

On the Web at [http://dust.ess.uci.edu/prp/prp\\_sei/prp\\_sei.pdf](http://dust.ess.uci.edu/prp/prp_sei/prp_sei.pdf)

NSF Science and Engineering Informatics (SEI) Proposal

Submitted: March 4, 2004

Last modified: Saturday 10<sup>th</sup> December, 2005, 20:56

Next Round Due: December 15, 2005

## **SEI(GEO): Scientific Data Operators Optimized for Distributed Interactive and Batch Analysis of Tera-Scale Geophysical Data**

Dr. Charles S. Zender

Department of Earth System Science  
University of California at Irvine

Dr. Phil Papadopoulos

San Diego Supercomputer Center  
University of California at San Diego

**News/Preface:** NSF funded the first NCO/SDO proposal. The next NCO/SDO proposal will tackle different issues, particularly native HDF support and possibly bioinformatics and multi-core processor extensions. This preface sketches out the second NCO proposal, and is followed by the fifteen-page body of the first proposal. Once we are ready to branch off and work on the second NCO proposal, probably around November, 2004, I will place a link to it here. So, just to be clear, this introductory information is an evolving document as we focus in on the second proposal. The next fifteen pages is the body of the first proposal and is fixed. Things got into this hybrid state because I expected to revise and re-submit the first proposal. I began the revisions before learning that the proposal was funded.

The second proposal will leverage the infrastructure and research provided by the first grant, and will develop whole new applications and/or enhancements for NCO/SDO. HDF, bioinformatics, and multi-core processor support are possible directions. The eventual focus needs a champion to Co-PI the next proposal. If you are interested in being Co-PI on the next proposal in one of these areas, or in being PI of the whole proposal, let me know. (I will not PI another proposal to NSF SEIII until the current grant is closer to expiring (say, 2006), but am happy to play a role in others' SEIII proposals.)

**Information for potential collaborators:** This is an NSF proposal to improve Distributed Data Reduction & Analysis (DDR&A) with NCO. The two main components of this proposal are NCO parallelism (OpenMP, MPI) and Server-Side DDR&A (SSDR&A) implemented through extensions to DODS/OPeNDAP and to netCDF4. This project would dramatically reduce bandwidth usage for NCO DDR&A. With this first NCO proposal out the door, the content of the next two NCO proposals is clear.

The second NCO proposal will be re-written and improved from the first. The certain changes include (1) a more convincing description of why CCSM IPCC data reduction exemplifies a computational, storage, and network-constrained problem common to many domains besides geophysics, (2) articulating the numerous required SSDR&A extensions to OPeNDAP and to netCDF4, and (3) deeper involvement (and funding) for distributed computing experts (e.g., UCSD/OptIPuter) on the parallelization and SSDR&A issues. Unfortunately, the NSF SEIII program has relatively limited funding. The average SEI three-year award size is \$500,000. Our budget is already ~ \$600,000 and I want to grow the SSDR&A component. If we are not funded this round, we will include these changes and re-submit to the next NSF SEIII round, due 20041215.

I anticipate sending a distinct (third?) NCO proposal to NASA. Its narrow technical focus will be NCO/SDO for native HDF speakers. I will ask Tom DeFanti and Mike Folk to clarify the issue

of whether HDF needs to be modified to be OptIPuter-aware. The two likely NASA programs which might support SDO are ESTO and REASON CANN.

Other significant new directions for NCO/SDO might merit full new proposals to NSF or to other agencies (e.g., DOE, NASA, NIH). Significant extensions under consideration include a bioinformatics component (led by Harry Mangalam, TACGI?), arithmetic speed-ups on multi-processor single-core nodes (led by Steve Jenks, UCI?), or really going hog-wild with the SDR&A idea (anyone?).

Please contact me if you wish to be involved with any future proposals. Comments on the proposals and letters of support are very welcome.

#### 1. Senior Personnel Inquiries for Intellectual Collaboration

- (a) Peter Cornillon (URI) OPeNDAP netCDF CL extensions (declined until next round, provided LOS)
- (b) Steve Jenks (UCI) multi-processor core performance improvements (no response)
- (c) Harry Mangalam (TACGI) bio-informatics (declined until next round, provided text)
- (d) Phil Papadopoulos (SDSC, Co-PI) parallelization?
- (e) Russ Rew (UCAR Unidata) netCDF4, API (provided LOS)
- (f) Takemura Sato (Japan ESC) DDRA demonstrations (will serve 1 TB)
- (g) Larry Smarr (Cal-IT2) OptIPuter (provided LOS)
- (h) Padhraic Smyth (UCI ICS) data mining (declined)

#### 2. Senior Personnel for Equipment Collaboration

- (a) Aaron Chin (Cal-IT2, Senior Personnel) OptIPuter server support

#### 3. Other Letters of Collaboration/Support to consider

- (a) Rajiv Bendale contact (IBM) bio-informatics
- (b) John Caron (UCAR Unidata) NcML aggregation, lazy processing
- (c) Brian Eaton (NCAR) CF conformance
- (d) Ian Foster (Argonne)
- (e) James Gallagher (URI) OPeNDAP netCDF CL extensions
- (f) Dan Holloway (URI) OPeNDAP netCDF CL extensions
- (g) Mark Taylor (Sandia) climate SDR&A

#### 4. Big Budget Items

- (a) One month summer salary per year for Zender
- (b) One month salary for Aaron Chin (OptIPuter)
- (c) P/A or Specialist II to architect/bulletproof/coordinate/release SDO (Butowsky?)
- (d) Grad Student to optimize computational geometry based on processor topology (ICS?)
- (e) \$17k for ESMF↔OptIPuter networking

#### 5. Next SEI proposal due December 15, 2004

## **Project Summary.**

Climate simulations for the Fourth Assessment Report of the IPCC will be performed and stored in netCDF format in multiple national and international HPC centers. HDF datasets from NASA, NOAA, and ESA satellites steadily accumulate in geographically disparate EOSDIS sites. These data are only as valuable as they are accessible to the working geophysicist at his or her desk. Concurrent interactive and scripted analysis of geographically distributed large scale scientific datasets is functionality critical to managing and interpreting the many streams of geophysical data.

We propose to develop a suite of Scientific Data Operators (SDO) for interactive and scripted manipulation of (locally and) widely distributed repositories of netCDF- and HDF5-formatted geophysical data. SDO's functionality will suit distributed, network-transparent, analysis of ensemble tera-scale datasets stored at many remote locations. SDO will meet the current and future needs of geophysicists and, potentially, bio-informaticists.

SDO integrates four existing, proven, Open Source software technologies: (1) netCDF—the de facto standard format for climate model data and Earth-bound geophysical observational datasets. (2) netCDF Operators (NCO)—the de facto standard for analysis of climate model and geophysical data. (3) DODS/OPeNDAP—the de facto standard for network-transparent client/server access to geophysical data. (4) HDF-EOS—the official storage standard for NASA EOS satellite data. We have taken advantage of existing synergies and convergence among these standards, and guidance from their initiators, to plan the extension of the existing NCO toolkit into the next generation SDO toolkit. SDO will (like NCO) be an indispensable software assistant to individual researchers and research centers for distributed processing of climate and satellite data.

This SDO project has three main parts: (1) Implement shared memory (OpenMP) and Message Passing Interface (MPI) parallelism to speed up and reduce latency of local and distributed Tera-scale data processing. (2) Design and implement automatic computational geometry load-balancing algorithms to exploit the intrinsic parallelism of frequently used arithmetic operations without user intervention. (3) Add, enhance, and simplify functionality for server-side processing of distributed netCDF data with the OPeNDAP protocol.

**Scientific Merit:** The proof-of-concept application, distributed analysis of NCAR CCSM IPCC assessment simulations within and across national boundaries, may potentially yield otherwise undiscovered patterns among different SRES scenarios for the same model, and among different climate models. New data reduction efficiencies will be gained by automatic configuration of computational geometry to optimize for the data servers' processor topology. The tera-scale climate datasets targeted for analysis will reveal any critical bandwidth, I/O, and client/server bottlenecks in distributed geophysical data processing.

**Broader Impacts:** Bringing distributed data analysis power out from the realm of the computational scientist to the desktop of the practicing geophysicist will leverage existing distributed capabilities by making their use transparent to the average scientific user. Improving tools to analyze and inter-compare geophysical model and observational data that reside in various national HPC centers increases the simulations' scientific value and decreases time to discovery and publication. SDO helps address the problem of growing bioinformatics data sets, especially gene expression data, in ways similar to the geophysics domain. It provides a mechanism for efficient storage and manipulation of the numeric data separate from the contextual or meta-data which is normally stored in XML.

