

Response to the DMAS Advisory Committee report
Benoît Pirenne, Associate Director, IT, NEPTUNE Canada
July 2008

The first DMAS Advisory Committee meeting took place in Victoria, on November 22, 2007. The committee was chaired by Bill St-Arnaud, CANARIE, Inc. The final report was received on May 12, 2008.

The following takes most of the findings of the committee and DMAS' answers. The extracts from the original report have been italicized for clarity.

Overall Comments:

The Committee was tasked with a series of questions on the system availability, file management system, DMAS commercialization and communal work environment for science. Overall the committee felt that DMAS has taken appropriate steps in terms of the system availability and reliability with the only additional suggestion that backup systems should be located somewhere else in the country other than Vancouver Island.

DMAS acknowledges the comment. Plans are currently underway to ensure that a copy of the data will be placed in a different part of the country, away from the earthquake-prone zone. See comments below.

The committee felt that direct commercialization of the DMAS raised serious concerns of practicality and the fact that attempts at commercialization may distract the DMAS team from its primary mission. However developing ocean cable research communal science tools that could be used on clouds services like Amazon might produce an indirect revenue stream when such tools were used by third parties where NEPTUNE/ONC could capture a percentage of those revenues. Using cloud services like Amazon might also address questions of scalability, file management and communal science platforms. The committee recommended that DMAS team explore platforms such as MyExperiment.org and Google Social amongst others.

DMAS is not fully pursuing commercialization, but remains open to opportunities that might arise in this area (see below). The question of compute cloud services is being investigated, but there is some resistance to rely on a U.S.-only service, which raises some issues of ultimate data ownership. Should *bona fide* commercial cloud services become available in Canada, they could certainly become contenders for the NEPTUNE DMAS needs. In this respect, it is also worth noting that, together with other partners at UVic and at NRC, we are looking at transforming somewhat the current WestGrid concepts and allowing its members to be seen as compute clouds. Current initiatives such as MyExperiment.org Google Social are being investigated.

Discussion Theme 1: High Availability

The committee was asked to report and give suggestions for future directions of the HA systems. Pierre notes that CISTI has issues for HA as well and suggested it is useful to have a snapshot capability to bring back data quickly.

DMAS is currently leaning towards having a fully redundant remote site, even though a more traditional backup with snapshot capabilities is clearly a more affordable avenue.

The committee suggested that the Data Repository not be located on the West Coast, and was assured that plans were being made to warehouse outside of the possible earthquake zone.

A disaster recovery working group has been created. It is composed of the members of the Systems group in DMAS and reports to the Associate Director, IT. The working group meets on a monthly basis and addresses issues dealing with vulnerabilities and impact assessment, definition of availability requirements, development of mitigation plans and finally testing and implementation considerations. The material produced is being collated on the internal Wiki as a continuous effort. As part of that activity, conversations with UVic have highlighted options for a secondary (backup) data centre at the University of Saskatchewan: this Institution's part of the WestGrid funding is primarily for storage. It will be approached in the coming months.

[ACTION: A status report will be presented at the next DMAS Advisory Committee meeting]

The Sun Systems architecture is sound and the committee suggested to stay with VMWare/Sun. As the volume grows, it is important to beef up the Web Server. Web crawlers may crawl the site and if the site is large this activity might impact performance of the servers, so this should be balanced. One option is to limit the number of connections for web crawlers at the OS level. This was already noticed for the Google crawler. It was noted that such crawlers are not to be prevented from their task.

The 64 bit tomcat implementation across multiple machines could be considered for the Web Server implementation.

DMAS is planning to have load balancing capabilities at the level of the web server to address peak demands on our services. Web crawler configurations will be considered in the set up.

[ACTION: Those features will be implemented by late summer of 2009, released in conjunction with the underwater infrastructure and its facilities]

The big irons and the small commodity machines are two options for providing HA. Virtualization of servers provides the ability to be agnostic about the physical resources. The Sun machines provide the basic power for the DMAS. Most DMAS applications do not need large processing power, although in future video analysis may be done.

The data that is available now is that collected for VENUS. Access is event-driven set up through triggers, and the DBMS uses caching. The feedback that the users want is to know how their job is doing, how large the result will be, how long will it take.

Small commodity machines are the preferred option for DMAS web servers. Video analysis would not necessarily be done on-line on our servers but farmed out to the Grid. The same would be true for other bulk data processing activities requested by our users through the web. Feedback to users on the progress of their requests will be implemented.

The general consensus is to spend money at the shore station. This station must be very reliable. The backhaul is second priority.

