

# CGD Compute Infrastructure Upgrade Plan

## 1.0 Introduction

The CGD Information Systems (IS) Group (ISG) is charged with providing a modern, progressive, and stable computing environment for the CGD user community. Scientists, engineers, and support staff should be able to concentrate on their individual duties without worrying about systems administration. The current infrastructure is being redesigned and upgraded to meet this goal. Achieving this on a restricted budget requires careful planning, time for implementation, and communication between the user community and the systems staff.

Services provided by infrastructure systems will be distributed to provide ease of recovery, ease of upgrade, and flexibility to meet the growing needs of the Division. Hardware will be reused where possible, and eliminated if beyond reasonable use. Automation of services (software builds, user addition/deletion, etc) will be key to managing the growing demands.

## 2.0 Future CGD Computing Trends

Written surveys from 2001 and meetings with each of the sections have revealed a wide variety of needs, desires, and deficiencies. Each section has a unique set of needs that often do not complement other sections. The one common factor among all sections is that more of everything will be needed to meet the common CGD goals:

- More disk space for project data, home directories, and e-mail.
- More Linux support to take advantage of the price/performance x86 hardware offers.
- More computational processing power for modeling development.
- More Microsoft products to support presentation/documentation efforts.
- More laptops (MS and Linux) for travel and home use.

## 3.0 General Problems

Computers have a useful life span that averages three years. The majority of the CGD computer infrastructure has exceeded this by at least one year, and in some cases by as many as three. The overall design of the computing facilities has not changed since its inception in 1996, but has been slowly added onto in a piece-meal fashion. The number of users (and therefore computers) to be supported has also grown. This combination has created a number of server interdependencies that cannot be easily maintained and have stretched the design to its useful limits.

Two specific problems face CGD computing: the concentration of services on three core servers; and the proliferation of small disks. The overall result has created system instability that can (and has) completely disabled computing division-wide.

### 3.1 Concentration vs. Distribution of Services

Several core functions/services must be provided in any computing environment. Examples include e-mail, time synchronizations, common disk access, DNS resolution, and web services. Two views exist regarding the proper organization of computers to provide these services: concentrate the services on the fewest number of servers possible, or distribute like services across multiple systems. Both methods have advantages and disadvantages.

Concentrating services on a single or a few compute servers has the advantage of low cost investment in hardware. This approach is generally acceptable in small organizations where downtime for maintenance or upgrades can be arranged without major impact to the user community. The extreme example of this situation is a home network where one (possibly a few) computers are involved, and the work can be performed in the evening or a weekend, without serious impact to the users.

Complexity and difficulty in maintaining computer systems grows as more services, users, and systems are added. Services are distributed across multiple systems to aid in stability. Redundant servers are often

implemented to provide load distribution and/or automatic failover. Upgrades of individual systems becomes easier when only a few services need to be addressed and possible interactions between software groups are eliminated.

Bearmtn provides 10 major services to CGD: e-mail, slave DNS, slave YP, print server, SAMBA server for PC interconnect, NFS for /home and /usr/local, DHCP server for PCs, NTP, backups, and autoclient server. The number of services concentrated on a single server makes upgrades extremely difficult: upgrading one service could easily cause the others to break. Upgrading the operating system would be a significant task that would involve upgrading the autoclients, another 25 systems, as well. The probability of a successful upgrade given these interdependencies is extremely low. Greenmtn and Goldhill are not as heavily loaded, but still provide too many services.

### **3.2 Small Disk Proliferation**

The majority of computer hardware problems are related to components with moving parts. The list is extremely short: disks, fans, and cable connections. External disk drives possess all three of these problems and were added to various systems in an unorganized manner as projects required them. No guidelines were provided for the purchase of new disks and no framework provided. The cumulative result is the current state of daisy-chained disks on servers and the proliferation of NFS mounted disks in the Division. The instability caused by this is apparent in a recent Greenmtn crash caused by the slight nudging of an external disk cabinet by a worker in the machine room. Division computing was restored to normal after 2 hours.

Additional problems with the disk infrastructure are caused by the use of software RAID<sup>1</sup> to support large data areas. Implementing RAID in software causes additional overhead for the operating system and provides only a small amount of data protection in the event of a crash.

### **3.3 Effects on Division Computing**

Service concentration and the number of small disks on systems currently affect all computer infrastructure upgrade and planning decisions. More central computers will be implemented to distribute the service load and provide redundancy where possible. Existing systems will be reused if appropriate.

A RAID framework for project data will be implemented. The new system will be hardware based to provide faster response and more reliable data access. Software RAID will be eliminated in the infrastructure systems.