# SEI(GEO): Scientific Data Operators Optimized for Distributed Interactive and Batch Analysis of Tera-Scale Geophysical Data

## 1 Introduction

Climate simulations for the Fourth Assessment Report of the IPCC will be performed and stored in netCDF format in multiple national and international HPC centers. HDF datasets from NASA, NOAA, and ESA satellites steadily accumulate in geographically disparate EOSDIS sites. These data are only as valuable as they are accessible to the working geophysicist at his or her desk. Concurrent interactive and scripted analysis of geographically distributed large scale scientific datasets is functionality critical to managing and interpreting the many streams of geophysical data.

We propose to develop a suite of Scientific Data Operators (SDO) for interactive and scripted manipulation of (locally and) widely distributed repositories of netCDF- and HDF5-formatted geophysical data. SDO's functionality will suit distributed, network-transparent, analysis of ensemble tera-scale datasets stored at many remote locations. SDO will meet the current and future needs of geophysicists and, potentially, bio-informaticists.

SDO integrates four existing, proven, Open Source software technologies: (1) netCDF—the de facto standard format for climate model data and Earth-bound geophysical observational datasets. (2) netCDF Operators (NCO)—the de facto standard for analysis of climate model and geophysical data. (3) DODS/OPeNDAP—the de facto standard for network-transparent client/server access to geophysical data. (4) HDF-EOS—the official storage standard for NASA EOS satellite data. We have taken advantage of existing synergies and convergence among these standards, and guidance from their initiators, to plan the extension of the existing NCO toolkit into the next generation SDO toolkit. SDO will (like NCO) be an indispensable software assistant to individual researchers and research centers for distributed processing of climate and satellite data.

This SDO project has three main parts: (1) Implement shared memory (OpenMP) and Message Passing Interface (MPI) parallelism to speed up and reduce latency of local and distributed Tera-scale data processing. (2) Design and implement automatic computational geometry load-balancing algorithms to exploit the intrinsic parallelism of frequently used arithmetic operations without user intervention. (3) Add, enhance, and simplify functionality for server-side processing of distributed netCDF data with the OPeNDAP protocol.

The proof-of-concept application, distributed analysis of NCAR CCSM IPCC assessment simulations within and across national boundaries, may potentially yield otherwise undiscovered patterns among different SRES scenarios for the same model, and among different climate models. New data reduction efficiencies will be gained by automatic configuration of computational geometry to optimize for the data servers' processor topology. The tera-scale climate datasets targeted for analysis will reveal any critical bandwidth, I/O, and client/server bottlenecks in distributed geophysical data processing.

### 1.1 Organization

This proposal is organized as follows. Section 2 describes the purpose, capabilities, and functional relationships between netCDF, HDF, NCO, and OPeNDAP. Section 3 describes the results

of our relevant, prior NSF-funded research. Section 4 describes the current barriers facing geoscientists who require Distributed Data Reduction & Analysis (DDR&A) capabilities. Section 5 details the specific objectives of the proposal and details our methods for achieving them. We then describe then prototype experiment for this project, DDR&A of climate simulations datasets. Section 6 describes our software engineering plans. Section 7 presents our project coordination plan, PI responsibilities, time-line, milestones, and software engineering methods. Section 8 concludes with a presentation of the broader impacts and synergies of our project. Three letters of support/collaboration and a list of acronyms and abbreviations are included as supplementary documents.

## 2 Background

The increasing size, number, and complexity of scientific data in the past decades has led to the development and use of self-describing data formats (SDDFs) and tools to manipulate these formats. The SDDFs replace less functional formats such as raw-binary or text-formatted data.

### 2.1 HDF

Two SDDFs currently dominate data archival in the geo-sciences. The first is the [Hierarchical Data Format](#) (HDF) (NCSA, 2004). HDF was developed at the National Center for Supercomputing Applications (NCSA) and adopted by NASA for Earth Science Enterprise (ESE) applications. HDF is the most commonly used archival format for ESE satellite data. This proposal does not involve any work directly with HDF and we mention HDF mainly due to its importance in observational geophysics. Although we would like one day to implement a native HDF-EOS back-end to NCO/SDO, that task would be extremely difficult and is beyond the scope of this proposal.

This proposal will, however, exploit and benefit HDF data indirectly thanks to a complementary proposal described in Section 8. Briefly, a fully-funded effort to layer the netCDF API on top of HDF5 is underway. Since NCO/SDO is completely netCDF-conformant, all NCO/SDO operations will soon work on any HDF file written with netCDF4.

### 2.2 netCDF

The second popular SDDF is the [Network Common Data Format](#) (netCDF), developed by Unidata at the National Center for Atmospheric Research (NCAR) (Unidata, 2004). netCDF has become the most commonly used archival format for large scale geophysical models, such as climate models. netCDF is less-powerful than HDF because it lacks features such as data compression, irregular grids, threading, and parallel I/O. However, netCDF is much simpler to program than HDF, and, as a result is widely used in the geophysical and climate modeling community by practicing scientists.

### 2.3 NCO: netCDF Operators

Tools to manipulate and view netCDF files are relatively easy to write since the API is much simpler than, say, HDF. The netCDF Operators (NCO) (Zender, 2004) is probably the best-known

toolkit for numeric and metadata analysis and manipulation of netCDF data.

Traditional processing of scientific data works with an intra-file paradigm. Users open a file, read a variable from the file, and manipulate it. The intra-file paradigm works well in cases where all the pertinent data are stored in one or a few files. In some disciplines, however, data storage requirements dictate that relevant data be spread over multiple files. Satellite-derived information, for example, may be stored in a file-per-day or file-per-orbit format. Data produced by geophysical time-stepping models is output every timestep or averaged over many timesteps. Climate models, for example, archive data once-per simulated day or month, and simulate years or even centuries producing hundreds or thousands of large files in a single simulation. In such applications, the inter-file paradigm becomes unwieldy and the optimal tool for data reduction must support an inter-file paradigm.

We developed some guidelines based on our extensive experience with geophysical and climate data and implemented them in NCO. NCO assumes that processing large numbers of geophysical data-files is most efficient and intuitive when:

1. Files are the fundamental unit of data. NCO makes it easy to add, subtract, and manipulate entire files.
2. Files to be processed in a single step are homogeneous. NCO assumes the structure of each file (i.e., the fields present and their dimensions) are identical to the structure of the first file in the sequence. The two exceptions NCO allows are that the record variable (i.e., time dimension) length, and, in some cases the number of variables present, may change from file to file.
3. Distinctions between *dimensions*, *coordinates*, and *variables* are maintained.
4. Operators have defaults that make sense and may be over-ridden with a simple, mnemonic command line switch.
5. Operators must provide an **audit trail that tracks data provenance**
6. Operators must be as generic as possible, imposing no limitations on data dimensionality, size, or type.
7. Conformance to metadata conventions is paramount

Apparently NCO's guiding philosophy, "do what a sane user would want" has succeeded! NCO runs on all modern operating systems, and its use is fully detailed in the [NCO User's Guide](#). To my knowledge, all established national and international climate modeling centers, including NASA, NOAA, NSF, and DOE centers install and maintain NCO for their system users. See, for example, NCO usage at [DOE ARM](#), [DKRZ](#), [LMD](#), [JISAO](#), [NCAR](#), [NOAA GFDL](#), [NOAA CDC](#), and [PRISM](#). In other words, NCO is widely used as middle-ware at geophysical institutions for data post-processing, hyper-slabbing and serving. The improvements and extensions to SDO proposed here will help weld HPC repositories into a **shared-cyberinfrastructure** that will benefit a substantial scientific community much larger than the proposers'.

## 2.4 DODS/OPeNDAP

The [Distributed Oceanographic Data Server](#) (DODS) data server protocol provides useful replacements for common data interface libraries like netCDF. The DODS versions of these libraries implement network transparent access to data via a client-server data access protocol that uses the HTTP protocol for communication. Although DODS-technology originated with oceanography

data, it applies to virtually all scientific data. In recognition of this, the data access protocol underlying DODS (which is what NCO/SDO cares about) has been renamed the [Open-source Project for a Network Data Access Protocol](#) (OPeNDAP). For the purposes of this proposal, DODS and OPeNDAP are used interchangeably, usually in the hyphenated form. Essentially DODS is being deprecated in favor of OPeNDAP, another acronym for the same technology. The [NCO User's Guide](#) and this [OPeNDAP Presentation](#) provide more details.

Any binary netCDF application (like NCO) may be OPeNDAP-enabled by linking to the OPeNDAP netCDF Client Library (CL) instead of the default netCDF library. Once NCO is OPeNDAP-enabled the operators are OPeNDAP clients. All OPeNDAP clients have network transparent access to any files controlled by a OPeNDAP server.

