 year window. Impact: very high.

Incident Mitigations: Do not evacuate the building unless the structure is critically weakened, as the danger from falling glass and debris is much greater than the danger of the building structure being critically damaged. Assign runners to report to the campus incident command post, located in the Campus Planning Office. After ensuring human health and safety, check integrity of chilled water piping (under *Incident Subtype: Chilled Water/HVAC* below). If the piping is physically secure, deal with the current power situation (see *Incident Subtype: Power Issues*) and then other chilled water issues, if any.

Incident Type: Facility Fire. Probability: very low. Impact: variable.

Incident Mitigations: In any fire alarm, the building should be evacuated to Moore Walk, on the north side of the structure. The automatic shutdown script should be activated to begin graceful shutdown of the server room; if this is not possible trigger the emergency room shutdown switch inside the server room to prevent additional damage to the equipment. The server room (1T1) is on a dry pipe standby system. It is therefore unlikely that a fire will trigger the sprinkler system unless it begins in the server room itself. Each server cabinet is somewhat sealed from major water damage as the enclosures provide a level of physical protection, unless the cabinets exceed their temperature threshold and the doors open automatically (which would indicate that the equipment inside is what has triggered the fire alarm in the first place), however physically powering the machines off provides an additional level of protection from water damaging the equipment. Should a cabinet suffer a leak, unrack any equipment that has been exposed to water, open the case, and allow the equipment to dry before attempting to power it on.

Incident Subtypes

Incidents below are listed in decreasing order of severity in each subtype. The first two subtypes include power and chilled water/HVAC. The third concerns major hardware failures. The fourth concerns network outages. The last is nefarious human agency (e.g., hacker attack).

Incident Subtype: Power Issues

Incident Type: Worker safety issue, personnel in 1T1 are suffering from electric shock.

Incident Mitigation: Immediately trigger the emergency power-off switch inside the main door. This will cut power to all electrical circuits in the server room, including the UPS-backed up power circuits. Call Public Safety at x5000 from the nearest physical phone; do not dial 911 from the campus phones, as the public safety office can direct emergency responders accurately to your location and the 911 dispatcher may not be able to provide accurate directions to EMT personnel due to the lack of street address references. If a physical phone is not available, call 626.395.5000 from a cellular phone, rather than 911 (for the same reasons). If you are trained, provide CPR or other first aid. Once the human safety issues are resolved, investigate the cause of the electric shock before reconnecting any power. Disconnect any suspect machines from power sources prior to re-engaging the PDU systems. Cooperate with campus Physical Plant and the Electrical Shop as they have authority in these incidents. *Only restore power when given authorization by an electrician.* Once power is restored, follow the Boot Order to restore normal operations.

Incident Type: Failure of the campus power grid, or failure of the dedicated transformer for the server room. Probability: extremely low. Impact: very high.

Incident Mitigation: Ensure timely execution of the automatic shutdown scripts for all machines in the server room, particularly the Network Appliance and the SpectraLogic tape library. If necessary, manually execute the Shutdown Order. Sufficient time should remain on the UPS battery power to enable graceful shutdown of the cluster. When building power is restored, follow the Boot Order to restore normal operations.

Incident Type: Interruption to the power service in one of the three main PDU units. Probability: extremely low. Impact: variable, likely low.

Incident Mitigation: If the PDU in question is unit #2 or unit #3, leave the unit inoperable until such a time that Caltech facilities can restore operation. Extended outages are acceptable. If the PDU in question is unit #1 and service cannot be restored in less than 2 business days, power down all equipment attached to PDU #2. Ensure worker safety and power down PDU #1 and PDU #2 on the main circuit panel. Disconnect the under-floor cables that feed power to cabinet(s) 2-3 through 3-3. Disconnect the under-floor cables that feed power to cabinets 1-1 through 1-5. Cross-connect the cables formerly connected to PDU#2 to cabinets 1.1 through 1.5. If time is critical, do not worry about reconnecting power cables from PDU #1, as the cables are self-terminating and can remain safely under the floor with no connections. Restore the main breaker for PDU #2 to the “on” position. This should re-energize the power distribution units in cabinets 1.1 through 1.5, and enable a restoration of the primary infrastructure. Follow the Boot Order to restore services to normal operations. Re-label PDU #2 to indicate which cables are connected to which cabinet using the existing naming scheme. Make changes in the Boot Order documentation to ensure proper restoration in the event of a second incident. Mark all printed copies of this document as obsolete and discard.

