# Driving E-Research Across the Pacific Workshop - DERCAP'09
# Sydney, Australia, November 12-13, 2010

(notes from the meeting, by Celeste Anderson)

The National Science Foundation project #0441119 Translight/Pacific Wave and AARnet, Australia's Academic and Research Network, co-sponsored a workshop, "Driving E-Research Collaboration Across the Pacific" (DERCPA), aimed at facilitating collaborative research efforts between the United States and Australia/New Zealand. This was a follow-on to the successful DERCAP '07 meeting held in conjunction with the APAC meeting in Perth in 2007. DERCAP'09 was held in conjunction with the eResearch Australasia conference (November 9-13). DERCAP09 took place November 12-13, 2009 and combined presentations and discussions in six categories: Cloud Computing, High Energy Physics, Astronomy, Geosciences, Marine Systems and Sustainability/Green IT.

**Workshop Attendees are listed below**:

| | | |
|---|---|---|
| Guido Aben | AARNet | guido.aben@aarnet.edu.au |
| David Abramson | Monash University | david.abramson@infotech.monash.edu.au |
| Celeste Anderson | Univ. of So. Calif. | celestea@usc.edu |
| Ian Atkinson | James Cook Univ. | Ian.atkinson@jcu.edu.au |
| Paul Avery | Univ. of Florida | avery@phys.ufl.edu |
| Tim Axelrod | LSST | taxelrod@lsst.org |
| Greg Bell | Berkeley Labs | grbell@lbl.gov |
| Steve Black | | steveblack48@primeronline.com.au |
| Paul Bonnington | Monash University | paul.bonnington@adm.monash.edu.au |
| Jacqueline Brown | Pacific Wave | jbrown@ms.uw.edu |
| George Djoradvski | Caltech | george@astro.caltech.edu |
| Tim Dyce | Univ. of Melbourne | tjdyce@unimelb.edu.au |
| Tom Fifield | Univ. of Melbourne | fifieldt@unimelb.edu.au |
| Ryan Fraser | CSIRO | ryan.fraser@csiro.au |
| Mark Gahegan | Auckland University | m.gahegan@auckland.ac.nz |
| Tony Haymet | UCSD | haymet@ucsd.edu |
| Chris Hancock | AARNet | chris.hancock@aarnet.edu.au |
| Ron Johnson | Univ. of Washington | ronj@washington.edu |
| Nick Jones | Auckland University | n.jones@auckland.ac.nz |
| Andrew Lewis | Griffith University | a.lewis@griffith.edu.au |
| Patricia McMillan | Univ. of Queensland | patricia.mcmillan@uq.edu.au |
| Louis Moresi | Monash University | louis.moresi@sci.monash.edu |
| Dietmar Muller | Univ. of Sydney | d.muller@usyd.edu.au |
| Chris Phillips | CSIRO | chris.phillips@csiro.au |
| Roger Proctor | Univ. of Tasmania | roger.proctor@utas.edu.au |
| Alex Reid | AARNet | alex.reid@aarnet.edu.au |
| Simon Richards | James Cook Univ. | simon.richards@jcu.edu.au |
| James Sankar | AARnet | james.sankar@aarnet.edu.au |
| John Silvester | Univ. of So. Calif. | jsilvest@usc.edu |
| Robin Stanton | Australian Nat'l. Univ. | robin.stanton@anu.edu.au |

| | | |
|---|---|---|
| Leon Sterling | Univ. of Melbourne | leonss@unimelb.edu.au |
| Steven Tingay | ICRAR | steven.tingay@icrar.org |
| Rod Tucker | Univ. of Melbourne | r.tucker@unimelb.edu.au |
| Tasso Tzioumis | CSIRO | tasso.tzioumis@csiro.au |
| Anthony Waugh | Univ. of Sydney | awaugh@physics.usyd.edu.au |

## Workshop Goals

1) Highlight the Science
2) Look at the barriers to collaborations from the infrastructure perspective
3) What is the picture for the next few years
4) Identify potential demonstrations we might be able to feature in your particular discipline

Chris Hancock and John Silvester kicked off the conference by welcoming everyone, having everyone introduce themselves and explained how the informal format of the workshop would work in exploring the wonderful commonalities our countries have. One objective for Translight Pacific Wave and AARNet is to explore how as service providers we can make research easier. John gave an overview of the NSF IRNC grants and what we hoped to accomplish.