### 3 Results from Prior NSF Funding on Related Projects

Zender is PI on ATM-0321380 "Acquisition of an [Earth System Modeling Facility](#) (ESMF) for Coupled Climate, Chemistry, and Biogeochemistry Studies". After negotiating the best price supercomputer through an open bid competition in summer 2003, we awarded IBM the ESMF contract in October 2003. The ESMF opened to early users in early February 2004 with a two day HPC-programming workshop attended by about 30 ESMF users. The ESMF is currently undergoing final acceptance testing by UCI and configuration by IBM prior to being fully devoted to coupled climate studies. The ESMF provides the computational power for Zender's graduate seminar [ESS 286B: Modeling Land Surface Processes](#). This SEI will use the ESMF as one source of tera-scale climate model data. This proposal will fund turning the ESMF into an OptIPuter node so that distributed data reduction of ESMF and SDSC data will test SDO performance between two geographically disparate nodes connected by the relatively high-bandwidth OptIPuter network.

Zender is a Co-PI on ATM-0214430, "Collaborative Proposal: Using Measurements from the Columbia Plateau Eolian System to Improve Global-Scale Models of Mineral-Dust Aerosols", 8/1/2002–7/30/2005. This project has resulted so far in four national meeting presentations with manuscripts in preparation ([Sweeney et al., 2002, 2003b,a](#); [Zender et al., 2003](#)). Our manuscript studies the range of uncertainty in LGM dust mass and radiative budgets to uncertainty in vegetation reconstruction. We show that a significant fraction of the observed LGM increase in Pacific Ocean dust deposition is attributable to vegetation change. Our paper in press ([Grini and Zender, 2004](#)) explains how the twin processes of saltation and sandblasting (SS) relate to loess formation. These SS physics were implemented in DEAD which is used in the NCAR CCSM (and other) IPCC simulations, the proof-of-concept application for this proposal. Since CCSM generates netCDF datasets, all analysis of CCSM data is relevant to this proposal.

Papadopoulos is Co-PI for ANI-0225642 "The OptIPuter", formulated to discover the impacts of ultra-high speed networks enabled by optically parallel wave-division multiplexing on system architecture, software architecture, and overall functionality. He is the chief OptIPuter systems and network architect for UCSD, responsible for the design and implementation of the UCSD experimental apparatus. The UCSD OptIPuter is designed around a high-speed packet switching network with a next-generation optical-core Chiaro router as its centerpiece. Six campus laboratories with clustered endpoints connect to the Chiaro through a private fiber plant in which each site has at least four parallel fibers connecting to the Chiaro router. UCI resources will signal through the Chiaro, allowing access to all OptIPuter resources including a 48-node, 21 TB storage test-bed.

Several invited talks and two peer-reviewed papers have been directly attributed to this project.

Papadopoulos serves as Senior Personnel for ACI-9619020, “NPACI: The National Partnership for advanced Computational Infrastructure”. He leads the design and implementation of the NPACI Rocks clustering toolkit used to build hundreds of clusters around the world and impacting several large-scale NSF programs. Rocks is a turnkey solution for rapidly building clustered endpoints. Papadopoulos has authored over one half dozen conference and journal papers and given more than twenty invited talks on Rocks.

Papadopoulos is a Co-PI for ANIR-0123973 “Designing and Building a National Middleware Infrastructure: Towards a National GRIDS Center” and serves as the site PI for SDSC. GRIDS produces integrated grid software releases. At SDSC, Papadopoulos oversees the development and architecture of a general purpose grid configuration tool.

## 4 Geophysical Domain Challenge for DDR&A

Although SDO, netCDF, and OPeNDAP apply to any gridded data, we will use the example of climate model data storage and reduction to concretely illustrate our project. We choose the field of climate modeling for two reasons. First, it involves tangible quantities (e.g., air temperature) and dimensions (latitude, longitude, height, time) which are familiar to all geophysicists. Second, our prototype application is data reduction and analysis of [Community Climate System Model \(CCSM\)](#) ([Blackmon et al., 2001](#); [NCAR, 2004](#)) climate simulations prepared for the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) ([IPCC, 2001](#)).

The IPCC AR4 is scheduled for release in 2006 or 2007. Simulations contributing to this report are underway now. Climate simulations from one model, CCSM, with differing initial conditions (ICs) and forcing scenarios will be performed and archived at geographically disparate High Performance Computing Centers (HPCCs), as described [here](#). These HPCCs include the National Center for Atmospheric Research (NCAR), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), the National Energy Research Scientific Computing Center (NERSC), and other national and international centers including the Earth Simulator Center (ESC) in Japan. Each of these centers will perform an ensemble of CCSM simulations. Typically, the different members of the ensemble correspond to Initial Condition (IC) perturbations needed to help quantify the internal variability of the model. The ensembles themselves span the spectrum of 21st century (and beyond) anthropogenic forcing scenarios agreed to in the IPCC Special Report on Emissions Scenarios (SRES) ([Houghton et al., 1995](#); [IPCC, 2001](#)). Characterizing climate change based on tens–hundreds of TB of CCSM results stored in [HPCCs](#) around the Globe is the **significant domain challenge** for geophysical modelers such as PI Zender.

The CCSM results of a single SRES ensemble occupy on the order of one TB of storage. Thus each HPCC mentioned above has tera-scale Local DR&A (LDR&A) needs for a single, local, CCSM ensemble. Much interesting science will be done on these results at each HPCC without any Distributed DR&A (DDR&A) component. Characterizing the internal variability, mean climate state, and transient evolution of both is a challenging DR&A problem potentially leading to new understanding (i.e., scientific discoveries) of the processes causing the observed changes in the frequency of occurrence and intensity of El Niños, and of abrupt climate transitions. LDR&A of many TB of climate data requires enormous computational and storage resources.

The barriers that face researchers who wish to perform LDR&A on climate data are (first) ag-

gregating the climate data from remote servers to the local host (which may have relatively limited storage), and (second) reducing the wall-clock time of the data reduction. Fortunately, NCO is available and is already **widely used** for LDR&A. Thus the first barrier can be lowered by promulgating use of OPeNDAP servers at HPCCs so that all researchers have access to the climate simulation data. This occurs organically as users learn more about NCO and OPeNDAP and request their system administrators to install them. Serving climate data via OPeNDAP does not address the network problem of bandwidth consumption by hundreds of researchers requesting the same raw datasets be transferred to their local machine. Thus high-bandwidth consumption, high-latency (assuming raw data are originally on remote machines), and un-optimized data-reducing clients (i.e., NCO) together form a **significant computer science problem that is a barrier to achieving the domain challenge** (i.e., characterizing climate by LDR&A methods).

Inter-comparing and aggregating CCSM results from geographically disparate HPCC centers requires Distributed Data Reduction and Analysis (DDR&A) or copying many TB of data en masse to one master storage location for LDR&A. The latter is a viable option for some important types of analysis (e.g., statistics of monthly mean results). However, the disadvantages of copying the distributed raw data to one center for LDR&A are numerous:

1. Raw model output data at high time resolution (e.g., daily average results for one century) for all the ensemble members would require hundreds of TB at a master HPCC storage location.
2. Copying distributed data to local storage is LDR&A in disguise. It offers none of the advantages of true DDR&A (where data are stored remotely). LDR&A reduces the pace of discovery by excluding researchers at other locations from performing entrepreneurial DR&A.
3. University researchers not affiliated with an HPCC have a very difficult time accessing the raw data directly in almost any scenario.

The scientific objectives of IPCC Working Group One (WG1 is in charge of describing the scientific basis of climate change) include quantifying climate sensitivity to various SRES scenarios. Establishing the sensitivity of internal and forced CCSM climate response to the spectrum of SRES forcings requires DDR&A because the results are stored at multiple HPCCs. Thus the prototype experiment for this project is DDR&A of CCSM climate data stored at HPCCs with network connectivity to UCI ranging from low ( $\sim 100 \text{ Mb s}^{-1}$  to NCAR) to high ( $\sim 10 \text{ Gb s}^{-1}$  to SDSC). CCSM DDR&A is an excellent proof-of-concept for SDO because the CCSM IPCC data are (first) in netCDF format, (second) of interest to hundreds of geophysicists (especially PI Zender!), and (third) helping to drive development of new HPC resources and tools.

## 5 Methods

In addition to the domain-specific scientific gains arising from characterizing CCSM IPCC simulation output, we have four overarching software and hardware engineering goals:

1. Increase speed of common rank-reduction operations
2. Reduce latency of distributed data analysis
3. Reduce bandwidth requirements of distributed data analysis

4. Do this all in an architecture-neutral, model-neutral, flexible and extensible software framework that works across many orders of bandwidth of network connectivity

The following methods are designed to attain our domain challenge by resolving the significant computer science barriers in our way.

## 5.1 Rank Reduction Efficiencies for DR&A

We will denote the rank  $R$  of a variable  $v$  with a left-superscript, e.g.,  ${}^Rv$ . Sample variables for our prototype experiment with climate model data are surface temperature  $T_s$ , surface pressure  $p$ , and top-of-atmosphere down-welling shortwave radiative flux  $F_{\text{SW}}^-$ . A time-series of, say, surface temperature  $T_s$  is a rank three ( $R = 3$ ) variable because it has two spatial dimensions (latitude and longitude) and one temporal dimension (time). The instantaneous value of air temperature  $T$  is also rank three since it has three spatial dimensions (latitude, longitude, and level). A timeseries of temperature data is rank four  $T(t) = {}^4T$ .

