DEFINITIONS

Backups: A place to duplicate important data. CDs, DVDs, USB drives, other computers, tapes, centralized disk servers are all viable methods of backing up. This document is concerned with centralized, IT-managed backup systems for administrative and research data which allows off-site data storage through tape or remote disk backup. For our purposes we’ll break this down into two types of backups:

- **Live Backups:** Duplication and storage of data which people are actively working.
- **Archival Backups:** Duplication and storage of data which is not being used and is kept mainly for data retention purposes imposed by the funding agency.

Backup Devices: As stated above, there are many ways to back up your data. For LAWR’s centralized storage, we currently use Tape Libraries which store data on removable AIT-3 tapes. In the near future, the plan is to perform live backups to drives in a Disk Array and archival backups to either a Disk Array or a Tape Library.

- **Tape Library:** A unit of varying size (we have two: one is as large as a dorm refrigerator, the other the size of two minitower computers side by side) which holds a limited number of tapes that are loaded into an internal tape drive robotically. The tape library's initial cost is separate from the tapes it uses. These tapes are implemented in groups and cycled through in varying frequency. Tapes can continue to be purchased and added to the cycles as the amount of data being duplicated increases. Example: A backup of the Beaumonts research server requires 17 tapes, which are then removed from the library and stored off-site until the next backup. In the intervening days, an "incremental" backup is done of Beaumonts, backing up just what has changed.

- **Disk Array:** A server-quality computer which holds a fixed number of hard drives (however many one decides to purchase). The entire case, internals, and drives are purchased as one unit. A disk array would be stored “off-site” in another building or department. There would be no tape swapping. As disk array sizes, once purchased, are fixed, there may be more than one disk array, depending on the size and/or purpose (admin, research, archive, etc) but in general there is effectively only one tape library (just different purposes for the tapes).

LAWR BACKUPS

History: LAWR currently uses two tape libraries to perform live backups of all research and administrative (business, staff, IT) data stored on servers. The size of the data being backed up is split 88/12 Research/Admin in terms of the amount of data duplicated. Over the past few months, a problem with one of the tape libraries forced the IT staff to spend 20 hours per week minimum making certain data was being backed up. The tape libraries are repaired to a limited extent with 4 of the possible 6 internal tape drives functioning. Tapes still need to be configured, swapped, managed, and purchased to handle the growing data needs. This is inefficient in terms of employee time and department money.

Current: Live backups are happening normally. A scheduled system is being put into place to minimize the time spent switching tapes, etc. The warranty on both tape libraries expired as the
cost to renew was over $3000 and that money could be better spent on a more advanced system. Spare tape drives were added to the larger tape library to allow more than one backup job to run per night.

**Proposal:** A method of backing up data that cuts down on the time IT needs to physically spend maintaining the system. Backing to disk has various advantages and disadvantages over tape:

Benefits of moving to Disk-to-Disk backups:

1) Faster backups, faster recovery
2) No tape maintenance
3) “Off-site” by putting device in another department’s building
4) 1 TB of disks << 1 TB of tape

Disadvantages of Disk-to-Disk backups:

1) Backup disk space must be considerably more than data disk space
2) Initial cost of disk array is comparable to tape library
3) Long term archival backup requires separate tape library or a disk array.
AMOUNT OF DATA

Below are charts showing the amount of data in gigabytes that is currently being backed up from the various servers at LAWR. This directly reflects the individual amount of tapes and/or disks required to house the backed up data.
As is shown, research-related data is far greater than the data for administrative purposes. The hope is to show that data backup is:

1) Necessary
2) Expensive
3) Good to be shared
BASIC PLAN

Convert administrative and research backups to disk-to-disk as funds permit. Retain tape-based backups as necessary for data that has outgrown the disk arrays and for backup system redundancy.

Current Setup:

![Current Setup Diagram]

Desired Setup:

![Desired Setup Diagram]

- Newer tape technologies are available for faster backups and less tape swapping
- Should our current enterprise backup licensing become too expensive, the disk-to-disk method offers the free BackupPC which is used by others on campus.
PLANS FOR SPENDING:

Here are presented two timelines for spending on the new backup systems. We’re currently testing the EMC Networker DiskBackup option as, with the current pricing at EMC and the fact it integrates with the tape backup system we use, it was the first step in either timeline.

The thought is that for the initial buy-in of a research disk array the department will pay for a portion of the system. If we outgrow the initial disk array due to research data sizes, it makes sense that new arrays will be mostly if not entirely funded by research funds.