## Session 1. Cloud Computing

The first session of the workshop was on Cloud Computing and was kicked off by David Abramson's presentation on *Mixing the Grids and Clouds: High-throughput Science using Nimrod or Is the Grid Dead*?

David first highlighted the collaborative work across the Pacific that is already in progress. Monash University has been involved in the PRAGMA (Pacific Rim Applications and Grid Middleware Assembly) collaboration which includes APAC and Monash University in Australia and University of California, San Diego (UCSD), CICESE, CSE-Online , Pacific Northwest GigaPop, Cray, the National Center for Supercomputing Applications (NCSA), located at the University of Illinois at Urbana-Champaign (UIUC), and StarTap in the United States. Another project is PRIME@Monash where projects range from bio-engineering, chemistry to computer science. And then there is MURPA (Monash University Research Program Abroad) which has been made possible by the massively increased bandwidth that Translight Pacific Wave offers. Seminars were conducted between Monash and UCSD with quality being so high it was "almost as if we were all in the same room".

Cloud Computing has been hyped quite a bit and reading the Gartner cycle it would seem that the hype cycle has "peaked". That said, cloud computing is a major shift in the provision and delivery of computing services. Universities usually rely on government-supported infrastructure, which is rarely controlled commercially, with access controlled by users, and high-end facilities subject to peer review. Switching to the cloud would be a major shift from distributed unmanaged services to scalable centralized services and commercial cloud services.

There are challenges in policy and technical areas surrounding the use of Cloud and Grid resources. Free resources will not disappear, but overflow capability might be provided by commercial providers. There is the potential to perform base-load computation on “free” resources and pay-as-you-go-services to meet user demand. However, to-date there are very few tools that can support both styles of resource provisioning.

If Grid enabled elasticity is compared to cloud enabled elasticity, there might be some advantages to using cloud resources. The Grid had too many people making it too complex[i]. Clouds might suffer the same fate if the resources don’t expand elastically or scale. It is possible that hybrid solutions might come into play – Grid for the wide area, Cloud for the local area[ii].

Nimrod is a widely adopted Grid middleware environment for building and managing large computational experiments over distributed resources. One of the goals of the Nimrod project is to create a user friendly interface to give scientists access to large computational experimentation. Global Grids support a wide range of scientific research from drug docking to antennae design.

Nimrod is built on Kepler (UCSD, UCD developed). Kepler doesn't handle large amounts of parallelism. So the two integrated together creates a flexible IO model. A recent experiment using MATLAB shows workflow using local resources and then getting resources from Amazon when needed to complete flow in set amount of time.

David mentioned a successful ARC Linkage grant with Leica to utilize a remote-controlled microscope in Germany using a workflow submitted by Australia (Opt-Portal). David suggested a strawman project: Grid Enabled Microscopy Across the Pacific (GEMAP). Use a mix of compute clusters from the US and Australia, remote microscopes from Leica and perhaps others, display the results on distributed devices such as OptIPortals. Issues that would need to be resolved is who pays for cloud time, which clouds are used, how is the network reservation handled, again who pays, and who pays for the project funding.

Greg Bell of Lawrence Berkeley National Laboratory presented *Cloud-Based Computation and Collaboration: the Challenges for IT Infrastructure* which outlined the current state of cloud utilization by Federal networks in the United States. The US Federal government is pushing the use of cloud services and Google has been responsive. At Berkeley Lab projects such as the Magellan project, collaborative and instructional services, and science pilots such as BLAST on Amazon Hadoop, climate analysis on MS Azure, Supernova Factory on Amazon EC2, and cloud benchmarking for HPC are underway.

Benchmarking has been done for several science areas with Amazon’s EC2 cloud and in the worst case it shows that EC2 is not ready for the kinds of workloads science can throw its way.

One of the implications for infrastructure providers is that cloud-sourcing raises the criticality of the network for bulk data transfer, for access to external services including email. Not only must the network have sufficient bandwidth, but the network connections need to be resilient and fault tolerant. Other things to consider are direct peering arrangements with major cloud providers, and the monitoring, management, and fault isolation of the networks involved. Other issues include identity management and federation, data management, and the calculation of true cloud costs (including bandwidth, storage, CPU consumption and other infrastructure costs.)