This is indeed the case: A number of the servers just purchased will be installed in the shore station. The plan is to have one serving each node and its instruments, plus 2 or 3 redundant enterprise service bus machines to transmit data. With the recent award of \$2.4M through the CANARIE Infrastructure Extension Program, the 10Gb/s backhaul connection between Port Alberni shore station and UVic has been completed.

[ACTION: A full network diagram of the NEPTUNE infrastructure including Shore Stations and data centre(s) will be presented at the next DMAS Advisory Committee meeting]

Discussion Theme Two: A new file management system for DMAS?

Which file system will serve best the growing needs of the ONC for storage, processing and search capabilities?

The Globus file system and Hadoop prototype were discussed. Hadoop is an open source clone of the Google file system.

Amazon offers a hosting service, EC2, which can be leased on an hourly basis and offers options for large storage.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. [<http://labs.google.com/papers/mapreduce.html>]

MapReduce can be run on EC2 using Hadoop.

Unfortunately the WestGrid does not share processors cycles and cannot access other networks, so it does not seem suitable for DMAS.

As much as possible the committee recommends use of external file management systems support by Amazon and other services, especially for access by researchers and users outside of Canada.

Ideally, the file management system will be independent of the underlying compute cloud infrastructure. Users will unlikely be allowed direct access to the file management systems. More reasonably, web pages and web services

will provide access to data products that will involve access to files and subsequent post-processing to take place before delivery. In the palette of possible systems to serve our needs, a newcomer from UCSD is iRODS which we are currently investigating.

[ACTION: The outcome of this file systems investigation will be presented at the upcoming DMAS Advisory Committee meeting]

Discussion Theme Three: DMAS Commercialization

- *What compelling features would make a commercial/paid support model successful?*
- *Do we need to limit ourselves to data management systems for ocean science?*
- *How do we develop documentation, online help and toll-free support lines?*
- *How do we exploit the time-to-market advantage we have over the competition (e.g., OOI)?*
- *Should we pursue all alternatives in parallel?*

Benoit: The palette is very wide for DMAS commercialization. It is an industry-ready sensor-based data-acquisition system. We can generalize its use to other environments: medicine, security, engineering, big science.

The Committee suggests that commercialization be explored very carefully. IBM and HP are all doing this and have the services of sales people with contacts. This effort would pull you away from your core mandate. The ability to sell even to the European and US ocean-based groups is low; they will do their own thing. There are a lot of similar products out there. The committee recommends that UVic determine what the offshore oil and gas exploration companies are using and see if there is an international market place for coastal ocean observation systems?

DMAS acknowledges the comments. DMAS certainly has distribution prospects as part of a "package" that would include sensor networks, be it for underwater use or otherwise. In this respect, DMAS has recently demonstrated a scaled-down version of its data acquisition system for deployment on vessels, opening up a much wider field of applications. but clearly, DMAS will not be able to embark on commercialization efforts on its own, without the appropriate resources. In this respect, and as an example, IBM have approached us and would like to gauge our system's "value" to them. Similarly, other avenues are being pursued by the ONC, NEPTUNE and VENUS management teams more widely and a LOI funding application to the CECR program has recently been made that includes support for commercialization efforts. A number of companies and organizations have already promised tangible support to these initiatives. Some other contacts involve the oil and gas industry.

[ACTION: DMAS, NEPTUNE and the ONC management will report on the commercialization efforts to date at the upcoming committee meeting]

The committee believes there are better ways to leverage your database and exploit your intellectual property. Use Google AdSense ads on your website. Develop tools

that can be used on Amazon. You generate some income and build upon Amazon's powerful toolsets.

The Partnership model is moving down a track where NEPTUNE is the new generation of wiring the oceans.

Whatever you do in commercialization, don't lose focus. You are looking at the big problems that affect the planet. Your justification is (1) public policy (2) Ecological development with spin off benefits for the marine SME's (3) public education and outreach. Can think of each of these as fulfilling a need and therefore a potential market. Perhaps the best fit for DMAS is (3).

DMAS might be better off to make tools available and let the community be supportive and develop the next layers. Use the open source model to bring the tools to the next level. Open source some of the components of DMAS. But make arrangements with Amazon and other virtualization and cloud suppliers to earn revenue when compute cycles are used in support of these open source applications

It is not yet clear that we would want to introduce advertisements on what is a public web site financed with public money; this issue needs further policy analysis. However, the development of tools to be made available for a small fee on Amazon or on smart phones (e.g., the iPhone application business model proposed by Apple) is appealing as a way to propose services for a modest fee to a growing arena of portable devices, hungry for localized information.