Incident Type: Failure of the Uninterruptable Power Supply in Cabinet 1.3, which houses the main infrastructure. Probability: extremely low. Impact: negligible.

Incident Mitigation: None immediately necessary. Machines dependent upon the UPS are all dual-power-supply capable, and are also connected directly to building power. Contact the UPS manufacturer for warranty repair (see the Appendix for service contact information).

Incident Subtype: Chilled Water/HVAC Issues

Incident Type: Failure of the campus chilled water supply. Probability: extremely low. Impact: moderate.

Incident Mitigation: Contact campus facilities and determine the likely duration of the outage. It is possible to run rudimentary services (email services, the primary department web server, one domain controller, and one interactive login server) without chilled water supply for an extended duration (even up to several weeks) with minimal risk to the equipment from thermal excess. Ensure a timely shutdown of all other equipment in the main server room, following the Shutdown Order, to reduce excess heat generation.

Incident Type: Failure of the main pipe into the central server room. Probability: extremely low. Impact: very high, potentially ongoing damage.

Incident Mitigation: A structural failure in the main coolant pipe represents a serious hazard to the equipment in 1T1, as the pipe is 8" in diameter and the water flow through the pipe is in excess of 100 gallons per minute. Note, however, that the under-floor power cabling is rated for aquatic use, so there is no immediate safety hazard represented by a major leak in the coolant system. Should the pipe be ruptured by structural weakness or earthquake, an audible alarm will be triggered by the HVAC unit and an alert will be sent out to the Physical Plant department via the BMS (building management system). A large rupture should trigger a shutdown by Physical Plant; if this does not occur then the systems administration staff have a key to the first floor janitorial closet. Enter the closet, and turn off the valves which control water flow through the pipe. Contact Physical Plant and report the incident. Without any water flow through the chilled water piping, heat exchange will be limited to air transfer. It should still be possible to run the skeleton services outlined in the previous scenario (email services, the primary department web server, one domain controller, and one interactive login server), but care should be taken to monitor the heat load.

Incident Type: Failure of the valves or cooling fans in one of the Rittal refrigeration units, or an entire refrigeration unit. Probability: extremely low. Impact: low.

Incident Mitigation: Likely none are immediately necessary. Since all of the cabinets in a single row are on a shared plenum, and since each cabinet is designed to handle a heat capacity of a compute node output projected out to 2025, failure of even an entire cabinet can likely be handled without downtime. Monitor the set points for all cabinets in the affected row. Contact Rittal authorized service personnel listed in the Appendix to affect onsite repair. A spare parts kit exists for the Rittal refrigeration unit and is stored in 112 Annenberg if the need to change parts is dire.

Incident Subtype: Major Hardware Failure

Incident Type: A failure of the Network Appliance central file server. Probability: extremely low. Impact: near full outage.

Incident Mitigation: With the central file server offline, most services will fail to operate in even a degraded mode. Follow the Shutdown order and turn off all services except the CMS content management server and the DNS server, which are not dependent upon the Netapp. Contact the IMSS web services group and have them update the main page to indicate that there is an unscheduled outage (contact information is in the Appendix). Open a trouble ticket with Network Appliance; we have a four-hour response time contract. When the Netapp is returned to normal operation, run integrity checks to ensure that no volumes need to be restored from tape before restoring any services.

Incident Type: A failure of the SpectraLogic tape library. Probability: extremely low. Impact: very low to negligible.

Incident Mitigation: Likely none are immediately necessary. Contact SpectraLogic an open a warranty case to restore normal operations. If repair is likely to extend past the automated backup schedule, run a manual level zero backup when normal service is restored.

Incident Type: A failure of one of the other rackmounted servers. Probability: moderate. Impact: low.

Incident Mitigation: Likely none are immediately necessary. Most of the services that reside higher in the dependency tree (LDAP and Active Directory) are multiply redundant at the host level. A failure of a single host thus does not represent a user-facing downtime and is acceptable. A failure of the other hosts represents in those other cases acceptable downtime for the individual service.

Incident Subtype: Network Outages

Incident Type: Campus DNS failure. Probability: low. Impact: very high to external-facing services.

Incident Mitigation: None possible (included for reference only). While CMS services will remain internally available (as the CMS department runs its own DNS service) outside services will be unable to find the proper referral for the CMS zone if the campus DNS servers are non-responsive. Users will be able to send email, for example, to domains outside Caltech, but will be unable to send email to other domains inside the Caltech namespace and it is very likely that outside (non-Caltech) email servers will be able to perform lookups for the CMS mail cluster.