In general, researchers are not exposed to the costs of electricity or cooling, so they do not appreciate the real costs. This can be a space for local innovation – but the keenest challenges are likely non-technical.

***Discussion****:*

What is the impact of cloud computing activities on the network? – Images needed by many of the disciplines thus far are not that large so moving them around have not strained the system or network yet. If you get into time-sensitive or large high definition steams, it could push the network requirements considerably.

The High Energy Physics (HEP) community needs data in a timely manner. They would need to be able to get 600 Megabytes of data per second out of Amazon or a similar Cloud Computing provider in order for it to be viable.

Some takeaways from the Cloud Computing talks were:

- Exploration of no-cost peering structures with cloud providers (the Australian research community might explore a joint procurement if that would make sense)
- Development of a cloud switch that allows movement from one VMware to another
- None of the current commercial cloud providers are providing high-performance clusters

*Possible Demonstration*

Mark Ellisman's lab (UCSD) might be a possible collaborator with Monash University for a grid-enabled microscopy project.

## Session 2: High Energy Physics

Anthony Waugh presented *Computing Challenges for High Energy Physics* giving an overview of what high energy physics is, how they conduct experiments and how much data they produce. Quite a bit of data is being produced and will increase dramatically when the Large Hadron Collider comes online.

Just for the ATLAS project there are 1900 scientific authors, 35 countries and 164 institutions all needing access to data. The traditional model was for each home institute to bring the data to its site for analysis. The new model is to use grid computing to spread the data around the world and send analysis jobs to the data. A very large international computing grid has been set up to enable data storage and access for these projects. The real challenge is data collection and storage, and not so much the processing of the information.

It is not possible to record 1 billion events per second. The trigger system selects only events of physics interest. Then the reconstructed data is processed for physics analysis.

Stored data doesn't need to be accessed that often but some of the data needs regular access by the 2,000 scientists working on the ATLAS project.

Tim Dyce's talk, *Australia-ATLAS: An LHC site in the grid outback* described the infrastructure issues that the ATLAS project in Australia has had to overcome.

It has been a difficult challenge for Australia to bring data home for analysis due to the lack of high capacity trans-Pacific links, so the new model is using GRID technology. University of Melbourne is the only grid site in Australia. The ATLAS computing model breaks the WLCG into clouds and tiers, the cloud and Tier1 is currently Taiwan, but data flows across the Pacific in order to attain the required bandwidth, data also flows to Australia from US and Canadian sites. There is going to be a lot more data movement than originally envisioned; 3.2 Petabytes of data per year is the current estimate. The data comes in waves and the faster researchers can obtain the data, the quicker they can use it (wasting less time).

Australia-ATLAS just completed a round of purchasing and currently has 250 Terabytes of storage and 250 cores. The resources are anticipated to double each year.

Graphs showing successful data transfers (56 Petabytes in two weeks) were shown. Network routing issues and fiber cuts have also created interesting problems making some European sites invisible for periods of time.

The University of Melbourne is waiting for an upgrade to a 10 gigabit connection into the AARNet network. They have set up PerfSONAR for testing, and Nagios for monitoring. Jumbo frames are enabled at all points on their path except the disk servers of the data sources.

Paul Avery was the final speaker in the High Energy Physics area and in his talk "*Petascale Distributed Computing Challenges in High Energy Physics*", he described the collaborative culture, the global distribution and how networking resources are cobbled together to allow flow of information.

From relatively small beginnings, the High Energy Physics community has grown to over 2,000 collaborators with costs now over $1Billion. The collaborators include not only the global physics community, but engineers and vendors working closely together with self-developed collaboration tools.

No one country can build the infrastructure to carry out the aims and goals of the Large Hadron Collider (LHC), so it has led to massive data collections belonging to the international research community. One of the challenges is in data storage, the other is CPU. The current collection is between 0.2 and 1.5 Gigabytes per second but by 2013 this is expected to be in the neighborhood of 100 Petabytes. A large part of the computing and storage capacity is not located at CERN, but distributed globally.

The Compact Muon Solenoid (CMS) project was used as an example of how global cyberinfrastructure will be used to move data from CERN in Switzerland to the 7 Tier One Sites, and then on to the Tier Two sites and so on. The reliability and availability of high capacity bandwidth is essential to the timely flow of data.

