Research Networks

Dr. Wayne Mallett
QCIF, JCU (HPC)

Traditional: HPC

- Engineering, Physical Sciences
- Mostly modeling or simulations
- Long running, multi-processor jobs
- *Small* number of large files generated
- Emphasis on CPU/memory interconnects
- Data *sometimes* shared (in a limited fashion)
- Data longevity not crucial (often project based)

Emerging: eResearch

- Life-Sciences & Climate
- Exploiting software written by others
  - E.g., MATLAB and R
  - Little knowledge of resource requirements
- A virtual infinity of short, memory intensive, single CPU jobs
- Millions of *small* files generated in short time
- International research collaboration
  - Large number of tools to achieve this
- Web interfaces so that results are publicly visible.

Emerging: eResearch

- Sensor Networks, Medicine, Indigenous Research
- Increasing demand for real-time capture/processing of life and events
- Raw and processed data to be available
- Data often to be retained *in perpetuity*
- Real-time data sourced from *slow* networks
- Much data comes in form of images or videos
- Data often needs to transmitted real-time to remote areas
- Data security is an issue (in many cases)

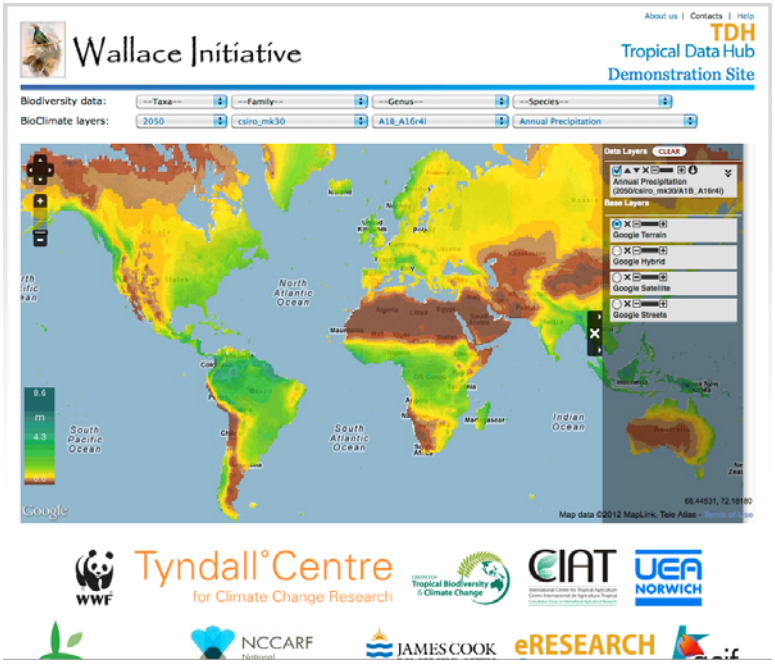
Data Movement Examples

- Smail
  - Researcher doesn't trust the network
  - Large amount of data, slow or unreliable network
- 3G
  - "Rural" locations, limited uploads/downloads, e.g., life capture
- DSL
  - "Regional" locations, e.g., remote area health
  - Remote research facilities
- Microwave/Satellite
  - Continuous data feeds, e.g., climate/environmental monitoring
- AARNET(/NBN)
  - Data archiving/replication
  - Data/Information sharing (collaboration)
- Server Interconnect (IB)
  - In-depth analysis of big data, Computational modeling (HPC)

Research Example (at JCU)

- ~66 million files at last count
- ~40TB of space consumed (expecting >1.3PB by EOY 2012)
- Submitting up to 50K jobs in *quick* succession
- Seemless multi-institutional processing desired (distances >1000km)
- Remote *cached* copies of raw files?
- Automatic recovery from *link* failures
- Data protection???

Climate Change / BioDiversity



Storage Configuration

- JCU: Multiple disk arrays
- JCU: Tape library (SpectraLogic T950, 8 LTO5 drives)
- QCIF/UQ: Storage Infrastructure (disk+tape)
- DMF (HSM) software from SGI
- Files either Regular, Dual-State, or Offline
- *Archived* files held on 2 separate tapes at JCU and 2 separate tapes at QCIF/UQ. Files transferred to QCIF/UQ via FTP (well over 50TB sent so far).
- Several gotchas in this DR environment
  - Time for migration
  - Size of cache at remote site
  - Service outages (anywhere) can require manual intervention

Network Critical

- Currently talking ~100TB at JCU. Expecting ~2PB of research data at JCU (awaiting capacity)
- Backup of *Big Data* is costly in many ways
  - How viable is a backup that takes over 7 days?
- Protection can be gained by:
  - asynchronous replication (e.g., to "the cloud")
  - version control, behind the scenes at least
- RDSI+NeCTAR to provide "research cloud" services
- Individual researchers already in "the cloud"
- Institutions investigating "cloud" alternatives

Summary

- If it exists, someone is probably using it
- Global collaborations are on the increase
- Amount of online data is rapidly increasing
- Researchers prefer their data "close" but realize it's not necessary any more
- IT services are required, not optional
- A lot of effort being directed toward **meta-data** and also presentation of "Big Data"
- Researchers live "close to the edge"