Executive Summary:

The purpose of this review/assessment is to identify the cause of the December 7, 2005 outage of the University of Alaska Fairbanks (UAF) email system. The review/assessment identified two major hardware failures as the main cause of the outage and four other factors that prolonged the recovery period.

The two major hardware failures are: The mirrored system disk on the mail server failed, requiring a replacement; a Storage Area Network (SAN) Input/Output (I/O) module was defective, causing extremely slow I/O response and confusion among responders trying to restore the data and system.

The other four factors affecting the recovery effort were:

1. The first responders to the outage did not communicate the problem in a timely manner to management, which could have allocated more resources and personnel to the recovery effort more rapidly;

2. There was no established procedure at UAF Division of Computing & Communications (DC&C) to address this type of crisis, which often leads to "shoot from the hip" remedies that serve to lengthen the recovery period rather than shorten it;

3. Within UAF DC&C technical expertise with the SAN is deficient;

4. Deployment of a redundant email server(s) was postponed due to staff shortages.
Review/Assessment Methods

The email from Fred Smits titled "Root Cause Analysis: UAF E-mail Problems (12/07/05-12/08/05) - Update #1" was reviewed.

A set of fourteen questions related to the outage and recovery effort was sent to each member of the two recovery teams formed by Fred Smits. The responses were reviewed and analyzed. The questionnaire is included in Appendix A.

Person-to-person and email discussions with some of the team members and help desk personnel were also included in the analysis.

Knowledge derived from the participation in the "UAF DC&C Email Stabilization Project" since May, 2005 also was used in the review.

Review/Assessment Analysis

Hardware failures are unavoidable. The failed system disk was replaced rapidly, but the defective I/O module inside the SAN was not positively identified until December 21, 2005. The defective module was replaced shortly afterward. The entire system appears to be whole again. The second response team built a redundant system for testing, and if necessary, to replace the primary email server.

Lack of communication at UAF DC&C is a continuing problem. It was noted a number of times during the "UAF DC&C Email Stabilization Project." Before the outage, when tape backup slowed significantly (an indicator of I/O problems) this should have triggered a warning and notification to upper management, but it did not. Also, during the outage, other staff experienced with SAN were not consulted until the next day and the SAN vendor support was not contacted for several hours. If upper management had been notified earlier, a better response plan could have been devised and response teams formed sooner to shorten the outage.

Technical expertise with the SAN is deficient. UAF DC&C staff are not proficient with SAN diagnostic tools. Hewlett Packard (HP), the SAN vendor, should have been consulted the moment the I/O problem surfaced. Training in this area needs to be a high priority.

The absence of a written procedure for constructing the email system, for resolving hardware, software and Operating System (OS) defects, and for managing this type of crisis were the contributing factors to the prolonged outage. For example, with proper documentation, the Network File Service (NFS) mount point issue should not have masked the real defect in the SAN I/O module. Without procedures, even informal ones, or some planning and testing, the first responders tried inadvisable remedies such as expanding the Redundant Array of Inexpensive Disk Set (RAIDset) on a defective SAN, which prolonged the outage by many hours. The absence of escalation procedures at UAF DC&C is also a contributing factor. If the problem was escalated earlier, additional technical assistance and advice could probably have stopped the RAIDset expansion and restore the system earlier. An escalation procedure, existing or newly developed, should be introduced to UAF DC&C staff if the merger does not.
A second email server (mail2) was scheduled to be in production on November 6, 2005. For various reasons, UAF DC&C failed to meet this schedule agreed to in the "UAF DC&C Email Stabilization Project" meeting in October, 2005. A second email server could have alleviated the severity of the outage since half of the users would be served by mail2 with a different and functional I/O module within the SAN.

A few members of the response teams voiced their opinion that the currently deployed SAN may not be a right fit and that it should be upgraded to an enterprise level SAN. HP engineers expressed similar views. While the current SAN can handle the I/O requirements of the mail server most of the time, during peak usage I/O traffic generated by the Internet Message Access Protocol (IMAP) clients can cause a bottleneck. The SAN should be upgraded to provide the extra headroom.

In general the Help Desks at UAF DC&C and Statewide (SW) worked well throughout the crisis. Initially information was not passed to the Help Desks in a timely manner. Communication improved as the crisis continued. However, the Help Desks would prefer more frequent updates from the response teams. There was also a misunderstanding of who was acting Chief Information Technology Officer (CITO) for the Office of Information Technology (OIT) during the crisis: The Director of OIT Systems Services or the Director of OIT User Services? This confusion had some impact on how well the Help Desks disseminate useful information to the consumer base. The Help Desks are revising and refining existing escalation and outage notification procedures as a result of this crisis.

