Future Generation Computer Systems 25 (2009) 124–131

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Building an OptIPlanet collaboratory to support microbial metagenomics

Larry Smarr [a], Paul Gilna [a,*], Phil Papadopoulos [a,b], Thomas A. DeFanti [a], Greg Hidley [a], John Wooley [a], E. Virginia Armbrust [c], Forest Rohwer [d], Eric Frost [d]

[a] California Institute for Telecommunications and Information Technology (Calit2), University of California, San Diego (UCSD), United States
[b] San Diego Supercomputer Center, University of California, San Diego (UCSD), United States
[c] University of Washington, Seattle, WA, United States
[d] San Diego State University, San Diego, CA, United States

ARTICLE INFO

Article history:
Received 11 March 2008
Accepted 12 June 2008
Available online 5 July 2008

Keywords:
Metagenomics
Microbial
CAMERA
OptIPuter
OptIPortal

ABSTRACT

We describe early experiments in the adoption of the OptIPuter architecture to provide data-intensive capabilities to several remote users of a large-scale, multi-year effort to organize and make publicly available data describing a wide variety of marine microbial ecologies, their genomic content, and the local environments in which they live—marine microbial metagenomics. Microbial genomes are millions of base pairs in length, requiring both a global view of the genome and the ability to zoom into detail interactively, enabled by the OptIPortal. We describe the design of a scientific data and compute server, enhanced by OptIPuter technologies, and early examples of its use in support of high performance science applications in this emerging scientific field.

1. Introduction

A new frontier field which studies the metagenomics of microbial ecologies [1] is emerging at the interface of genomics, environmental sciences, and information technology. This new field examines the interplay of hundreds to thousands of microbial species present at a specific environmental location in space and time. Each individual organism's genome sequence is studied as a tightly coupled part of an entire biological community. This means that each individual sequence can now be considered from the worldview of the ecological sciences: the composition of the rest of the community, the environmental conditions in which it is found, and its relationships with other species with which it is found at other times and places. It also sets the stage for many new breakthroughs to occur in basic science, medicine, alternate energy sources, and environmental cleanup.

In this paper we will describe the use of the NSF-funded OptIPuter [2] technologies to provide high bandwidth data-intensive capabilities to a multi-year effort to organize and make publicly available data describing a wide variety of marine microbial ecologies, their genomic content, and the local environments in which they live. This project, funded over seven years by the Gordon and Betty Moore Foundation (GBMF) [3], will develop an innovative state-of-the-art Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA). The participants include the California Institute for Telecommunications and Information Technology (Calit2) [4], the J. Craig Venter Institute (JCVI) [5], the Scripps Institution of Oceanography's Center for Earth Observations and Applications (CEOA) [6], and the San Diego Supercomputer Center (SDSC) [7].

2. CAMERA and marine microbial metagenomics

Carl Woese [8] and his collaborators have shown using 16S RNA [9] phylogenies that all living things can be evolutionarily [10] resolved into three "kingdoms"; Bacteria, Archeae, and Eukarya. While the macroscopic biological world of animals and plants belong to the Eukarya, the vast majority of organisms on Earth are single celled organisms. In fact, most of what we often think of as evolution – namely, the evolution of multi-cellular life – is but a small piece of life's history. Single-celled life seems to have been present on Earth for at least 3.5 billion years, whereas today's familiar multi-cellular life rapidly evolved after the Cambrian Explosion of 0.5 billion years ago [11], and the evolutionary steps that led to humans began only after a celestial impact removed dinosaurs from the world just 0.065 billion years ago.

Thus, microorganisms on land and in the ocean play fundamental roles today in every ecosystem on Earth and hold the key to

* Corresponding author. Tel.: +1 858 337 3187.
E-mail addresses: lsmarr@ucsd.edu (L. Smarr), pgilna@ucsd.edu (P. Gilna), philip.papadopoulos@gmail.com (P. Papadopoulos), tdefanti@ucsd.edu (T.A. DeFanti), ghidley@soe.ucsd.edu (G. Hidley), jwooley@sdsc.edu (J. Wooley), armbrust@ocean.washington.edu (E. Virginia Armbrust), frohwer@gmail.com (F. Rohwer), eric.frost@sdsu.edu (E. Frost).