### **4.0 Operating System/Architecture, Software, and Vendor Hardware Support**

The number of supported operating systems and hardware platforms directly affects the complexity of the computing environment and the number of IS staff members necessary to maintain it. The diversity of supported systems is determined by a compromise between the computing requirements of the scientific and support staff, and the number of IS staff available. What is supported is flexible and will change to keep pace with the Division needs.

The definition of "support" established in CGD is "...the software is kept up-to-date regarding new releases of the software, and that system problems reported in the software will be analyzed and, when possible, repaired. The systems staff cannot be expected to be experts in the function and usage of all supported software, although there are varying degrees of expertise with different packages."<sup>2,3</sup>

### **4.1 Operating Systems and Architecture**

Currently supported operating systems and architecture consist of the following:

---

<sup>1</sup> RAID: Redundant Array of Inexpensive Disks. RAID-0: Striping. RAID-1: Mirroring. RAID-3: One parity disk per set. RAID-4: One asynchronous parity disk per set. RAID-5: Distributed parity across all disks.

<sup>2</sup> Sitongia, Leonard, CGD Supported Software, 1999, page 1.

<sup>3</sup> This document needs to be updated to reflect software currently in use, and expanded requirements.

- Sun Microsystems Solaris 2.6, 2.7, 2.8 on SPARC architecture.
- SGI Irix 6.5 on MIPS architecture.
- RedHat Linux 7.3 on x86 architecture.
- Microsoft Windows 2000 on x86 architecture.
- Microsoft Windows 98 on x86 architecture.

x86 architecture includes servers, desktops, and laptops.

Apple Macintosh OS X support is planned.

Support for Windows XP is not planned at this time due to recent results of testing by the IS staff.

#### **4.2 Software Support (Solaris, Linux; /usr/local, /contrib)**

The configuration of software into operational installations in /usr/local and test software in /contrib is a good method for introducing upgrades. This method in CGD disintegrated after being ignored for a long period of time. Incompatibilities are beginning to surface with old software compiled under Solaris 2.6 not running under Solaris 2.8 (ex. Meeting Maker).

Both the operating systems and the software need to be updated at the same time to solve the problem. (OS upgrades are addressed in section 4.8.) /usr/local and /contrib will be rebuilt from scratch on a dedicated build system running Solaris 2.7. The IS staff will be distributing a list of software to be built and maintained. Once the new /usr/local and /contrib are built, they will be mounted on a few production servers for testing. The current /usr/local and /contrib will be held in reserve if fall back is necessary. Automated software will be introduced to find new software releases and install them as necessary.

#### **4.3 Windows and Unix Co-habitation**

Windows and Unix co-habitation problems revolve around 2 central issues: data exchange and reading attached e-mail documents. Three possible scenarios exist for working in both worlds:

- 1) Use a Unix system with tools for reading e-mail attachments
- 2) Use a PC or Mac system that also has software to interface with Unix systems
- 3) Purchase both Unix and PC/Mac systems
- 4) Dual boot a system

All methods are being used within CGD to varying degrees of success. Option 3 is ideal from a support and usage standpoint, but can be prohibitively expensive. Dual booting a system is time consuming and rarely satisfies user needs: the system spends most of its time booted in a primary operating system and is rarely booted into the secondary because of the time involved.

Servers must also be correctly configured to interact with other operating systems. Bearmtn is the current SAMBA server used for sharing Unix file systems to PCs. Bearmtn is not fast enough to service PC requests and the SAMBA software is not configured for correct authentication. Additional problems arise with the Admin staff having their home directories mapped to Bearmtn. Overall performance has degraded to the point of annoyance, and on occasion, dangerousness, with the loss of work.

A new Unix server (Tablemtn) has been purchased to replace Bearmtn as the SAMBA server. The current Division Windows 2000 server (CGDW2K) will be rebuilt as a true domain server, and SAMBA will be configured to authenticate off of the new CGDW2K. This will solve the authentication, browser, and stability problems with the PCs. Admin staff home directories will be moved from Bearmtn to CGDW2k for performance and storage capacity.

#### **4.3 Preferred Hardware Vendors**

The choice of a hardware vendor is primarily focused on x86 architecture due to the numerous vendors available. (SPARC hardware is only available from Sun Microsystems). Each vendor has its own proprietary method for designing systems. Reducing the number of vendors drastically reduces the number

of quirks to be dealt with. Use of an established major vendor also increases the amount of hardware support available through service contracts.

Dell has been chosen by the CGD/ISG (and much of UCAR) as the preferred vendor for x86 servers, desktops, and laptops. UCAR/Contracts has negotiated favorable purchasing contracts, and Dell offers standard on-site service for hardware issues.