***Discussion:***

The HEP community has found lots of networking issues when doing these experiments. The experiments are very sensitive to network tuning. Many scientists have had to learn more about networking than they would wish.

Some takeaways from the HEP talks were:

- The Middleware for HEP is also still a “research experiment”.
- User support will be an issue in the future.
- Can an institution draw from more than one Tier1? The University of Melbourne might draw more traffic from US Tier1 site rather than Taiwan Tier1 site. Available bandwidth was not taken into consideration when the original geographical selections were made.
- Are there likely to be collaborative projects that will involve transfers between Tier2 and Tier3's? This is a bit of an unknown. What users and groups of users might do with the data is known.

*Possible demonstration:*

Performance testing over the trans-Pacific links from University of Melbourne to locations accessible through Pacific Wave could be performed both before and after the upgrade to 10 gigabit.

**Session 3: Astronomy**

Tim Axelrod's presentation "LSST and the Cloud: Astro Collaboration in 2016" presented some of the opportunities that may arise once the LSST project becomes operational.

Once built, the 8.4-meter Large Synoptic Survey Telescope (LSST) will survey the entire visible sky deeply in multiple colors every week with its three-billion pixel digital camera, probing the mysteries of dark matter and dark energy, and opening a movie-like window on objects that change or move rapidly: exploding supernovae, potentially hazardous near-Earth asteroids, and distant Kuiper Belt Objects.

The LSST is wide, deep, and fast making its field of view much greater than the Keck Telescope. It has six different filters. A comparison with the Keck telescope on Mauna Kea shows that its field of view is enormous. Once in operation, the data stream to be collected is about 18 Terabytes per night (1/2 gigabyte per second) and the survey lasts for 10 years. The Data Management System will turn this data stream into products. This one survey will drive all sorts of science programs; programs such as weak lensing, supernovae and transient astrophysics, milky way structure, solar system inventory, and individual science collaborations.

With the exception of Chile and an Institute in France, the institutional members are all in the United States. That is a huge geographical difference from the High Energy Physics research community.

Potential collaborations with the LSST project would be real-time spectroscopic analysis following up detection of a transient event using big telescopes, combining optical and radio such as ASKAP and SKA, pixel-level combination with surveys in other wavelengths, or ad-hoc combination of information with other surveys using virtual observatories.

Projects like GAMA, which will bring together data from the Anglo-Australian Telescope (AAT), the VLT Survey Telescope (VST), the Visible and Infrared Survey Telescope for Astronomy (VISTA), the Australian Square Kilometre Array Pathfinder (ASKAP), and the Hershel and Galaxy Evolution Explorer (GALEX) space telescopes to construct a state-of-the-art database of over 250,000 galaxies in the local Universe, will become more common.

Some obstacles to overcome are:

- the relatively constant science budget
- the relatively constant number of funded scientists
- the ration of data to humans is inexorably increasing
- data flow from experiments is exploding
- automated techniques are only partially successful
- the quality of the data

To overcome such obstacles we will need high-speed networks as a foundation, enabling cloud computing, which in turn enables human computation.

Stephen Tingay's talk, *Going coast to coast at the speed of light: continental scale real-time radio astronomy,* showed the national infrastructure that connects up many of the astronomy resources (ASKAP, SKA, etc.) in Australia.

Radio Astronomers would like to see the SKA project get underway.

The connections need to support real-time flow of data. The network transfer of data for real-time or near real-time processing is a lot cheaper than recording to hard disk at the telescope. There are still challenges in getting bandwidth to the right places and removing the remaining bottlenecks. For ASKAP the fiber connection is from the site to Geraldton and then to Perth. The usefulness of ASKAP is depending on the bandwidth available.

Australian astronomers would love to collaborate with the US but there is a lack of decent bandwidth connectivity to telescopes in the United States. Most VLBI antennae have bad network connections. The aim should be 10 gigabit if possible. It is possible that US researchers might wish to use the data stored for ASKAP which will be stored in Perth and made available using the Virtual Observatory interface.