0167-739X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2008.06.009

knowledge of the first three billion years of life's evolution that took place in the oceans of our planet. If we are to discover the great majority of genes that record life's evolutionary experiments with a constantly changing environment, we must look to ocean microbes.

Microorganisms play fundamental roles in every ecosystem on Earth. As Norman Pace, one of the pioneers of microbial evolution observed, "Life's diversity is mainly microbial in nature. Although the biosphere is absolutely dependent on the activities of microorganisms, our understanding of the makeup and natural history of microbial ecosystems is, at best, rudimentary. One reason for the paucity of information is that microbial biologists traditionally have relied on laboratory cultures for the detection and identification of microbes. Yet, more than 99% of environmental microbes are not cultured using standard techniques. As a consequence, the makeup of the natural microbial world remains largely unknown" [12]. The field of research known as metagenomics (or environmental genomics) has recently emerged largely as a means of circumventing the need to culture a given microbe in order to study it.

As an example, in April 2004, a major breakthrough in the scale of application of the techniques in metagenomics was published by Craig Venter and his team at the J. Craig Venter Institute in Rockville, Maryland in Science [13]. Whole-genome shotgun sequencing [14], which had been developed to sequence the human genome, was performed on samples taken from an entire microbial ecological community in the Sargasso Sea off Bermuda. Micron-scale pore filters were used to remove from the sample most eukaryotes (larger than 3 microns) and the smaller dissolved DNA and viral particles (smaller than 0.1 micron). The particular region was chosen because it is nutrient-limited and therefore was thought likely to have limited biological content. In spite of that, the genomic analysis revealed at least 1800 genomic species, including 148 previously unknown bacterial phylotypes. Compared with known genes previously reported in the National Institutes of Health's GenBank [15], the water sample from this single "eco-niche" in the ocean yielded an amazing 1.2 million previously unknown genes.

Following up this research, the Venter Institute then conducted the most comprehensive study of marine microbial biodiversity ever carried out. The Sorcerer II Expedition [16] collected and sequenced samples from around the world's oceans, creating the Global Ocean Sampling (GOS) project. Using the Sargasso Sea experiment as a prototype, the Sorcerer II Expedition took samples roughly every 200 miles, with more detailed sampling in sites such as the Galapagos Islands and the coast of Australia. At each site, the water samples were filtered and the specimens frozen and flown to the Venter Institute. There robotic sequencers performed the shotgun sequencing of the communities' collective DNA, funded by the GBMF.

The highly efficient use of shotgun sequencing of microbial ecologies, as compared with the traditional approach of culturing individual microbes in the laboratory, has produced a rapid increase in the number of genes sequenced [17]. Besides the Sargasso Sea and GOS studies, communities investigated have come from soil, deep sea sediments, acid mines, and human feces. Venter Institute has plans, funded in part by the Sloan Foundation, to sequence the DNA from bacteria, viruses, fungi, and other microbes in New York City air [18]. CAMERA is building a state-of-the-art computational and collaborative data-analysis facility to house all this metagenomics data, with the option of using unprecedented high-performance OptIPuter access to the end-user.

This transition from individual organism genomics to ecological metagenomics is having a dramatic impact on biological research today. The discovery of so many new genes gives us the opportunity to consider innovative biological approaches to alternate energy [19], pharmacology [20], environmental cleanup [21], and even climate-change adaptation strategies [22]. Given the scientific advances and economic benefits that the metagenomic study of microbial ecologies can produce, it is not surprising that it has taken off as a new branch of science.

3. The CAMERA OptIPuter data server and the campus Lambda-Grid

The NSF-funded OptIPuter project is focused on establishing "end-to-end" dedicated light waves (1- or 10 Gbps wavelengths on fibers termed "lambdas") from the end user's laboratory, across campus and wide-area networks (National Lambda Rail [23], Internet2, and the Global Lambda Integrated Facility [24]) to remote data repositories, compute resources, scientific instruments, or colleagues, dramatically improving the flow of data to the end-user.

In addition to providing web access to genomic data over the shared Internet to over 2000 scientists from more than 50 countries, CAMERA was architected so that it could also enable high end users to take advantage of OptIPuter technologies. Within one month of the award, Calit2 had assembled a 64-processor development and pre-production validation cluster with 12 Terabytes of storage. On a longer time scale, Calit2 staff members increased the size and upgraded the capabilities of the Calit2 server room [25], and procured and installed a larger production server. Both development and production clusters are used heavily for different aspects of the project.