#### **4.4 Autoload Servers (Solaris Jumpstart and Linux Kickstart), Sun Autoclients**

Solaris and Linux both support network installation and upgrades. (The Solaris method is Jumpstart; the Linux method is Kickstart.) The automation of the operating system installation process shortens the time of systems installation, and reduces the support time required by IS staff. This template approach provides a uniformity that is desirable for both the ISG and the user community: what works on one system will work on all systems.

Sun has discontinued support for the Autoclient process. CGD Autoclients are being converted to standalone systems (and the OSs upgraded) via the Jumpstart process. This will provide a supported means for upgrading CGD workstations and increase system stability in two areas: workstations will run local operating systems, decreasing the reliance on the servers; and servers will become more stable with the elimination of the NFS traffic caused by serving multiple copies of the operating system to workstations. /usr/local, /contrib, and some /opt software will still be NFS mounted from central servers for uniformity and ease of maintenance.

New procedures and software will be implemented by the IS staff to prevent stand alone systems from drifting away from standard installations. (Compute nodes on the cluster are currently being maintained in this manner.)

#### **4.8 Multiple OS Levels (Solaris 2.6, 2.7, 2.8)**

CGD is now supporting three different versions of the Sun Microsystems, Inc. operating system: Solaris 2.6 (1998), Solaris 2.7 (1999), and Solaris 2.8 (2001). This has created several problems for both the systems staff and users. Instances of software compiled under 2.6 breaking under 2.7 or 2.8 are becoming more frequent. SSH (secure shell) and Meeting Maker are two such examples. Features of the more recent operating system (primarily 64 bit addressing and file system journaling) are also being denied on computers still running Solaris 2.6.

There are two problems preventing the upgrade of central servers and desktop autoclients: the autoclients' software has no clear upgrade path from 2.6 to 2.7 (or higher), and the autoclients download their OS from Bearmtn, which also serves DNS, YP, SENDMAIL, IMAP, NFS (/home), SAMBA (PC access), and NTP. Any attempt at upgrading the operating system on Bearmtn in its current configuration would result in a long week of downtime for the entire Division trying to rebuild the autoclients.

Services are being migrated off Bearmtn to new servers. This will allow individual systems to be upgraded without the additional complication of software interaction. The autoclients are being converted to stand-alone systems that will download the OS ("jumpstart") off of a separate server once at installation time. A Division roll-out of Solaris from 2.6 to 2.7 began in December, 2002, and will be completed in Spring, 2003. All Sun workstations will be stand alone (approximately 40) at this time, and should be more stable.

Servers running Solaris 2.6 (Bearmtn, and Greenmtn) will be upgraded last to avoid autoclient dependencies. Once all Solaris systems are up to 2.7, a Solaris 2.8 upgrade campaign will begin. The time estimate for this is Fall, 2003.

#### **5.0 Upgrades**

Infrastructure upgrades will focus on distributing common services across several machines and upgrading existing hardware to support computing growth. This will make future maintenance of the systems easier by removing dependencies, stabilizing servers, and limiting the damage caused by a server failure. The prime example of this are Bearmtn and Greenmtn.

Bearmtn and Greenmtn were originally configured with the intent of providing hardware redundancy for Division computing: if one system went down, the other could assume responsibility for compute functions. What developed was quite different. Each system has evolved to providing unique core services without the possibility of redundancy. Critical interdependencies were created between Bearmtn and Greenmtn. The two systems became entwined to the point that if one system crashed, the other would, also.

A large portion of the dependencies, but not all, were fixed in May, 2002, by the addition of disk and migration of critical services to each system. The servers have been significantly more stable since then and can be rebooted independently of each other.

### **5.1 Bearmtn**

Bearmtn is a Sun E3000 server that is 5 years old and hosts 10 common services<sup>4</sup> for the Division. Upgrading the operating system on this server has been impossible because the autoclient software is not supported by Solaris 2.7. This has prevented all other services, including SMTP and DNS, from being upgraded, leaving known security holes and stability issues unresolved.

Bearmtn will be replaced by two new systems in FY03 and the core services distributed between them. Operating system, software, and hardware upgrades will be easier to accomplish given fewer software interactions on each individual server.

### **5.2 Greenmtn Disk Space and Disk Configuration**

CGD servers currently have 2 terabytes of disk storage in the machine room. There are over 30 disks that range in size from 2 GB to 70 GB, and in age from 6 years old to less than 6 months. The disks have been added a few at a time in an unorganized, daisy-chained manner. Each connection in the chain is a potential failure point which degrades overall stability. The physical number of server connections is near exhaustion. Finally, all RAID 0 (mirroring) and RAID 1 (striping) functions are being performed by software, which degrades performance and causes additional load on the system.