Another potential is to form partnerships with large companies, such as Alcatel-Lucent when they lay cable. Another example is partnership with instrument manufacturer.

These are the sort of arrangements that have already been considered and have in two cases involved royalty payments from future product sales. Again, it is at the level of ONC and the NEPTUNE Canada management where the initiatives are taking shape as in the recent CECR LOI application.

Discussion Theme Four: A communal work environments for science

The project is relatively ambitious and we would like to hear from the Advisory Committee on the following points:

- Which technologies are available today to implement a social networking environment in which scientific processing is to take place?*
- Which data processing middleware should be considered that provides transparent access to Grid processing environments?*
- Which workflow tools should be leveraged that supports well the execution of procedural tasks expected from scientists?*
- Which data processing environment should be made available as a default system for users to run calculations and scripts? Matlab? R? ...*
- Advice on adopting/building upon community standards such as those provided in Canada by DFO?*

This field is changing rapidly with the growth of popularity of social web sites like MySpace and FaceBook. Web 2.0 MashUps were considered. The suggestion was made to look at iGoogle and Google OpenSocial and MyExperiment.org:

- *iGoogle supports the use of specially developed "gadgets" to display content on a user's page. The gadgets interact with the user and utilize the Google Gadgets API. The Google Gadgets API is public and allows anyone to develop a gadget for any need. [<http://en.wikipedia.org/wiki/IGoogle>]*
- *OpenSocial provides a common set of APIs for social applications across multiple websites. With standard JavaScript and HTML, developers can create apps that access a social network's friends and update feeds. [<http://code.google.com/apis/opensocial/>]*

DMAS with its rich data source and interested community may be able to offer the scientific community the means to help each other to develop, share and enhance common apps and gadgets. In particular graduate and undergraduate students are interested in developing these kinds of applications. The Google Summer of Code program was used by CISTI last summer. [<http://code.google.com/soc/2007/>]

Google is indeed one of the current major players in novel software approaches. A DMAS delegation attended the recent Google I/O developer's conference. It was very instructive in that it provided an indication of the trends and directions of the industry for the coming few years. The trend is indeed toward the production of open-source toolkits that can then be used to create applications more efficiently or "democratize" the preparation of attractive applications. Our view of "Oceans 2.0" is clearly one where this approach will be facilitated, in particular in conjunction with the Service Oriented Architecture already in place that makes the access to functions and services so much easier and affordable for students and other interested parties through Web Services.

[ACTION: DMAS proposes to show some of the possibilities available through the new Web 2.0 technologies that we intend to use at the next Advisory Committee meeting. Advice will be sought from the Committee on the value of those in the context of the development of a participatory environment aiming at the engagement of students]

If this can be an activity that is perceived to be led by scientists, it will have more traction. Tony Bailetti of Carleton University is leading an effort for open source development of communication-enabled applications. One important difference between that work and DMAS is that DMAS has access to content, which makes it more immediately useful.

It is certainly true that DMAS has access to rich, multimedia content and is being driven by scientists. We take good note of the suggestion. As a matter of fact, Ocean Networks Canada, parent organization for NEPTUNE and VENUS, had provided a letter of support for Tony's proposed initiative with the Centre of Excellence for Commercialization and Research (CECR) last year. It was however not successful.

Other topics were mentioned:

- *Java Portlet: JSR 168: The specification defines a common Portlet API and infrastructure that provides facilities for personalization, presentation, and*

security. Portlets using this API and adhering to the specification will be product agnostic, and may be deployed to any portal product that conforms to the specification.

[http://developers.sun.com/portalserver/reference/techart/jsr168/pb_whitepaper.pdf]

- My Yahoo is a customizable web page with news, stock quotes, weather, and many other features.
- The Dojo toolkit is a modular open source JavaScript toolkit (or library), designed to ease the rapid development of JavaScript- or Ajax-based applications and web sites.

A primary task is visualization. Thus a primary tool for the collaborative environment is the use of visualization tool kits, including

- MatLab.
- ManyEyes [<http://services.alphaworks.ibm.com/manyeyes/>] and
- LabView [<http://en.wikipedia.org/wiki/LabVIEW>].

CISTI is also working on a similar sort of experimental application.

DMAS takes notes of all the references and will check them out.

The DMAS team would again like to thank the Advisory Committee and in particular its chair Bill St-Arnaud for their work. We are looking forward to another fruitful exchange at the occasion of the next meeting to take place in the early fall of 2008.