The current CAMERA production configuration revolves around a 128-node (512 CPU core) Intel Woodcrest Cluster ($\sim$5 Teraflops) with 512 GB of aggregate memory. All nodes are interconnected via 1 GbE and half those nodes are interconnected with 10 Gbit Infiniband. All nodes, except for storage, run Linux CentOS 4 and is managed using SDSC Rocks [26]. Eight Sun Microsystems "Thumpers" (24 TB raw storage servers) run Sun's ZFS file system on a Solaris OS and are each connected at 10 GbE, providing nearly 200 Terabytes of storage dedicated to the production CAMERA server (see Fig. 1). Storage is arranged in a replica pairs. Using ZFS snapshots and physical replication means that several logical and physical copies of data are always available. A collection of Postgres Database servers, cluster management, grid login, and web servers make up the balance.

The CAMERA complex is embedded into the UCSD OptIPuter which has dedicated optical links connecting several major laboratories on campus. A companion NSF award, Quartzite [27] (which added optical switching to packet switching in this infrastructure), previews what campuses' infrastructure needs to evolve to: immense bandwidth, optical circuits on demand, and reconfigurable end-point systems. A production CAMERA switch router, a Force10 E1200, connects all CAMERA nodes and has 10 Gbit external connections to UCSD campus optical networking and to the remote OptIPuter fabric. The UCSD optical infrastructure was designed so that other campuses could clone it, thus establishing similar "on-ramps", as is now happening at Calit2's other campus at UC Irvine. Work is underway to get similar optical infrastructures on many other campuses around the world.

As a practical example of what this OptIPuter campus-scale implementation enables, we consider CAMERA computing challenges. While CAMERA data sets are physically small (e.g. fitting on a commercial drive), it is actually bandwidth to data that dominates the calculation. User's entering CAMERA through the online BLAST submission run a parallel program to calculate results. The degree of parallelism depends on what is actually being searched, but a typical search consumes about 100 CPUs for a single query

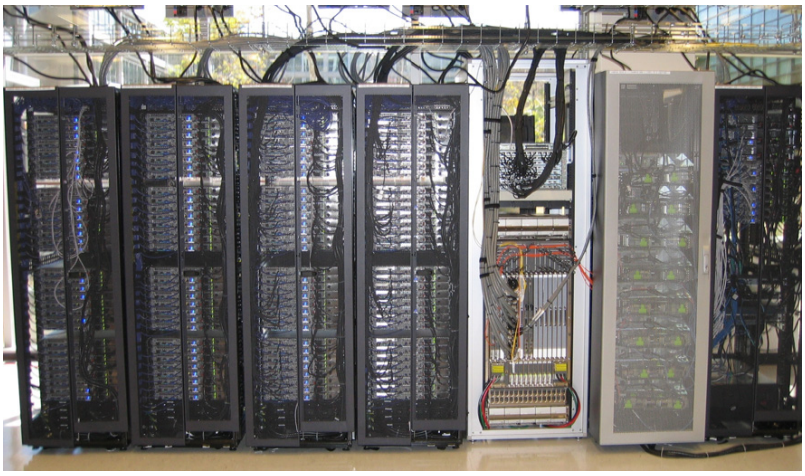


Fig. 1. The CAMERA OptIPuter server complex at Calit2@UCSD.

(or about 25% of our production server). For single sequence alignments, the entire submission, parallel search, and return results to user is completed in ∼10 s.

However, sometimes our production complex does not provide enough horsepower to complete a particular analysis. Using our OptIPuter dedicated optical network on campus for connectivity, we can acquire other computing resources on an as-needed basis. For example, in preparation for a new release of software a large number pre-computed BLAST results are needed for the JCVI-authored fragment recruitment viewer (FRV). To support, the FRV calculations, we temporarily dedicated an additional 200 processors of a remote cluster at SDSC. By connecting at 10 GbE to the CAMERA database and storage, these remote nodes appeared local and became an extension to the CAMERA complex. In this configuration, FRV calculations ran on the SDSC cluster for ∼1 week (about 4 CPU-years of calculation).