We need eVLBI to follow up on the transient projects -- there are some objects changing hourly. LSST, Skymapper, Fermi can all provide triggers that will need to be followed-up quickly. Some takeaways:

- Australia is leading the world in trans-continental radio astronomy
    - (10 Gbps) - demonstrating Australia’s path to SKA readiness:
    - FACTORS OF 10 - 1,000 IMPROVEMENT IN FIBRE CAPACITY NEEDED;
    - FACTORS OF 2 -3 IMPROVEMENTS IN FIBRE REACH REQUIRED;

    - NATIONAL IT INFRASTRUCTURE [National Broadband Network].
- Inclusion of ASKAP (and other existing antennas) will improve current array by a factor of four in image quality;
- Collaborate internationally with New Zealand, China, India, Japan, Korea, USA, Europe, South Africa in global radio astronomy projects
- Australia would like to collaborate with USA, but the connectivity to telescopes is the primary issue.

***Discussion:***

The Catalina Sky Survey (CSS) is a project that includes telescopes in the United States and Australia working to find killer asteroids before they hit. It looks for near-earth objects (NEOs) and has discovered 70% of the found objects in the last three years. Six observers use the University of Arizona's 1.5 meter, or 60-inch, reflector telescope at Steward Observatory's Mount Lemmon site and the 0.7 meter, or 28-inch, Schmidt telescope on Mount Bigelow in the Santa Catalina Mountains north of Tucson. Two observers use Australian National University's 0.5 meter, or 20-inch, Uppsala Schmidt telescope at Siding Spring, New South Wales, Australia. Will this type of project need bandwidth across the Pacific?

Last year, the Caltech team wrote a National Science Foundation proposal to expand what's called the **Catalina Real-Time Transient Survey**, or CRTS, into a true, fully open synoptic sky survey. Thanks to the $890,000 NSF grant awarded this month, the CRTS team soon will construct a Web site that will make roughly 10 terabytes of data taken by the Catalina Sky Survey over the past five years –- as well all new CSS data that continues to stream in -- available over the Internet to astronomers worldwide, professional and amateur.

## Session 4: Geosciences

Louis Moresi from Monash University's School of Geosciences led off the presentations for the Geosciences group. His presentation, *Around the world with Underworld - distributed development and collaboration in computational geodynamics* featured AuScope, a national geoscience infrastructure program funded under NCRIS with support for geospatial research, earth imaging, simulation software, core library, data grid.

Computational Infrastructure for Geodynamics (CIG) is a coordinated effort to develop reusable, well-documented and open-source geodynamics software. It is supported by the U.S.'s National Science Foundation. Under this program, 2-D/3-D code called "Gale" has been developed for the long-term tectonics community and is currently installed at the Texas Advanced Computing Center (TACC). You can get a free account by filling out an application on the website. Gale is a joint effort between CIG, the Victorian Partnership for Advanced Computing, and Monash University.

The Australian-funded Underworld code is equivalent to the US-funded Gale code. It is modeling all tectonic deformation, over times of 10's to 100's millions of years and distance as well.

Collaborations across the Pacific take place across many time zones and there are not too many opportunities to communicate in real-time. So collaboration tools need to be able to track what each group is doing by storing many different files and keeping them in sync, organizing the discussion around those files, coordinating and reporting on HPC calculations which may queue and run over several days, sharing the writing of technical documents and allow easy sharing with other collaborators.

What the community needs is very simple collaboration tools. Some examples of simple tools are collaboration wikis (Tiddlywiki + jsMath), Dropbox, email & Chat, Skype & Video Chat, the Phone & SMS. What doesn't work is technology that gets in the way such as Googlewave.

There are stable releases of Underworld on the ARCS Grid and the ability to plug in your own modules/toolboxes. Currently, we haven't figured out best way to do that. Also, how much do you need to know about the guts of the machines in order to make it work?

What the community has discovered is that performance is quite sensitive to underlying architectures and a certain amount of performance tuning is required. Someone with experience usually has to check each machine in advance. Three examples of machines were analyzed.

Most vizualizations are done at run time and shipped back as images so there are no real strains on the network at the moment. But this could change as the projects and collaborations ramp up.

The next presentation in Geosciences was by Dietmar Muller from the School of Geosciences at the University of Sydney entitled "*Building a Virtual Geological Observatory*".

Virtual Observatories have revolutionized astronomy by giving the world access to a vast array of sky observations and tools. But we know more about the surface of Mars than that of Earth. For earth observation we have a multitude of observation methods. Plate tectonics makes the job more difficult since things shift over time.