Rank reduction is often the first step in geophysical DR&A. For example, a researcher may want to derive the time-mean temperature field  ${}^3T = \bar{T}$  (vertical overbar denotes temporal averaging) from the model-predicted  ${}^4T = T$ . This requires a rank reduction from  ${}^4T \rightarrow {}^3T$ . The time-mean, global-mean surface temperature  ${}^0T_s = \langle \bar{T}_s \rangle$  (angles denote area-averaging) is the scalar obtained by performing a reduction of  ${}^3T_s \rightarrow {}^0T_s$ . More generally, rank reduction is  ${}^{R_i}v \rightarrow {}^{R_f}v$  where  $R_i$  is the initial variable rank,  $R_f$  is the final variable rank, and, for geophysics,  $[R_i, R_f] \in [0, 1, \dots, 5]$ . Usually  $R_i > R_f$  but temporary rank expansion (broadcasting) is often required to perform arithmetic efficiently between variables of different ranks.

In modern climate modeling applications, each rank reduction reduces the data size by about two orders of magnitude! For example, CCSM data typically have 100–200 points in the longitudinal dimension. Reducing  ${}^4T = T(t)$  into zonal-mean temperature  ${}^3T = [T]$  (brackets denote zonal-averaging) shrinks the dataset size by a factor of 100–200. Rank reduction on a remote server, followed by transfer of reduced data, versus transfer of raw data, followed by rank reduction on a local machine, has multiple-order of magnitude implications for the bandwidth required by DDR&A processing.

### 5.1.1 Algorithmic Description of Rank Reduction

The algorithms NCO (and, hence, SDO) use for rank-reduction make clear the intrinsic parallelism of many LDR&A operations. It is this parallelism we will exploit with Shared Memory Parallel Programming (SMPP) via [OpenMP](#) directives.

The masked, weighted average of a variable  $x$  can be generally represented as

$$\bar{x}_j = \frac{\sum_{i=1}^{i=N} \mu_i m_i w_i x_i}{\sum_{i=1}^{i=N} \mu_i m_i w_i} \quad (1)$$

where  $\bar{x}_j$  is the  $j$ 'th element of the output hyperslab,  $x_i$  is the  $i$ 'th element of the input hyperslab,  $\mu_i$  is 1 unless  $x_i$  equals the missing value,  $m_i$  is 1 unless  $x_i$  is masked, and  $w_i$  is the weight. This formidable looking formula represents a simple weighted average. NCO uses various permutations and extensions of this formula to compute related statistics such as masked, weighted sums, extrema (i.e., minima and maxima), and standard deviations.

When  $\mu_i = m_i = w_i = 1$ , the generic averaging expression above reduces to a simple arithmetic average. Currently,  $m_i = w_i = 1$  for all NCO operators except `ncwa`. These variables are included in the discussion below for completeness and because this project will add masks to other operators (cf. Section 6.1).

The size  $J$  of the output hyperslab for a given variable is the product of all the dimensions of the input variable which are not averaged over. The size  $N$  of the input hyperslab contributing to each  $\bar{x}_j$  is simply the product of the sizes of all dimensions which are averaged over. The input hyperslabs are independent of one another. Thus  $N$  is the number of input elements which *potentially* contribute to each output element. For a complete discussion of the conditions under which input elements contribute to the output hyperslab, see the [NCO User’s Guide \(Zender, 2004, p. 37\)](#).

### 5.1.2 Analytic Load Balancing (ALB) in LDR&A

We now describe the intrinsic parallelism of rank reduction. First, we note that no input element  $x_i$  contributes to more than one output element  $\bar{x}_j$  in (1). the outermost loop over the output hyperslab  $\bar{x}_j$  is the least rapidly varying dimension not averaged. We will investigate the efficacy of OpenMP directives placed around this loop. Both the input and output array may be stored as shared data since each input hyperslab maps to a single output element.

Consider our prototype climate model application where the user wishes to derive  $\langle \bar{T} \rangle(z) = {}^1T$  (the global area-mean, time-mean vertical temperature profile) from the raw data  $T(x, y, z, t) = {}^4T$ . In this case,  $N = N_x \times N_y \times N_t$  where  $N_x$ ,  $N_y$ , and  $N_t$  are the number of elements in the  $x$ ,  $y$ , and  $t$  dimensions respectively. For a typical NCAR CCSM IPCC simulation one hundred years in length stored at daily temporal resolution and T85×L40 resolution,  $(N_x, N_y, N_t) = (256, 128, 36500)$  and  $N = 1196032000$  ( $\sim 1.2 \times 10^9$ ). Thus each of the forty points in the output variable distills an average of about one-billion input points. Given that (1) involves about five floating point operations per input point (to handle masking, weighting) and additional logical operations, it seems fair to estimate about ten billion floating point operations per output point. This is well within the realm ( $N \gtrsim 10^9$ ) where OpenMP parallelism is likely to increase computational throughput rather than decrease it due to the overhead of setting up the threads themselves (Jim Tuccillo, IBM, personal communication, February, 2004). In other words, throughput efficiencies would be achieved by spawning up to  $N_z = 40$  threads, one per vertical level.

We call exploiting this parallelism Analytic Load Balancing (ALB) because its efficacy follows from analytic considerations that depend on the size of the hyperslab to be reduced to a scalar. This size will be evaluated automatically at runtime without user intervention, and will differ for variables of different ranks. OpenMP formalism within the rank reduction (1) will not always enhance throughput. Thus ALB will be invoked only if short run-time checks verify its efficacy. Implementing and tuning this algorithm will take place in Year 1.

### 5.1.3 Per-Thread Variables (PTV) in LDR&A

Per-Thread Variables (PTVs) are another promising way to parallelize rank reduction operations on multi-file operators. Assume each input file in (1) contains the same  $V$  variables  $v_1, v_2 \dots v_n \dots v_V$ . Computing each  ${}^{R_i}v_n \rightarrow {}^{R_t}v_n$  on a separate OpenMP thread inside the “variable loop” of the multi-file operators may be easier to implement, though not necessarily more efficient, than ALB

parallelization (Section 5.1.2). The PTV method appears to be robust for a mixture of  $R_i$  in the input file. Reductions on variables with smaller  $R_i$  will finish more quickly than those with larger  $R_i$ , and the OpenMP thread will simply proceed to the next variable requiring reduction. Since a typical CCSM simulation has  $V \sim 100$  the amount of idle thread time relative to total thread time is likely to be small since most computational geometries have eight or fewer CPUs per node. Implementing and tuning the PTV algorithm will take place in Year 1.

#### 5.1.4 Computational Geometry Load Balancing (CGLB) in LDR&A

The last form of parallelization possible for rank-reduction operations involves spawning a separate MPI process for reducing files. Consider  $M$  input files in (1):  $\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_m, \dots, \mathbf{F}_M$ . The simplest form of Computational Geometry Load Balancing would be to spawn  $M$  MPI tasks. Each task would obtain and perform DR&A on a least one file. In all likelihood,  $M$  is an upper bound on the optimal number of MPI tasks to spawn. A better optimized DR&A might spawn no more MPI tasks than local computational nodes that are available. Suppose there are  $H$  computational nodes,  $\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_h, \dots, \mathbf{Q}_H$ , available for the DR&A. Then a better number of MPI tasks to spawn might be  $\min(M, H)$ .

As discussed above in Section 5.1.2, it makes sense to parallelize code branches with more than about one billion floating point operations. Thus the most general case of LDR&A parallelization would use a hybrid approach relying on both OpenMP and MPI. For example, OpenMP could be used for PTV (Section 5.1.3) within a file. The number of OpenMP threads would be  $\min(V, L)$  where  $L$  is the number of CPUs per local node. MPI could spawn a separate task for each file (assuming the reduction operation commutes). Since the optimal balance of OpenMP (e.g., PTV) and MPI parallelization depends on the precise computational geometry available to the SDO client, we call this Computational Geometry Load Balancing (CGLB). Implementing and tuning the CGLB algorithm will take place in Years 2 and 3.

## 5.2 Latency and Bandwidth Efficiencies in DDR&A

We now characterize the computational barriers we must overcome to obtain efficient DR&A with large datasets stored at geographically disparate locations. In simple terms, traditional LDR&A is expressible as a sequence of operations on local files. With NCO, typical LDR&A multi-file operations are representable as

```
operator [options] file1 file2 ...fileM fileout (2)
```

Here `operator` stands for any NCO multi-file operator such as `ncra`, the netCDF Record Averages, which would time-average the input files and store the results in `fileout`.