Current Status of UAF DC&C Email System

There is no known defect with the current system. The reliability issue a year ago was caused by a known Linux kernel bug associated with the ext3 journaling file system with user file quota activated. This issue was resolved in September 2005 when the operating system was updated from Red Hat Release 3 to Red Hat Release 4. The outage in December 2005 was caused by a failed system disk and prolonged by a defective I/O module in the SAN connected to the mail server. The defective part was replaced in late December and I/O performance is verified by HP to be normal. The system should be stable and ready for the spring semester.

During the December 2005 outage a redundant mail server (mail2) was built for testing and verification, and if necessary, replace the primary mail server in case of failure. The redundant server was originally scheduled for deployment in November 2005. UAF DC&C failed to meet this schedule due to a staffing shortage according to the principal analyst responsible for building mail2. The redundant server probably would have alleviated the severity of last month's outage.

Comments on the UAF DC&C Email Stabilization Project

In mid-May 2005, the UAF DC&C Email Stabilization Project had its first meeting. For 4 months UAF DC&C was not able to pinpoint the cause of extremely poor performances when user file quota was enabled. After gathering the necessary technical information, a Google search on the key words "ext3 + quota" identified a documented and well known kernel bug in Red Hat
Release 3, it took about 15 minutes to find this crucial information. After a few more meetings, the team drafted an action plan with benchmarks and schedules for resolving this problem. UAF DC&C personnel were charged with implementation and SW personnel were charged with providing assistance if necessary. The basic plan was to upgrade the operating system to Red Hat Release 4, which contains the kernel bug fix, by the end of July. However this plan was not completed until October. Various technical and staffing problems at UAF DC&C were the cause of the delay. However these problems were not shared with other team members from SW until repeated queries. The next phase is to build redundancy into the system just in case the primary system is down. The basic plan drafted in October is to create an identical system by the end of November. UAF DC&C was charged with building the redundant system. Again, UAF DC&C missed the date and did not communicate why, nor asked for help. It was only after repeated queries did UAF DC&C acknowledge staffing deficiencies and reassignment that caused missing the timeline. Then the December 7 outage happened. In retrospect, UAF DC&C personnel should not have been placed in charge of implementing the plans.

The "Project" revealed a number of problems at UAF DC&C. Foremost of all are poor communication, inability to solve the problem, unwillingness to ask for assistance, lack of urgency, poor or no planning, and insufficient testing and verification. These problems indicate poor management and leadership. These problems were discussed and emphasized at Project meetings many times. Yet they continue to surface, not surprisingly even during the December outage. These problematic practices are ingrained among most UAF DC&C team members and will be difficult to eradicate. However, the Project did introduce better practices to them. The continuation of the Project involving them, as well as the merger, over time, could have positive effect if SW team members play a more prominent and leadership role.
## Appendix A

The following questions were sent to:

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doug Knight</td>
<td>UAF DC&amp;C, no response</td>
</tr>
<tr>
<td>Daniel LaRoe</td>
<td>UAF DC&amp;C</td>
</tr>
<tr>
<td>Walker Wheeler</td>
<td>UAF DC&amp;C</td>
</tr>
<tr>
<td>Kerry Digou</td>
<td>SW</td>
</tr>
<tr>
<td>John Mitchell</td>
<td>SW</td>
</tr>
<tr>
<td>Jo Knox</td>
<td>UAF DC&amp;C</td>
</tr>
<tr>
<td>Jim Williams</td>
<td>UAF DC&amp;C</td>
</tr>
<tr>
<td>David DeWolfe</td>
<td>SW</td>
</tr>
<tr>
<td>Ricky King</td>
<td>SW</td>
</tr>
<tr>
<td>Richard Machida</td>
<td>UAF DC&amp;C, no response</td>
</tr>
</tbody>
</table>

1. What caused the outage? (In your opinion)

2. Was there any early warnings of an impending failure?

3. Were there hardware deficiencies? If the answer is yes, what was deficient?

4. Was hardware/software mis-configuration a factor?

5. Were there any software or application defects that contributed to the outage or delayed the recovery?

6. Were established procedures for addressing this type of failures adequate? Was there a written remediation plan?

7. Were staffing levels sufficient to diagnose the failure and to perform the recovery effort in a timely manner?

8. Were there any institutional or departmental processes and procedures that may have contributed to the outage, or may have delayed the recovery effort?

9. Did the "UAF DC&C E-mail Stabilization Project" have an impact on the outage and recovery?

10. Would placing a second mail server (mail2) in production alleviate the severity of the outage?

11. Were there other human or technical factors played a role in the outage or recovery?

12. Would preventive maintenance help?

13. Was there technical knowledge deficiency among staff responding to the outage and recovery?

14. Comments and suggestions?