Global climate change and shrinking resources have heightened our sense of dependence on the earth and focused attention on this area of research. However, we are now facing a flood of data making us data rich, but information poor. The balance is worsening every year due to the increase in remotely sensed data. The data sets are orders of magnitude larger and more complex. Earth scientists need to amalgamate data on millions of years of plate tectonic motions, mountain building, erosion, climate and sea-level change, and connect them to analysis and process modeling tools. The Virtual Observatory is a potential unifier in this area.

The Virtual Observatory is a potential way to reconstruct the earth, share complex data, create sets of tools for visualization and processing models using high-performance computing. Google Earth is an example, but the standards (KML) to share data are poorly conceived.

One example of an observatory prototype is GPlates (www.gplates.org) which includes nodes at University of Sydney, Caltech, and the University of Oslo. GPlates integrates plate tectonic and geodynamic models with geological and geophysical data in a "Plate Tectonic GIS".

A science application such as studying the earth's paleogeography as it relates to sea level change can tell us what drove similar changes in the ancient past. To assimilate the many data sets required to come up with a 4-D equivalent to plate tectonics it is necessary to link plate kinematics with the mantle which includes testing alternative plate kinematic models, modeling plate boundary dynamics, modeling vertical motions due to mantle convection, modeling global and regional sea change, modeling palaeoclimate, computing mantle heat flow and then assimilate the geological data into model. Michael Aivasis' group at Caltech developed the Geoframework that helps create mantle convection models (CitcomS).

Many interesting models come out of this collaboration. For the future, more work and support is needed for **VIRGO (Virtual Geological Observatory)**. There are many challenges ahead.

The third and final presentation in the geosciences area was given by Mark Gahegan from the University of Auckland, New Zealand.

**GEON (**www.geongrid.org), the geosciences network, is an open collaborative project that is developing cyberinfrastructure for integration of 3 and 4 dimensional earth science data. GEON is developing the OpenEarth Framework (OEF) to facilitate the integration a variety of multi-dimensional data in response to natural requests from users.

Some examples of how the community is using the resources is

- Alistair Rees, University of Arizona - people share their fossil collections
- Dogan Seber, SDSC GEON SYNSEIS Integration (SCEC also)
- 3-D Earthquake Modeling using HPC: New Zealand doesn't have the supercomputing capacity to do these types of simulations in depth (TeraShake)
- Rock Taxonomy (ontologically based) - geologic map integraton

Lessons learned during the development of GEON

- engage key domain stakeholders early and often
- equal number of GEO and IT project leaders
- 2-3 face-to-face team meetings per year
- specialist meetings with domain scientists and IT staff
- do strong outreach back into science community not directly involved. React constructively to their angry feedback
- hold annual all hands meeting, leads to an annual conference see http://www.geoinformatics2009.org
- summer school for grad students: IT and GEO (free)
- international outreach (I-GEON)
- embedded sociologists observing their process (listen to them early)
- ongoing funding depends on your perceived value to wider group

***Discussion:***

Some takeaways from these discussions are:

- geologists want collaboration tools that are simple and available offline with the ability to resolve editing conflicts straight away
- don't want to spend time setting up and maintaining servers
- as seen in the high energy physics and astronomy discussions, performance is sensitive to the underlying infrastructure

**Session 5: Marine Systems**

Roger Proctor of IMOS kicked off this session with *The Australian Integrated Marine Observing Systems: Present and Future Possibilities.* IMOS is a five year initiative of the Australian Government (2006-2011) to establish an observing system which can be maintained in the long term to inform on climate change in the ocean.

There are four objectives:

- improve capacity to understand ocean variability
- monitor interaction of the ocean boundary currents
- deliver systematic capability to understand
- provide a marine information system for researchers to easily access legacy and new data

There are four regional nodes plus a bluewater node. The vision is to have regional monitoring by the 11 national facilities providing information for regional node science. Each component has its own science questions. Facilities that are used to gather data are ARGO floats (last 3 years, 3261 in the world), autonomous vehicles (AUV), sensor nets, southern ocean moorings, satellites, coastal radar, fish tagging, ships of opportunity, gliders ( deep water and coastal shallow gliders), and the mooring network. The eleven facilities are listed below:

- **ARGO** (Argo Australia)–measures vertical profiles, physical & O20-2000m
- **SOOP** (Ships of Opportunity Programme)–surface & sub-surface physical & biochemical data along ship-tracks
- **SOTS** (Southern Ocean Automated TimeseriesObservations)–extreme climate, the southern ocean moorings
- **ANFOG** (Australian National Facility for Ocean Gliders) –gliders, ocean and shelf repeat transects, phys & biol
- **AUV** (Autonomous Underwater Vehicle)–underwater vehicle, close up view of benthic biodiversity + phys
- **ANMN** (Australian National Mooring Network)–coastal (NRS) & shelf-edge moorings, surface, sub-surface & acoustics
- **ACORN** (Australian Coastal Ocean Radar Network)–HF radar, surface currents & waves to 75km offshore
- **AATAMS** (Australian Acoustic Tagging and Monitoring System)–acoustic tagging, fish, sharks, sea mammal detection

- **FAIMMS** (Facility for Automated Intelligent Monitoring of Marine Systems)–networked sensors on the GBR (Scott Bainbridge)
- **SRS** (Satellite Remote Sensing)–satellites, SST and ocean colour for Australian region
- **eMII** (eMarineInformation Infrastructure)–data management, integration, bringing it all together

Collection and storage of the data is using open source, standards driven tools. The IMOS Metadata Entry and Search Tool (MEST) is the backend catalog to the IMOS Ocean Portal which gives access to the data either directly for a simple search or directly from MEST for more complex or advanced searches.

DataTurbine, originally a proprietary NASA code, is now a robust open-source streaming data middleware system that fulfills the requirements of an sensor-based observing system. Its development by SDSC/UCSD is sponsored by the US's National Science Foundation.

Much of the equipment used is highly dependent on equipment from the US, and other international collaboration.

Potential future facilities include drifting buoys, sediment deposition, microbial Observatory, bio-acoustics, fast ice etc. Potential new data types include floc imagery, gene sequences, sound spectra, and ice properties. Integration with ocean forecasting a la BLUElink might be a next step.

Possible collaborations with the US-IOOS and the EU MyOcean might be possible.

Tony Haymet (Scripps) was the second speaker.

The Southern California Coastal Ocean Observing System (SCCOOS) brings together coastal observations in the Southern California Bight to provide information necessary to address issues in climate change, ecosystem preservation and management, coastal water quality, maritime operations, coastal hazards and national security. As a science-based decision support system, SCCOOS works interactively with local, state and federal agencies, resource managers, industry, policy makers, educators, scientists and the general public to provide data, models and products that advance our understanding of the current and future state of our coastal and global environment.

People can see data on the Southern California coastal region on this site. It is the longest biological record of climate change due to collection of data along the coast due to the crash in sardine population.

With the Ocean Observatories Initiative (OOI) we are now moving from a data poor environment to a data rich environment. Just learning how to comprehend all the additional information is a challenge.

*OOI 'Regional Scale Node' + bits on 'Coastal' and CI* presented by Ron Johnson highlighted the Regional Scale Node project that was just recently funded by the US for deployment on the Juan de Fuca plate. Neptune Canada is already built on the northern extent of the Juan de Fuca plate, fiber across an entire plate. The coastal observatory in Oregon will be connected to it.

The plan is to send data back in real-time. Researchers are interested in putting robots and sensors out on this deployment as well. One researcher wants to do gene sequencing off a robot near a volcanic vent!

The network requirements were built in to the project from the start of OOI. The primary infrastructure will go in around 2012-2013, and the secondary around 2013-2014, so the network infrastructure will not be ready for a while or at least will not be online for a while. This much power has not been pushed this far for this length of time, so there are still some engineering issues to resolve. Other potential challenges: too many disciplines, need to be coherent and comprehensive environmental data at the global scale.

**Session 6: Green IT/Sustainability**

The first speaker was Rod Tucker - Univ of Melbourne (CUBIN).

Why is energy efficiency and the network important? It is an operational expenditure. It is a major part of telecommunications expenditure, it has a greenhouse impact, and energy limited capacity creates bottlenecks.

The most energy used in the network comes from the core routers and the data center. Energy efficiency in routers is improving at about 20% per year. But Data Center consumption is increasing. It doubled between 2000 and 2006 (*Revolutionizing Data Centre Efficiency—Key Analyses" McKinsey & Company, April 2008.)*

Fiber to the Home (FTTP) is the most efficient way to deliver services in terms of energy, graph with energy per bit, can add them up to get you how much it costs to send data.

Note about cloud computing -- we must consider the transport energy cost to really know what the cost per bit is. Cloud Computing does not solve all problems.