With NCO compiled as OPeNDAP clients, file names may be URLs accessible through one or more OPeNDAP servers. With NCO OPeNDAP clients, DDR&A is possible and typically expressible as a sequence of operations on (local and) remote files.

```
operator [options] http://server1/file1 http://server2/file2
...http://serverK/fileM fileout (3)
```

Currently, the DDR&A example (3) requires high bandwidth because none of the averaging takes place on the OPeNDAP server. Instead, OPeNDAP transfers the raw input files back to

the local `ncra` client for processing. A classification of possible client-server DDR&A scenarios helps to clarify the barriers to more efficient DDR&A. We classify the scenarios based on four criteria:

1. Client-Server Traffic Protocol (CSTP): “Parallel” or “Serial”?  
DDR&A traffic is serial for a multi-file operator (e.g., `ncra`) that requests data synchronously, i.e., one file at a time, operates on it, then asks for the next file. Many useful DR&A operations (e.g., multi-file averaging) are commutative, and thus could be performed in any order to yield the same answer (to round-off precision, anyway). DDR&A traffic is parallel when a multi-file operator (e.g., `ncra`) requests multiple files asynchronously, and operates on them in some pre-defined or random order. CST volume is high for dumb servers and low for smart servers.
2. Latency: “High”, “Medium”, or “Low”?  
DDR&A is high latency if it often forced into wait states for other data. Serial CSTP is high latency by definition, since the operator goes into a wait state after processing each file, until the next input file request completes. The source of what we call “Medium” latency in DDR&A is arithmetic. Multi-file operations that do not commute (e.g., averaging with a temporal stride) must complete their operations in a specified order. Even if all files are requested in parallel, the arithmetic must always be executed synchronously and thus delays are possible obtaining data from slower servers. Commutative operations do not suffer from this constraint if the user foregoes bit-for-bit answer reproducibility. The option to forego bit-for-bit reproducibility for commutative operations with asynchronous I/O defines what we call “Low” latency.
3. Server-side Data Reduction & Analysis (SSDR&A): “Smart” or “Dumb”?  
Currently, the OPeNDAP servers broker client data requests without performing SSDR&A. These are “Dumb” servers since they return the full requested data hyperslab to the client with no intermediate processing. The client may only ultimately need reduced data (e.g., file averages). The cost of transmitting raw rather than reduced data is described in Section 5.1. Our goal is to move all feasible DR&A (e.g., rank reduction) from the client to the server to reduce the high-bandwidth requirements of transferring raw data.

The technical goal of this project is to develop SDO so that it transparently (to the user) functions with the desirable capabilities mentioned above and in Section 5.1.

Table 1 outlines our planned progression of DDR&A Scenarios starting from the present NCO↔OPeNDAP capability. Implementing SSDR&A is the most difficult task as it may involve changes to the netCDF API and to the OPeNDAP-netCDF client library (CL). By the end of Year 3, SDO will automatically identify logically segregable parts of multi-file jobs so that arithmetic and data reduction are done on the servers whenever possible. The results of the intermediate SSDR&A will be relayed to the SDO client for final aggregation and processing.

## 6 SDO Software Engineering

We will apply modern software engineering techniques to SDO. Already the NCO code-base is clean and well-documented internally. We make heavy use of metadata and [systematization](#) in variable and function names to impart a high level of literacy to the source code.

**Table 1: DR&A Scenarios**

Scenario	ALB <sup>a</sup>	CGLB <sup>b</sup>	CSTP <sup>c</sup>	Latency	SSDR&A <sup>d</sup>	Expected
NCO Clients lack OpenMP, MPI. Clients synchronously request data from servers. Clients do all DR&A.	No	No	Serial	High	Dumb	Present
SDO Clients gain OpenMP parallelism, request data asynchronously.	Yes	No	Parallel	Medium	Dumb	Year 1
SDO Clients gains MPI.	Yes	Yes	Parallel	Low	Dumb	Year 2
OPeNDAP Servers perform intermediate DR&A.	Yes	Yes	Parallel	Low	Smart	Year 3

<sup>a</sup>Analytically Load-Balanced: Requires OpenMP in SDO clients.

<sup>b</sup>Computational Geometry Load-Balanced: Achieved with client-side OpenMP and/or MPI.

<sup>c</sup>Client-Server Traffic Protocol: Requires client-side OpenMP and/or MPI.

<sup>d</sup>Server-Side Data Reduction & Analysis: Requires extensions to OPeNDAP netCDF Client Library and (possibly) netCDF4 library. Exploits client-side efficiency and load-balancing improvements.

## 6.1 Enhancing Robustness

The following standard software engineering practices will be applied to the NCO code-base during the execution of this project:

1. Unit and Regression Testing: Addition of self-diagnosing tests that no existing features break as new features are added.
2. Profiling: Analysis of code bottlenecks and scalability
3. Memory purification: Detection and elimination of dangling pointers, un-freed memory (e.g., with [Valgrind](#)).

All of these practices have been applied ad hoc to NCO in the past. The Scientific Programming Specialist will rigorously enforce, build into the code, and automate these practices where possible.

## 6.2 NCO/SDO new features

Dozens of major and minor features would make NCO more useful and robust and are enumerated on this [TODO list](#). This SDO project would help address many of these deficiencies as part of its robustness and standards conformance emphasis. The items that would advance NCO/SDO DDR&A functionality are candidate tasks for the Specialist position funded by this proposal. The top five candidate tasks not mentioned previously include:

1. Geographic [masks](#) for all operations (e.g., masking oceans with `--mask=Atlantic`)
2. [Climate & Forecast](#) (CF) Metadata Convention conformance
3. Rank-reduction and logical constraint operations for [ncap](#) and [ncwa](#)
4. Multi-file input capability for [ncap](#) and [ncwa](#)

5. More pre-defined complex arithmetic operations in [ncap](#), e.g., standard deviations, chi-squared tests, linear regressions.

This functionality would greatly enhance SDO's power and reduce the effort the working scientist needs to put into DDR&A to obtain useful results.

## 7 Project Coordination

PI Zender will take overall responsibility for project coordination. NCO was first publicly released in 1995, and became an Open Source project in 2000. There are currently three active NCO developers. Unlike related the base SDDF projects [netCDF](#) and [HDF](#), NCO has never received based institutional support. Rather like [OPeNDAP](#), NCO has grown organically because users need NCO functionality. [OPeNDAP](#) has been quite successful at obtaining institutional support (from NASA, NOPP, NOAA, and NSF) via the peer-review process. NCO/SDO will strive to emulate and learn from the [OPeNDAP](#) practices that have let it successfully transition to a robust, institutionally supported Open Source project.

### 7.1 Personnel

Zender will continue to lead the NCO/SDO development, establish SDO development priorities and policies, continue to entrain the Open Source community, and coordinate architecture decisions with [netCDF](#) and [OPeNDAP](#) developers at UCAR/Unidata and University of Rhode Island (URI), respectively (see attached letters of support from [Rew](#) and [Cornillon](#)). Zender will work with a full-time Specialist to help design SDO and to implement the optimizations for distributed data reduction and analysis (Section 5.2) and to enhance the software engineering and robustness of NCO/SDO (Section 6). Zender will work with and [ESS/ICS](#) graduate student to identify and optimize techniques that exploit the intrinsic parallelism that pervades tera-scale geophysical data reduction (Section 5.1).

Co-PI [Papadopoulos](#) will consult and advise on issues of SDO parallelism, load balancing, and network connectivity. His experience developing open source software on high performance computers is the key bridge between the existing, homegrown NCO software and the proposed robust, parallelized, distributable SDO. [Papadopoulos](#) helped develop the [Parallel Virtual Machine \(PVM\)](#) and [High-Performance Virtual Machines \(HPVM\)](#) libraries. He leads development of the [Rocks Clustering Toolkit](#) at [SDSC](#). [Papadopoulos](#) is a Co-PI on the [NSF Middleware Initiative GRIDS](#) center which has produced four public releases of integrated and tested grid software. [Papadopoulos](#) is also [OptIPuter](#) Co-PI and [Network Architect](#). As such he will oversee integration of the [ESMF](#) with [OptIPuter](#) that is essential to developing, testing, and optimizing SDO for high-bandwidth connected [HPCCs](#).

### 7.2 Management Style

Zender directs two other multi-investigator projects, the [Earth System Modeling Facility](#), larger in scale than this proposed SDO project, and the [netCDF Operators](#). NCO is the smaller scale, never-funded, OpenSource software project which SDO will leverage. Zender makes efficient use

of project coordination software such as [SourceForge](#) (for complete project coordination), [Mailman](#) (for project mailing lists), [wreq](#) (a work-request tracking system for prioritizing tasks), and extensive documentation on project [Home Pages](#). These techniques maximize project transparency and minimize confusion that arises through misunderstood responsibilities, requests, and goals.

All SDO software design, construction, and modification will employ Concurrent Versioning System (CVS) or its successor ([Subversion](#)) to facilitate distributed development. To facilitate collaboration, all SDO software and data made available for this project will be publicly accessible (read-only) through CVS and OPeNDAP servers, respectively. This will encourage the Open Source community to remain involved in SDO development. We believe strongly in unfettered exchange of software and data.