Should this SDSC remote cluster, which is part of the OptIPuter experimental complex, prove to be insufficient, there is enough network bandwidth to extend to the larger resources in Teragrid [28] in a nearly identical manner. Making the CAMERA resource directly available at high-bandwidth eliminates complexity for the applications and management — essentially CAMERA extends into resources for these occasional large-scale calculations. This looks like "cloud" computing, but in reality is not nearly as complex on the software side. Our very capable (and reconfigurable) network allows us to easily connect to other physical resources and make them appear (temporarily) as if they exist within the same machine room.

In summary, the core CAMERA computing and storage complex is built from commodity components and software so that existing parallel applications, web servers/services front-ends, database servers, large flat-file storage, and grid-based authentication can all function without requiring significant changes. However, where CAMERA differs from the majority of cluster complexes is that it is also "OptIPuter enabled" by 10 Gbit Ethernet dedicated connections to external labs through NLR. This enables the subclass of CAMERA remote researchers who desire much higher performance to treat the CAMERA data and computing complex as a "peripheral" to their lab without being handicapped by insufficient connectivity. For authenticated labs and users, they can have read-only 10 Gbps lambda direct access to CAMERA database servers and flat-file data farms. We will discuss two such implementations in Section 5.

4. The software architecture of the CAMERA OptIPuter data server

We use the Rocks cluster toolkit to define all Linux nodes in CAMERA for scalability, reliability and reproducibility. There are currently eight different node configurations, called "appliances" in Rocks, which include Compute, front-end, database, portal/web-server, authentication, registration, application servers, and monitoring consoles. The complexity of CAMERA required changes in the Rocks toolkit to simplify the management of "staging" and "production" versions. When preparing a new release of software, a single administrator builds a smaller-scale, but fully-functional set of servers for testing – when validated, only a small number of variables are changed in a single MySQL database to define the production environment. At this point, the production servers are rebuilt in exactly the same manner as our validation/staging complex. Our mechanisms and methods allow for production environments to be built programmatically and dramatically reduces the variability of human administration.

CAMERA data is held in SQL databases to provide a well-defined internal data model. However, most bioinformatics codes operate on standard flat-file "databases". NCBI-BLAST (which is the core BLAST algorithm used in CAMERA) requires its own binary formatted files from text-based FASTA files. Today, when a user requests a BLAST through the portal, we make the BLAST execution parallel (e.g. up to 135-way parallelism when running against all ORFs) and segmented binary files must be available. When designing CAMERA, our first design included replication of these binary data files on all local disks. However, using Sun Fire X4500 storage servers connected at 10 Gbit, holding these binary files on a traditional network file server, provides excellent performance and eliminates another layer of data management. In addition we take advantage of ZFS' built-in snapshot capability to replicate data among storage servers for data integrity. The large available storage (∼200 TB raw) allows us to keep several versions of data as well as output from users so that they can easily retrieve past results.

As CAMERA data expands in size, performance expectations may require more than the 1 GB/s (10 Gbit/s) delivered by a single X4500 file server. Should it be necessary, data can be replicated across X4500s or striped using a parallel file system to achieve the necessary aggregate performance. One possible solution could be to use the internal drives of compute nodes and carefully manage where data is replicated. However, our current deployment of

storage servers contains almost 400 disk drives or about 3X the number of drives in the entire cluster. The high-level of disk integration in the X4500s allows us to use a much simpler design and achieve excellent performance.

Our overall systems goal is one of complete reproducibility of the data products and the computing complex itself. To this end, we integrate a large number of externally authored tools and external data and make both available to our users. Through Rocks, we have an excellent mechanism for reproducing the infrastructure itself so that users have a transparent view into the system that they are using to analyze data. This reduces uncertainty in the process. With the large number of metagenomic datasets that CAMERA is now planning to house, we are developing a similar methodology for ingesting raw data from external projects (''providers'') and then providing this data in a variety of formats as needed by various tools.

CAMERA data has two-parts — voluminous sequence data (and assemblies) and environmental metadata. Both types of data are of interest to scientists, but the environmental data (e.g. geo-location, salinity, temperature, chlorophyll, time-of-day, and more) is essentially unstructured with few, if any, metafields being common to every single sequence. For example, chlorophyll counts make sense in marine genomic datasets, but have less relevance to air or intestinal genomic sets. Enabling scientists to sort through, select, and operate against subsets of sequence data based on meta-data attributes is at the core of what CAMERA is trying to accomplish.