IS is installing a framework of hardware RAID devices on the servers. The hardware RAID will devices provide an easier interface for management, and off load processing from the CPU. The end result will be greater reliability and uptime. As appropriate, IS will replace, at no cost to the projects, existing individual daisy-chained disks with equivalent quota in the RAID framework. Future additions to the framework will require specific disks to be purchased by project to maintain stability. The current devices being tested can hold 3 TB of raw disk space (usable space will be slightly less).

Disks attached to individual workstations are not at a critical point. However, future additions to workstations need to be for local processing only. Disk necessary for use in conjunction with infrastructure compute resources need to be in the infrastructure. NFS cross-mounting within the Division is becoming a problem.

### **5.3 Sun Rays**

The Sun Ray is a proprietary x-terminal (thin client) product developed by Sun Microsystems, Inc. The thin terminals have no disk and provide an X-window environment by communicating with a server. The goal is to provide compute resources and uniform processing environment to a large number of users while reducing the number of machines (servers) to be maintained.

Sun Rays have had qualified success within CGD. The total number of Sun Ray clients within CGD was 25 as of March, 2002. These were supported by two servers that were over-taxed with users. Performance suffered from memory leaks created by the Xsun, Netscape, and TWM processes. Since then, another 20 Sun Ray clients have been deployed, a third server added, and the primary server upgraded.<sup>5</sup> Normally, this

---

<sup>4</sup> Bearmtn services: NFS /home, NFS /usr/local, SMTP (e-mail), e-mail aliases, DNS, NIS, backups, NTP, Mailman (listserv), Autoclient server, printer server, ?, ?.

<sup>5</sup> The primary Sun Ray server (sunray1 – E250R 2x 400MHz, 2 GB memory) was replaced with a Sun Fire 280R (2x 900Mhz, 2 GB memory) in October, 2002. The new Sunray1 has provided a suitable level of performance for users, and has been able to support the remaining 20 Sun Rays.

provides an adequate environment to support 45 users. Problems still exist with run-away processes, but the additional memory has mitigated the problems somewhat. An additional 20 Sun Rays are on the shelf and ready for deployment. There are no plans for the purchase of additional Sun Rays.

The Sun Ray servers are also suffering from OS creep. Currently Sunray1 and Sunray3 are at Solaris 2.8, while Sunray2 is at Solaris 2.7. Sunray2 will be upgraded to Solaris 2.8 as part of the Division 2.8 upgrade in Fall, 2003.

#### 5.4 Compute Servers

The designated Division compute servers are Sanitas, Flagstaff, and Neva.

Host	OS	Memroy	CPU	Purchase Year
Flagstaff	Solaris 2.6	576 MB	2 x 296 MHz	1996
Sanitas	Solaris 2.6	768 MB	2 x 296 MHz	1997
Neva <sup>6</sup>	Solaris 2.7	2 GB	2 x 400 MHz	2001

Sanitas and Flagstaff have not been upgraded in at least 5 years and are no longer useful for model development. The lack of memory forces a reduction in data sets that is unacceptable. The servers have been reduced to running occasional Matlab and IDL processes.

A third compute server, Neva (UltraSPARC, 2x400MHz, 2GB memory), was added by in Oct, 2002. Additional servers will be purchased if justified. The architecture (x86, SPARC, etc) and operating system (Linux, Solaris, etc) of future purchases will be determined by the need of the user community. An experimental Linux server is planned for fiscal 2003. The timeline for additional upgrades has not been determined.

#### 5.5 Compute Clusters

A small amount of money became available during FY02 for the purchase of "small item" (i.e., < \$5000) hardware. The spending restriction for this money was that it should benefit the Division as a whole. A cluster was chosen as a means to fill a computational and architectural gap within the Division. If the cluster proved unsuccessful, the components could be broken up and re-deployed elsewhere. The cluster (Anchorage) is up and running. Future expansion/additional clusters depends on the success of this one.

#### 5.6 Backups

The addition of central servers, more Division staff, and larger data sets has stretched the current backup system to its limits. Data retention times are getting shorter as more data is being saved.

A new backup server and LTO (Linear Tape Open) robot were ordered and have arrived. The initial capacity of the system will be over 2TB of data and will provide additional capacity for backing up more user data. The danger of such a large system is that backups may be running continuously, affecting the performance and stability of other servers.

The current DLT robot will be moved to the Division Windows 2000 server to begin backing up data located on the server.