### 7.3 Schedule and Milestones

**Year 1. Milestones:** 1a. OpenMP parallelization for arithmetic operators; 1b. ESMF becomes OptIPuter node with 1–2 Gb s<sup>-1</sup> connectivity; 1c. Release NCO version 4 (exploits netCDF4+OpenMP)  
*Tasks:*

1. Parallelization with OpenMP
  - (a) Independent variables concurrently processed on separate threads
  - (b) Asynchronous Client→Server data requests (CSTP)
  - (c) Analytic Load Balancing (ALB)
2. Develop unit testing for crucial code paths
3. Develop benchmarking suite to assess LDR&A performance

*Travel:*

1. One-week visit by Specialist to Unidata to coordinate integration of netCDF4 features into NCO/SDO.

**Year 2. Milestones:** 2a. ESMF↔OptIPuter connectivity increases to 10 Gb s<sup>-1</sup> (depends on outside funding); 2b. Demonstrate DDR&A of CCSM data between/among multiple HPCCs: ESMF, NCAR, SDSC, ESC; 2c. Release NCO version 4.2 (exploits MPI)  
*Tasks:*

1. Analytic parallelization with MPI (CGLB)
2. Enhance SDO benchmark suite to assess DDR&A and CGLB performance gains

*Travel:*

1. One-week visit by Specialist to Unidata to coordinate netCDF and DODS netCDF CL API additions with Unidata netCDF and OPeNDAP projects.
2. PI Zender and Specialist attend ACM Supercomputing meeting to present science and to demonstrate DR&A technology.
3. PI Zender and ESS Graduate Student to attend Fall AGU meeting to present science and to demonstrate DR&A technology on CCSM IPCC simulation data.

**Year 3. Milestones:** 3a. SDDR&A on OPeNDAP servers and netCDF4; 3b. Prototype SDDR&A applicability to Genomic data; 3c. Release NCO version 4.4 (exploits SDDR&A)

*Tasks:*

1. Place SDDR&A hooks in netCDF4 library
2. Use SDO operators to add “file-out” DR&A to OPeNDAP netCDF CL
3. Hybrid parallelization with OpenMP/MPI (CGLB)
4. Measure DDR&A, ALB and CGLB performance gains
5. Profile CCSM DDR&A to identify remaining bottlenecks

*Travel:*

1. PI Zender and Specialist attend ACM Supercomputing meeting to present science and to demonstrate DDR&A and SDDR&A technology
2. PI Zender and ESS Graduate Student to attend Fall AGU meeting to present science and to demonstrate DDR&A and SDDR&A technology on CCSM IPCC simulation data.

## 7.4 Education Plan

This project will train one graduate at the interface of computational and geophysical sciences. The Specialist and graduate student will both conduct training workshops with Zender at appropriate national scientific conferences (Supercomputing and AGU) in Years 2 and 3. This will help ensure widespread institutional awareness of the project. Where possible we will pay work-study undergraduates at UCI to assist with programming tasks.

Interestingly, the primary NCO developer besides Zender for many years grew from novice programmer to expert writer of parsers and lexers by contributing to NCO in the Open Source fashion. With a true Open Source project, knowing how much outreach one is doing is impossible because the code is freely available. We have accepted source code contributions from about a dozen people in the last ten years. With a funded base, we have high hopes for entraining others, including programmers in less developed countries, and passing on our geophysical skill-sets to them.

## 7.5 Open Source Software

This SDO project, devoted to improving analysis of data often produced for pure research and/or public policy (e.g., geophysical measurements, climate models), will only thrive if it remains Open Source. NCO has been supported by the Open Source community for five years. The Open Source model provides for wider testing of the software. Public access to source code makes finding and fixing bugs much easier (e.g., [Raymond, 1999](#)). Volunteer developers now solve 50–75% of NCO bugs. SDO will remain, true to its roots, Open Source Software released under the GNU General Public License.

# 8 Broader Impacts and Synergies

All sciences utilizing LDR&A and DDR&A on netCDF data will benefit from efficiency improvements to NCO/SDO outlined in Sections 5 and 6, respectively. Geoscience fields which use data

storage formats other than netCDF or HDF will also benefit from SSDR&A improvements outlined in Section 5. netCDF and HDF have already been embedded in some bio-informatics computational languages to reduce data storage and access costs associated with XML and relational databases (RDBs) (Stein, 2004). All DR&A improvements are thus usable by appropriately formatted bioinformatics data sets, especially gene expression data, in ways similar to the geophysics domain.

This proposal is synergistic with a funded NASA REASON project “Merging the NetCDF and HDF5 Libraries to Achieve Gains in Performance and Interoperability” (Rew, 2003). This project is building netCDF3 on top of HDF5. The result, to be called netCDF4, will exploit the many performance advantages of HDF5 (parallel I/O, chunking, data compression) while retaining the simpler, more intuitive netCDF API. The PI, Russell Rew of Unidata (see attached letter of support) supports our proposal and will be involved in yearly meetings. The Co-PI, Mike Folk of the National Center for Supercomputing Applications (NCSA), has in the past expressed his support for bringing NCO functionality to HDF data.

Many of the DDR&A improvements we propose clearly fit the Information Integration (II) portion of this NSF RFP, especially the emphasis on **decentralized data-sharing**. This proposal supports Papadopoulos at a nominal level only. We plan to seek additional funding from NSF or NASA to support heavier involvement by his SDSC group to help exploit DDR&A parallelism. Our next project would be complementary to this NSF project. We will propose to completely abstract the SDO I/O layer, develop a native HDF back-end to SDO, and to fully exploit the MPI2 parallel I/O library to further reduce latency and bandwidth requirements of DDR&A.

# References

## Bibliography

- Blackmon, M., B. Boville, F. Bryan, R. Dickinson, P. Gent, J. Kiehl, R. Moritz, D. Randall, J. Shukla, S. Solomon, G. Bonan, S. Doney, I. Fung, J. Hack, E. Hunke, J. Hurrell, J. Kutzbach, J. Meehl, B. Otto-Bliesner, R. Saravanan, E. K. Schneider, L. Sloan, M. Spall, K. Taylor, J. Tribbia and W. Washington, 2001: The Community Climate System Model. *Bull. Am. Meteorol. Soc.*, **82**(11), 2357–2376. 4
- Grimi, A. and C. S. Zender, 2004: Roles of saltation, sandblasting, and wind speed variability on mineral dust aerosol size distribution during the Puerto Rican Dust Experiment (PRIDE). *J. Geophys. Res.*, **109**(D7), D07202, doi:10.1029/2003JD004233. 3
- Houghton, J. T., L. G. M. Filho, J. P. Bruce, H. Lee, B. A. Callander and E. F. Haites, Eds., 1995: *Climate Change 1994: Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios*. Cambridge Univ. Press, New York. 4
- IPCC, 2001. Cambridge Univ. Press, Cambridge, UK, and New York, NY, USA. 4
- NCAR, 2004: *Community Climate System Model*. National Center for Atmospheric Research, Boulder, CO, <http://ccsm.cgd.ucar.edu>. 4
- NCSA, 2004: *Hierarchical Data Format*. National Center for Super-Computer Applications, Champaign-Urbana, IL, <http://hdf.ncsa.uiuc.edu>. 2.1
- Raymond, E. S., 1999: *The Cathedral & the Bazaar*. O'Reilly Inc., Sebastopol, CA. 7.5
- Rew, R. K., 2003: Merging the NetCDF and HDF5 libraries to achieve gains in performance and interoperability. "<http://www.unidata.ucar.edu/proposals/NASA-AIST-2002/Description.pdf>". 8
- Stein, L. D., 2004: *Down with Species-Specific Database Projects, Up with Data Services*. Cold Spring Harbor, NY, <http://stein.cshl.org>. 8
- Sweeney, M. R., A. J. Busacca and D. Gaylord, 2003a: High accumulations rates and the generations of thick Palouse loess via topographic traps, Juniper Canyon. *Proc. Geol. Soc. Amer. 2003 Meeting*. 3
- Sweeney, M. R., A. J. Busacca, D. R. Gaylord and C. Zender, 2002: Provenance of Palouse loess related to late quaternary glacial outburst flooding in the Pacific Northwest. *Eos Trans. AGU*, **83**(47), H22B–0899. 3
- Sweeney, M. R., A. J. Busacca, C. A. Richardson, M. Blinnikov and E. McDonald, 2003b: The Columbia Plateau dust engine during the last glacial maximum: Trouble with cold starts. *Proc. XVI International Quaternary Association (INQUA) Congress*. 3
- Unidata, 2004: *Network Common Data Format*. Boulder, CO, <http://www.unidata.ucar.edu/packages/netcdf>. 2.2
- Zender, C. S., 2004: NCO User's Guide. "<http://nco.sf.net/nco.pdf>". 2.3, 5.1.1
- Zender, C. S., M. Flanner and J. Adams, 2003: LGM dust distribution and radiative forcing: Sensitivity to vegetation reconstruction. *Proc. XVI International Quaternary Association (INQUA) Congress*. 3

# Index

coordinates, [3](#)

DDR&A, [i](#)

dimensions, [3](#)

Earth Science Enterprise, [2](#)

ESE, [2](#)

HDF, [2](#)

National Center for Supercomputing Applications, [2](#)

NCSA, [2](#)

netCDF4, [i](#)

NSF, [i](#)

OptIPuter, [i](#)

SEIII, [i](#)

SSDR&A, [i](#)

variables, [3](#)

## 8.1 Budget Justification

% NB: Do not use LaTeX formatting in Budget Justification since must  
% upload into Liz's Word document

### Salaries and Wages

One month of summer salary support for three years is requested for Prof. Charles Zender, the PI at UCI, who will have primary responsibility for the proposed research.

