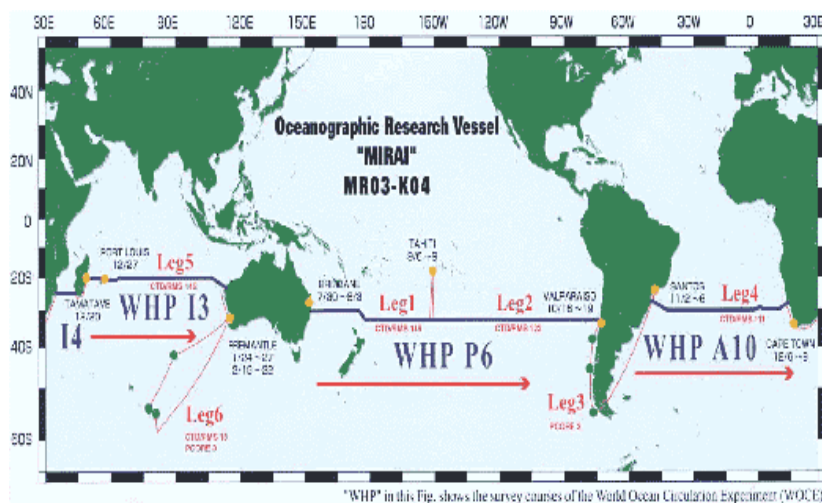


Duplicate Checking of Large Datasets at CSIRO Marine Research



Collaborators: Robert Bell, CSIRO, Ann Thresher, CSIRO

Objective

The project aims to research possible improvement in efficiency of an existing suite of duplicate checking programs that are used in CSIRO Marine Research. It also aims to introduce a more extensible software architecture to facilitate future extension. This project was conducted as a collaboration between CSIRO High Performance Computing Centre and VPAC.

Background

The Ocean Observing Networks at CSIRO Marine Research manipulates large datasets obtained from various sources. The datasets are important in various applications such as global warming study. However, the datasets contains overlapping data.

In order for scientists to fully utilize the datasets, the duplicates inside the datasets must be eliminated. However, the existing programs that perform such a task are inefficient. To improve the workflow of the end user and to provide useful data to the user in a timely manner, it is imperative to improve the efficiency of current programs.

Outcomes

<i>Number of Processors</i>	<i>Execution Time</i>
2	338.200068950653s
3	205.849731922150s
4	164.152899026871s

Table 1: duplicateNCaddons

A Redesign resulted was undertaken to decrease the cost of I/O operations. There was a three-fold speedup of the program after the program was optimized. The duplicateNCaddons algorithm was parallelized and the results above demonstrates that the new program exhibit good scalability. Further improvements to the parallel database organisation are planned.