#### 5.7 Computer Monitors (Screens)

A majority of the monitors currently in use within the Division were originally purchased with the Sun Ultra 1s. Theses monitors are aging and beginning to break down. The monitors are attached to the Ultra 1s, which were removed from maintenance during 2001. Eventually, the hardware maintenance vendor will stop repairing the monitors free of charge.

Replacement monitors (LCD flat panels) need to be purchased to replace the aging monitors. Where the budget for this will be found is unknown at this time.

---

<sup>6</sup> Neva was added as a compute server in October, 2002. Neva is the decommissioned Sunray1 server.

### **5.8 Monitoring Software: Security, Health, Performance**

Systems are monitored for three reasons: security; errors and health; and performance for capacity planning. The IS group has installed a dedicated system (<http://lookout.cgd.ucar.edu>) for monitoring the central Division servers. The primary software used for monitoring performance (SARGE) and health (Netsaint) have been installed. Notification via e-mail and pager has been added and already proven useful for correcting problems before they become catastrophic. Additional software for security and capacity planning will be installed, and the server upgraded to support the additional load.

Monitoring of user's Unix systems for performance can be added by request.

### **5.9 Web, FTP, and SSH Servers within CGD**

The former web server (Goldhill, 1x167 MHz, 256MB memory) was also being used as the secure shell server (SSH) for external access to CGD systems, a general use compute server, and the Division FTP server. The actual web data was NFS (cross) mounted from Greenmtn. This configuration and high levels of use created a server that was slow, interdependent on other systems, not stable, and presented security problems.

The web and FTP server functions have been separated onto a dedicated Linux system, with the web and FTP data disk being part of the server. Additional disk space was added for web and FTP data.

SSH is being installed by default on all new Unix installations and upgrades to provide a secure communications means within the UCAR perimeter.

Additional servers for SSH and secure e-mail access from outside of UCAR will be added by end of year 2003.

### **5.10 Machine Room Electrical**

The Uninterruptible Power Supply (UPS) in the CGD machine (ML 315) room is at its limit. The addition of a single disk resulted in the UPS entering alarm mode. The UPS cannot be upgraded with additional capacity. A replacement must be purchased before the Division returns after the A-Tower portion of the MLUR project. This is expected to be a major expense. Whether the money will come from the Division or UCAR infrastructure funds is unknown.

### **5.11 Machine Room (ML 315) Space**

The CGD machine room (ML 315) is near space capacity. Future compute additions will have to be placed in individual offices, away from conditioned power (UPS) and dedicated cooling. (The CGD cluster is currently located in an office.) Additional space requirements have been submitted to the space consultants working with UCAR/NCAR for future needs, but no decisions have been made.

The MLUR project and associated moves will keep the Division in a state of turmoil for at least 2 years. During this time, new compute systems will continue to be purchased and fit into the infrastructure. Space will have to be made available once the Division has made its final move.

## **6.0 Security**

Computer security is becoming more important as hackers become more innovative. The security goal in a research environment is to increase secure computing practices without interfering with work. This involves additional monitoring, robust recovery capability, and providing secure methods for work.

### **6.1 User Login Disabling/Removal**

User accounts with active passwords (open accounts) provide access points for hackers to gain entry. The more usernames, the more chances for breaking. Open accounts also impact system resources. More usernames in the password file requires more time to authenticate users. It is also common practice to leave user files on the system as long as the account is still active. This can consume a significant amount of disk quota that could be better utilized.

Since 1996, over 800 usernames have been added to CGD systems. Over 600 are still active in the password file. Recent studies have shown approximately only 250 users have logged in during the past year.

A login monitoring system has been installed to keep track of the last time a username was accessed. Usernames older than 6 months will be disabled, usernames that have not been accessed for more than 12 months will have files saved to tape, all files removed from the system, and the username deleted. Division admins and account sponsors (if possible) will be consulted before action is taken.

## **6.2 Network Monitoring**

Discussed in section 5.2.

## **6.3 Server Security**

All systems on the network with remote login capability are vulnerable to attack. Systems located in the perimeter network are especially vulnerable. Programs to scan for changes in the system directory will be installed. Servers providing necessary but potentially insecure services (WWW, FTP) will be isolated to limit damage in the event of a break-in.

## **7.0 Summary**

Numerous upgrades have either been completed, are underway, or are planned for the CGD infrastructure compute systems. Stability and security issues are a primary concern. Additional requirements for more CPU cycles on different platforms and the explosion of data sets reinforce the need for new equipment. New equipment will also require new services to be provided by the ISG. Meeting these goals with the same level of staff and maintaining a low CSC rate will be difficult.