Zender will lead the NCO/SDO development, establish SDO development priorities and policies, continue to entrain the Open Source community, and coordinate architecture decisions with netCDF and OPeNDAP developers at UCAR/Unidata and University of Rhode Island (URI), respectively (see attached letters of support from Rew and Cornillon).

To Be Named---One full-time Specialist Step I will share responsibility for SDO library design, and will have primary responsibility for library implementation, server-side extensions, profiling, regression testing, debugging, and SDO releases. The Specialist will work with the graduate student to profile, test, and improve the OpenMP and MPI modifications.

A 2% cost of living increase was applied each year of this proposal as well as a 5% merit, where applicable.

To be Named---Graduate Student Researcher III. Funds are requested to support one non-resident graduate student each year of the project. The graduate student support is requested at 50% for 9 months during the academic year and 100% for 3 months during the summer. The graduate student will work with Zender on optimizing OpenMP and MPI parallelization to exploit the intrinsic parallelism of common data reduction arithmetic, with both local and distributed data. All salaries and wages were estimated using UCI's academic and staff salary scales.

### Employee Benefits

Fringe Benefits were estimated using the composite rates agreed upon by the University of California Office of the President and the DHHS Audit Agency, the Cognizant Audit Agency for the University of California. Benefit rates used in this proposal are:

Faculty - summer - 12.7%

Academic (Specialist) - 17%

Student employees - summer - 3%

Student employees - academic year - 1.3%

Fees are requested for one nonresident student for the duration of the project. University of California policy requires award payment of fees for any student with more than 25% support from a grant. In the first year, \$21,147 is requested for non-resident fees and tuition, \$22,579 in the second year, and \$24,111 in the third year. Fees & tuition are excluded from indirect cost assessment.

#### Equipment

Equipment funds are requested for the first year only for two dual Opteron workstations at \$5,000 each. These workstations will be dedicated to the Specialist and the graduate student for use on this project.

The Cisco Catalyst WS-3550-12T switch and accompanying components provide a 1 gigabit/second Ethernet uplink or 2 gigabit/second Ethernet "Etherchannel" uplink to a connection on campus for Optiputer. An additional 2 GBICs for the second uplink are included to increase to 2 GE speed. The switch provides multiple GE copper interfaces and acts as a node connection point. In order to tie to the ESMF, the Cisco SW-C3508G GE switch and components is included. Equipment prices include tax, and shipping and handling charges. Equipment is excluded from indirect cost assessment.

#### Materials and Supplies

None are requested.

#### Travel

Domestic: One round-trip per year at \$1500 per trip is requested for the Specialist to travel to Denver/Boulder to visit with the Unidata netCDF and OPeNDAP projects. Each trip includes roundtrip travel from Irvine to Denver, one-week hotel and per diem. Travel support is requested in years 2 and 3 for the PI and graduate student to attend the ACM Supercomputing meeting to present science and to demonstrate distributed climate data reduction technology. \$1000 is requested each year for travel per person plus \$500 in shipping/rental fees at a total cost of \$2500 per year. Travel support is also requested in years 2 and 3 for the PI and Specialist to attend the Fall AGU meeting to demonstrate distributed climate data reduction technology. \$1000 is requested each year for travel per person plus \$500 in shipping/rental fees at a total cost of \$2500 per year. These trips include estimated conference registration, abstract submission fees, RT airfare, lodging, meals and ground transportation. Travel estimates are based on historical usage.

#### Other Direct Costs

Charges for journals, photocopying, long distance phone, fax and postage charges pursuant to this project are requested each year. Included in these expenses are long-distance charges for usage directly related to the project, such as communication with colleagues, journals, and vendors. Photocopying of research materials including publications and results of this sponsored research project are requested as well as mail and shipping for materials related to this project. Support is requested in years 2 and 3 for publication costs pursuant to this project, which include utilization of expensive color figures. Costs were estimated according to historical usage.

Subaward to UCSD: \$12,918 is requested in the first year. This subaward will fund one month of salary for Aaron Chin, the UCSD OptIPuter project manager, to install a OPeNDAP server on the UCSD OptIPuter, and to configure it for benchmark studies of SDO in a high-bandwidth distributed data mode.

#### Indirect Costs

Facilities and Administrative costs were estimated in accordance with UCI's approved indirect cost rate agreement. The indirect cost rate of 51.5% of MTDC through 6/30/05 and 52.5% of MTDC effective 7/1/05 was based upon the nature and location of the work proposed. Graduate student fees and tuition and equipment are excluded from indirect cost assessment. The subaward to UCSD is not subject to indirect cost assessment due to a University of California multicampus agreement UCI's indirect cost rate agreement was approved by DHHS, the Federal Cognizant Audit Agency for UCI on 12/5/01.

## 9 Facilities, Equipment, and Other Resources

### 9.1 Computer and Networking

Our SEI(GEO) project is well-situated to take advantage of the fastest fastest network connections at UCI and UCSD. The UCI Network Infrastructure provides researchers with  $1.0 \text{ Gb s}^{-1}$  access to the high-performance network of Cal-(IT)<sup>2</sup> and to the Gbbackbone of UCINet. UCI will upgrade this link to  $10 \text{ Gb s}^{-1}$  in the near future. This will remove one potential bottleneck to the ESMF $\leftrightarrow$ OptIPuter connection.

PI Zender is director of the Earth System Modeling Facility (ESMF), an NSF-supported MRI facility dedicated to coupled global climate, chemistry, and biogeochemistry simulations. The ESMF is an 88-CPU Power4+ IBM supercomputer with 192 GB RAM and 32 TB of RAID storage. Since this SEI proposal is highly complimentary the ESMF mission, the ESMF will be made available for NCO/SDO/OPeNDAP development, benchmarking, and test. Funding for bi-directional  $2 \text{ Gb s}^{-1}$  connections between the ESMF and UCI's Campus portal is requested as part of this SEI proposal. The Tera-Scale distributed data reduction to be optimized in this SEI proposal will be demonstrated between two or more geographically disparate supercomputer data-farms. The 30 TB ESMF RAID storage will typically be one of those data-farms. The ESMF will place at least 1 TB of storage under the control of a OPeNDAP server in support of this SDO project.

Co-PI Papadopoulos is Program Director for Grid and Cluster Computing at the SDSC. As Co-PI of the Cal-(IT)<sup>2</sup> OptIPuter, Papadopoulos manages the storage, clusters, and grid part of the UCSD OptIPuter. The targeted OptIPuter node for the OPeNDAP server is an IBM storage cluster (see attached letter of support from Larry Smarr). We plan to partition a portion of the cluster for OPeNDAP services. The cluster consists of 48 storage nodes with a single management node. Each node is an IBM xSeries 345 2U rack mount server with dual 2.8 GHz Xeon Processors. There is 1.5 GB RAM and six 73 GB drives for a total of 2.19 TB of storage each. The applications will access the storage via PVFS. The IBM storage cluster is connected to the OptIPuter network today at  $4 \text{ Gb s}^{-1}$ .

Our project will conduct distributed data reduction at two other sites in addition to the ESMF and UCSD, which share a high-bandwidth connection. The other sites are NCAR and Japan's Earth Simulator Center (ESC). These are also world-class supercomputer facilities. Facility managers or directors at these sites have expressed great interest to Zender in making available about 1 TB of storage available through OPeNDAP servers for proof-of-concept benchmarking for our study (Personal communication, 2004, Dr. Tetsuya Sato, Director-General, Earth Simulator Center, Japan; personal communication, 2004, George Fuentes, Head, High Performance Systems Section, SCD, NCAR). Accessing these sites via the standard research Internet requires no additional networking support or hardware. Deployment and operation of the OPeNDAP servers at these sites will be worked out on an informal basis once the SDSC connection is complete.