Incident Type: Failure of the main building router. Probability: extremely low. Impact: very high.

Incident Mitigation: Outside the scope of this document (included for reference only). If the main building router is offline, no network traffic is possible and both the outside-facing services (including the department CMS site) will be accessible. There is no service level agreement with campus IMSS, and thus no contractual obligation for the campus network group to respond within any particular time window. It is, however, unlikely that this will cause a major disruption unless caused by an Institute-wide event such as a major earthquake, as the network team prioritizes building router outages above all other service requests except border router outages.

Incident Subtype: Malicious Human Activity

Incident Type: A loss of data integrity due to long-term malicious activity. Probability: extremely low. Impact: very high, constrained to a subset of the user community.

Incident Mitigation: If a user reports corrupted files, their account should be immediately locked out by changing the shadowExpire entry in the LDAP service, and a new password should be given via methods where authentication can be reasonably guaranteed. The user's data store should be forensically examined, starting with time stamps on the known-to-be-bad file(s). Contact campus Information Security and report details that are available, as the Information Security group keeps extended logs on the border routers and may be able to assist in tracking down the hacker's pattern of behavior. If it can be reasonably determined that the activity is very recent, the user's data store can be restored out of the snapshot capability of the Netapp. Otherwise, restoration from tape backup may be necessary. It is also possible that the user will prefer to examine their data store themselves to determine which files need to be replaced.

Incident Type: A security incident in a CMS-maintained service. Probability: low. Impact: low.

Incident Mitigation: Contact campus Information Security and request that the host involved be blocked at the campus border routers, the building router, or the building switch, depending upon the severity of the incident. Typically incidents of this type represent resource hijacking, as opposed to corporate espionage which is more common in the private sector, so cutting the machine off from the outside world, but allowing access from the internal network, can facilitate forensic investigation. If possible dump the disk image to a file for later analysis. Re-install the suspect machine using the canonical install, and restore the configuration using the configuration management system. Check for zero-day vulnerabilities in all public-facing services running on the host. If patches or workarounds are available from the software vendor, apply them and restore the service. If they are not available, the service may have to be restricted to the campus network or the building to allow the service to be re-established with reasonable security, until security fixes become available.

Incident Type: A security incident on a CMS workstation. Probability: moderate. Impact: low.

Incident Mitigation: Virtually all incidents of this type represent end-user activity that is not easily reproduced by an attacker (a successful phishing attack, or a drive-by script on a malicious web site, etc.) Also, in most cases, substantial investigation will yield little constructive results. If possible dump the disk image to a file for later analysis. Re-install the suspect machine using the canonical install, and restore the box to normal service.

Appendix

Contact Information

Campus Contact: Public Safety Office
X5000 in emergencies

Campus Contact: IMSS Web Publishing Service
Ariel Fox x2524, Cynthia N. Kiser x5710

Campus Contact: Information Security Office
Ruthanne Bevier x2671, Greg Grasmehr x3576, security@caltech.edu

Campus Contact: Physical Plant
Phil Vaziri x 6096 – mechanical engineering (HVAC, chilled water)

Mike Anchando x4999 – electrical shop supervisor

Vendor: Network Appliance (Netapp central file server)

Service Tag/Serial Number:

Date of Purchase: June 2009, current warranty expires in June 2012

Vendor Contact: **Carol Orosco, Insight Investments, corosco@insightinvestments.com, 949.439.0052**

Vendor: SpectraLogic (Tape Library)

Service Tag/Serial Number:

Date of Purchase: June 2009, current warranty expires in June 2012

Warranty/Service Contract Number (if applicable):

Vendor Contact: **Carol Orosco, Insight Investments, corosco@insightinvestments.com, 949.439.0052**

Vendor: Rittal (Server Cabinets and Refrigeration Units)

Service Tag/Serial Number: Various, not applicable

Date of Purchase: June 2009

Vendor Contact: James Jung, Rittal Corporation, jjung@rittal-corp.com, 310.989.6949

Vendor: Dell Computer, Inc. (Assorted Servers)

Service Tag/Serial Number, Date of Purchase, Warranty/Service Contract Number:

Various, see online hardware inventory at http://inventory.cms.caltech.edu/glpi

Vendor Contact: Jeff Cochran, Dell Account Executive, jeff_cochran@dell.com, 949.637.6506