TIMELINE ONE

Phase 1:
- Research continues to use Tape Backup for at least FULL backups
- Administrative Backup moves to Disk-to-Disk
  - Software: EMC Networker DiskBackup Option: $1100 (Already purchased)
  - Hardware:
    - Rackmount Case/Internals: $1500
    - RAID Card: $500
    - Drives (3x750 GB) = (2.25 TB @ RAID 5) = (1.5 TB): $1000
    - Total: $3000
Phase 2:
- Research and Dept split costs on larger disk array
  - Software
    - EMC Networker DiskBackup Module Upgrade: $6000
  - Hardware
    - Rackmount Case/Internals: $4000
    - RAID Card: $1000
    - Drives (16x750 GB) = {12 TB RAID 6} = {10.5 TB} = $3200
    - Total: $8200

Phase 3:
- Research and Dept split costs on archive disk array
  - Hardware
    - Total: $6000
Phase 4:
- As needs arise (current tape drive fails, more backup space needed) AIT-3 Tape Drive replaced with LTO3 Tape Library or larger disk array
- Tape Library:
  - LTO3 Library Cost: $6000
  - Tape Cost: $4000
  - Total: $10000
- Disk Array:
  - Total: $4000-$8000
TIMELINE TWO

Phase 1:
- Research continues to use Tape Backup for at least FULL backups
- Research and Dept split costs on larger disk array
  - Hardware
    - Rackmount Case/Internals: $4000
    - RAID Card: $1000
    - Drives (8x750 GB) = 5.25 TB RAID 5 = 4.8 TB = $3200
    - Total: $8200
- Larger Research Servers remain on Tape Backup
Phase 2:
- Research and Dept split costs on archive disk array
  - Hardware
    - Rackmount Case/Internals: $3000
    - RAID Card: $1000
    - Drives \( (4 \times 1 \text{ TB}) = (4 \text{ TB RAID 5}) = (3 \text{ TB}) = $2000 \)
    - Total: $6000

Phase 3:
- AIT-3 Tape Drive replaced with LTO3 or larger disk array
  - Tape:
    - LTO3 Library Cost: $8000
    - Tape Cost: $4000
    - Total: $12000
  - Disk:
    - Total: $6000
SUMMARY

The timelines show two paths, the latter putting a large sum of money forward at first. If we can purchase an initial disk array for both administrative and research data that is quite large: 10 TB or more, this will handle live backups for a few years unless the needs grow larger than expected. For new research clusters that will have large data backup needs, it will be requested that the PI set aside money for backups, either for separate arrays or tapes, depending on the phase and the amount of disk array space remaining.

Other Considerations: E-Mail

Desktops and non-server computers for staff and faculty are not being centrally backed up. Much of the critical data for users is already stored on servers (Z Drives for example), however one critical area is at risk: e-mail. With these new disk arrays, we can work out methods of copying user e-mail to disk at night, where it will be officially backed up by the software. Current licensing is too high and really unnecessary for backing up workstations directly with our enterprise software (they charge per computer). There are other methods which will be cheaper.

As always, local backups to USB disks, etc. are a good idea. If you feel that losing a particular piece of data would be extremely bad, then it should be in at least two places.
APPENDIX

NOTES for IT:

SHORT TERM

Backup Tape Drives
-- Add more tape drives to ADIC Library as one of the two tape drives in library is faulty
---- We have spares from the original ADIC troubleshooting
  ■ Keep the tape drives backing up major data sets for as long as they last
---- IN EFFECT NOW

Disk-to-Disk Backup
-- Purchased EMC DiskBackup option to test (DONE)
-- Integrated into current backup software for tape cloning, etc.
-- Add arrays of disks as needed for backups
---- Incrementals for research on disk also
---- Fulls for research on tape
---- More on disk as research contributes
-- Disk arrays housed off-site
-- Look into hot-swap raid arrays for less troubleshooting
---- More expensive
-- Look into expandable archiving systems
---- More expensive

-- BackupPC Linux option
---- If EMC Networker fails or becomes too pricey
---- To be tested on Moe, spare machine
---- Another way to backup to disk, free software

-- Archival
---- Software that moves untouched-after-a-date to a tertiary server
---- Another disk array that can be added to tape backup for archival but will accumulate old data
---- Structure of directory important for keeping track of what's what.

Plan:
Disk-to-Disk for administrative servers, staff, faculty
Disk-to-Disk for incrementals of research
Tape backup for Full backups of research
Tape backup for some incrementals
Tape backup as needed for administrative servers
Archive as above

LONG TERM

Disk-to-Disk-to-Disk

EMC Networker or BackupPC
Data is backed up to disk arrays, who's cost is split with research/department
As data amount increases, more disk arrays added by whoever needs it
Look for expandable solution
Archive to disk as well

Backup Plan:
Disk-to-Disk with expandable options
If EMC pricing fails, we move to BackupPC
--Faculty pay for hardware
--Dept pays for software, maintenance, setup
Archival will be a secondary array for longer-term non-active storage

ADIC backup will continue to meet research needs until failure
--After failure, faculty need to chip in for disk-to-disk