### 9.2 Maintenance and Technical Support

Network and Academic Computing Services [NACS](#) is the largest IT organization at UCI. Dr. Frank Wessel manages the NACS Research Computing Support Group (RCS). Dr. Wessel is the NACS project lead for the Earth System Modeling Facility (ESMF), a recently funded NSF MRI with C. Zender as PI. RCS provides customized support and facilitates user access to high-performance

computing (HPC) resources, software, training, and development of the UCI research infrastructure. NACS RCS staff led by Dr. Wessel and Garrett Hildebrand will facilitate and oversee the dedicated network infrastructure to link the ESMF to the UCSD OptIPuter. To implement the required network connection for the high-speed ESMF↔OptIPuter network, NACS will upgrade network facilities with additional switches and interconnects provided for in the budget. UCSD OptIPuter project manager Aaron Chin will ensure smooth connectivity on the UCSD end.

## 10 Acronyms and Abbreviations

**Table 2: Acronyms and Abbreviations**

Abbreviation	Description
ABI	Application Binary Interface
ACM	Association for Computing Machinery
AGU	American Geophysical Union
AMWG	(CCSM) Atmospheric Model Working Group
API	Application Programmer Interface
CAM	Community Atmosphere Model
CCM	Community Climate Model
CCSM	Community Climate System Model
CENIC	Corporation for Education Network Initiatives in California
CF	Climate & Forecast (metadata convention)
CL	Client Library
CPU	Central Processing Unit
CST	Client-Server Traffic
CSTP	Client-Server Traffic Protocol
CVS	Concurrent Versions System
Cal-(IT) <sup>2</sup>	California Institute for Telecommunications and Information Technology
CalREN-2	California Regional Network
CalREN-XD	California Research Network, Experimental Development
DDRA	Distributed Data Reduction & Analysis
DEAD	Dust Entrainment And Deposition Model
DKRZ	Deutsches Klimarechenzentrum
DODS	Distributed Oceanographic Data Server
DRA	Data Reduction & Analysis
ESA	European Space Agency
ESC	Earth Simulator Center
ESE	Earth Science Enterprise
ESMF	Earth System Modeling Facility
ESS	Earth System Science (Department)
FAR	First Assessment Report
TAR	Third Assessment Report
SAR	Second Assessment Report
AR4	Fourth Assessment Report
GB	Gigabyte
GCM	General Circulation Model
GFS	Global File System
GPFS	General Parallel File System
Gb	Gigabit

**Table 2:** (continued)

Abbreviation	Description
HDF	Hierarchical Data Format
HIPerWall	High-Performance Visualization System for Collaborative Earth System Sciences
HPCC	High Performance Computing Center
HPVM	High-Performance Virtual Machines
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICS	Information and Computer Sciences
II	Information Integration
IPCC	Intergovernmental Panel on Climate Change
JISAO	Joint Institute for the Study of the Atmosphere and Ocean
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
MM5	PennState/NCAR Mesoscale Model version 5
MPMD	Multiple Program Multiple Data
NACS	Network and Computing Services
NASA	National Aeronautic and Space Administration
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
NCVweb	NetCDF Viewer Web
NERSC	National Energy Research Scientific Computing Center
NOAA	National Oceanic and Atmospheric Administration
NOPP	National Oceanographic Partnership Program
NPACI	National Partnership for Advanced Computational Infrastructure
NPOESS	National Polar-orbiting Operational Environmental Satellite System
NSF	National Science Foundation
NUG	NCO User's Guide
ONI	Optical Networking Initiative
OPeNDAP	Open-source Project for a Network Data Access Protocol
ORNL	Oak Ridge National Laboratory
OpenMP	Standard API for SMPP
OptIPuter	Optical networking Internet Protocol computer
PI	Principle Investigator
PVFS	Parallel Virtual File System
PVM	Parallel Virtual Machine
RAM	Random Access Memory
RCS	Research Computing Services
RDB	Relational databases
RFP	Request for Proposals

**Table 2:** (continued)

Abbreviation	Description
SDDF	Self-describing Data Format
SDSC	San Diego Supercomputer Center
SEI	Science and Engineering Informatics
SMP	Symmetric Multi-Processing
SMPP	Shared Memory Parallel Programming
SP	Senior Personnel
SPMD	Single Program Multiple Data
SRES	Special Report on Emissions Scenarios
SSDRA	Server-Side Data Reduction & Analysis
TB	Terabyte
UCAR	University Corporation for Atmospheric Research
URI	University of Rhode Island
WRF	Weather Research and Forecasting (model)
XML	Extensible Markup Language
netCDF	network Common Data Format
p655	8-CPU IBM computer node in ESMF
p690	32-CPU IBM computer node in ESMF

## 11 Project-Wide Combined Collaborator and Advisor List

All Personnel Associated with Proposal, Collaborators and Co-Editors of Project Senior Personnel, their Post-docs, and their Thesis Advisors:

Ammann, C. A. (NCAR)  
Bian, H. (NASA/UMBC)  
Bonan, G. B. (NCAR)  
Busacca, A. (WSU)  
Chien, Andrew, University of California, San Diego  
Colarco, P. (GSFC)  
Collins, W. D. (NCAR)  
Cooper, W. A. (NCAR)  
DeFanti, Tom (UIC)  
Dongarra, Jack, University of Tennessee, Knoxville  
Famiglietti, J. (UCI)  
Foster, Ian, Argonne National Laboratory  
Garrett Hildebrand (UCI)  
Gaylord, D. (WSU)  
Geist, George, Oak Ridge National Laboratory  
Grimshaw, Andrew, University of Virginia  
Grini, A. (U. Oslo)  
Kesselman, Carl, ISI, University of Southern California  
Kiehl, J. T. (NCAR)  
Kohl, James, Oak Ridge National Laboratory  
Kuester, F. (UCI)  
Mahowald, N. M. (NCAR)  
Maxine Brown (UIC)  
Messina, Paul, Caltech  
Moore, J. K. (UCI)  
Nachtigal, Noel, Sandia National Laboratories  
Okin, G. (U. Virginia)  
Pajarola, R. (UCI)  
Pratt, Thomas, Sandia National Laboratories  
Rasch, P. J. (NCAR)  
Riesen, Rolf, Sandia National Laboratories  
Sanderson, James, Los Alamos National Laboratory  
Semeraro, David, National Computational Science Alliance  
Shelton, William, Oak Ridge National Laboratory  
Smarr, Larry, University of California, San Diego  
Stammer, Detlef, Scripps Institution of Oceanography  
Sunderam, Vaidy, Emory University  
Thomas, G. T. (CU)  
Torres, O. (NASA GSFC)  
Valero, F. P. J. (Scripps)

Yu, S. (Duke)

List is Alphabetical by Surname.

Collaborators of Zender:

Ammann, C. A. (NCAR)  
Bian, H. (NASA/UMBC)  
Bonan, G. B. (NCAR)  
Busacca, A. (WSU)  
Colarco, P. (GSFC)  
Collins, W. D. (NCAR)  
Famiglietti, J. (UCI)  
Gaylord, D. (WSU)  
Grini, A. (U. Oslo)  
Kiehl, J. T. (NCAR)  
Kuester, F. (UCI)  
Mahowald, N. M. (NCAR)  
Moore, J. K. (UCI)  
Okin, G. (U. Virginia)  
Pajarola, R. (UCI)  
Rasch, P. J. (NCAR)  
Valero, F. P. J. (Scripps)  
Yu, S. (Duke)  
Torres, O. (NASA GSFC)  
Thomas, G. T. (CU)  
Kiehl, J. T. (NCAR)  
Cooper, W. A. (NCAR)

Collaborators of Papadopoulos

Sunderam, Vaidy, Emory University  
Dongarra, Jack, University of Tennessee, Knoxville  
Geist, George, Oak Ridge National Laboratory  
Kohl, James, Oak Ridge National Laboratory  
Nachtigal, Noel, Sandia National Laboratories  
Pratt, Thomas, Sandia National Laboratories  
Shelton, William, Oak Ridge National Laboratory  
Riesen, Rolf, Sandia National Laboratories  
Sanderson, James, Los Alamos National Laboratory  
Semeraro, David, National Computational Science Alliance  
Chien, Andrew, University of California, San Diego  
Smarr, Larry, University of California, San Diego  
Stammer, Detlef, Scripps Institution of Oceanography  
Grimshaw, Andrew, University of Virginia  
Foster, Ian, Argonne National Laboratory  
Kesselman, Carl, ISI, University of Southern California  
Messina, Paul, Caltech

Collaborators of Chin:

Maxine Brown (UIC)  
Tom DeFanti (UIC)  
Garrett Hildebrand (UCI)

## 11.1 Supplementary Documents

1. `${DATA}/prp/prp_sei/prp_sei_ltr_cornillon.pdf`
2. `${DATA}/prp/prp_sei/prp_sei_ltr_smarr.pdf`
3. `${DATA}/prp/prp_sei/prp_sei_ltr_rew.pdf`
4. `${DATA}/prp/prp_sei/prp_sei_clb.pdf`
5. `${DATA}/prp/prp_sei/prp_sei_abb.pdf`