A multi-faceted approach needed to build a green Internet:

- Improved efficiency in electronic and photonic devices
- Low-energy switching techniques
- Improved architectures
- New protocols
- New approach to network design and management

Greg Bell's presentation, *Recovering Stranded Capacity in the Data Center,* discussed the efforts of Lawrence Berkeley National Laboratory to increase the efficiency of its data center operations. Most data centers today are over-cooled and inefficient. This leads not only to additional resource costs, but a waste of scientific capacity. The Berkeley Lab added 100kW load to a legacy data center without needing to add any additional cooling.

One of the first steps is to determine what your current footprint is by collecting data. They put in 700 environmental monitors to determine temperature, pressure, current and humidity. This can lead to tuning of the room by removing perforated floor tiles, shutting off humidity features on the air conditioning units. A second strategy is to try some simple inexpensive solutions to air flow management such as blanking panels or transparent curtains but you need to make sure the solutions work with fire code, fire suppression systems etc. A third strategy is to move to liquid cooled racks.

There are products on the horizon that will tolerate warmer air intake. Also there are new liquid cooling products but how do you choose among them? There is a chill off competition in its second year (sun/lbnl/svlg) and the results should be available March 2010. Water cooling without chillers using waste water from HVAC units used to cool cluster will be available soon.

What has resulted from these studies is recovery of capacity for science and avoidance of additional expense in building out a new facility. They are working on an instrumented chiller-less container at the moment. Their ultimate goal on the instrumentation side is to come up with an efficiency dashboard. They are open to discussions with other groups and their discoveries as well.

If you do need to build a new data center, you will need to include instrumentation for real-time efficiency calculation from the start. It is much easier to have installed at the outset than to retro-fit this later on.

How do we motivate energy efficiency in the data center space? Just saving electricity is not enough. However, adding in the benefits of increasing computing capacity without additional mechanical infrastructure investment thus avoiding a costly data center build-out can be a persuasive argument for retrofitting existing data centers.

***Discussion:***

How does this fit in with political considerations for the National Broadband Network?

Do we think this might push use of multicast? Is multicast more green? It certainly consumes less power.

For video distribution - multicast IPTV that brings this to exchange in a more efficient way. Australia might be able to take the lead on this since they will have a national broadband network capable of doing this.

## Workshop Summary

The workshop attendees found the presentations of very high quality and valued the opportunity to hear about new developments in the US and Australia. Most important was the opportunity for cross-fertilization of ideas and identification of potential collaborations and joint experiments.

There are certainly some areas where researchers in the United States and Australia are already pursuing collaborations utilizing advanced networks, and many routine international collaborations that “take the network for granted” on a day to day basis. All of these activities are greatly enhanced by the increased bandwidth and enhanced performance provided through R&E networks; but it is relatively small set of researchers and projects that currently need high bandwidth connections on a routine basis. The day is coming when everyone will need the resources that currently only these few disciplines enjoy today. Between 2012-2016, huge loads will be coming onto the network in many disciplines and there is some concern among researchers who already rely on good connectivity for their research that there will be additional strains on the network in the future if plans are not made to increase capacity and to make sure the connections are reliable, fault tolerant etc.

However, some researchers are planning on running their experiments and work flows remotely and just bringing back the results. Data storage may turn out to be a larger issue especially when it comes to determining who is going to pay to store the data over the long term. Another issue along these lines is ownership of the data and how open it will be to other researchers. The HEP community has resolved some of this for their upcoming work, but in some areas this has been an obstacle to more collaborative efforts.

Middleware and ease of use was an area that was identified as a challenge for many research groups. There is still work needed to make the networks and systems easier to use. Scientists and researchers would like to be able to use high-capacity networks without having to know a lot about the underlying infrastructure.

---

[i] Goscinski, W. and Abramson, D. “An Infrastructure for the Deployment of e-Science Applications, in “High Performance Computing (HPC) and Grids in Action”, Volume 16 Advances in Parallel Computing, Editor: L. Grandientti, March 2009, approx 540 pp. hardcover, ISBN: 978-1-58603-893-7.

[ii] Abramson, D., Giddy, J. and Kotler, L. “High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid?,” International Parallel and Distributed Processing Symposium (IPDPS), pp. 520-528, Cancun, Mexico, May 2000.