We will ''master'' all data in a traditional SQL database and then produce various formats of the master data as a combination of database reports and well-defined programs. For example, raw sequence and metadata will be held in a database and then a FASTA format report will be written as output from the database. A segmented database for parallel NCBI BLAST is then created from these files. The data, no matter the format, can be traced back to the database master copy. For this workflow, any database will work (e.g., MySQL, Postrgres, Oracle, IBM DB2, and others).

A dedicated NLR gigabit/s path was researched, designed, and implemented between Calit2 and JCVI. This has been essential for the efficient transfer of applications software and large data sets. When UCSD receives a release candidate of software from JCVI (or other providers), it often has data associated with it. We set out to make ∼200 GB long-haul transfers commonplace (and more efficient than the traditional FedEx). Using our dedicated link and UDT from Grossman Lab at University of Illinois, Chicago, we are easily able to sustain about 35 MB/s for disk-to-disk transfer, but disks are 2500 miles (∼4200 kM) apart! We compared to using the identical software over a more congested route and saw speeds average about 5 MB/s. Total data transfer consumes about 100 min over NLR (vs. 700 min or 11.5 h over a congested public network). Additional work on using more optimal disks systems (both source and sink) would result in further time reductions when using a dedicated circuit.

5. Two working examples of remote metagenomics OptIPortals

It was recognized early in the OptIPuter project that users of 1- or 10 Gbps dedicated optical networks needed to use their laboratory Linux clusters as the end-points to terminate the optical circuits (as a telephone terminates the telephonic circuit). So the OptIPuter project developed a standard recipe for how to modify a Linux cluster to make it an ''OptIPortal'', OptIPuter end-point customized for scalable computing, storage, or visualization [29].

Combined with 10 Gbps campus optical fiber (discussed above) ''on-ramps'', connecting the end-user OptIPortal Linux cluster laboratory analysis facility to the NLR and then to CAMERA, the user will see up to a 10–100 fold increase in bandwidth accessing our CAMERA facility compared to access over the traditional shared Internet using a web browser. The OptIPuter project is working with a few CAMERA early adopters who have deployed OptIPortals in their labs and are customizing them to the needs of microbial metagenomics researchers (see Fig. 2). To establish full OptIPuter end-to-end cyberinfrastructure to their laboratories required detailed studies of the networking connectivity between Calit2 and these sites, as well as across their campuses. In this section, we will illustrate how in two such laboratories the display walls were set up and used for metagenomics, linking to the Calit2 CAMERA server. The two labs are the Forest Rohwer Lab at San Diego State University (regional-scale OptIPuter) and E. Virginia Armburst's Lab at the University of Washington, Seattle, WA (national-scale OptIPuter).

In the Summer of 2007, the Rohwer Lab installed a 9600×3600 pixel (∼35 Mpixel) OptIPortal tiled wall driven by [30] 8 DELL PCs running ROCKS. The cluster is being used to analyze metagenomic datasets generated by the Rohwer group and to develop new bioinformatic metagenomic data analysis tools, which are slated for deployment on the CAMERA compute server. In essence, through it's adoption of the OptIPortal architecture, the Rohwer group has established a local CAMERA cluster development environment that allows newly developed software to be tested and modified before integration into the CAMERA central server.

The visualization wall is being used to develop tools for co-displaying metabolic pathways predicted from metagenomes onto a global map. These metagenomic datasets represent nine major biomes, 100+ samples, and >17 million individual sequences. The initial goal of this project is to visualize co-occurrences of metabolisms and geochemical data from coral reefs around the world. The geochemistry data is mostly derived from remote sensing platforms (>100 Terabytes) and is stored at the SDSU visualization center and other datacenters. Soon, these datasets will be supplemented with more than 200 hours of high-definition video from coral reefs. When complete, a user will be able to ''swim'' through coral reefs and see the pertinent metagenomic and geochemical data. The advantage of the display wall is the ability to see the details of these extremely large datasets in context of the ''larger'' picture.

The Rohwer lab members range from undergraduates to post-docs and mathematicians to biologists. Together, they envision using the visualization wall in the same way as a communal white board, with all of the advantages of real time mashups of the terabyte datasets. Future plans invoke the use of the the OptIPortal, tied over a 1 or 10 Gbps optical link to the CAMERA server at Calit2, to examine 300+ sample metagenomic datasets from all over the world. The high I/O of the CAMERA servers coupled to the large pixel count in the OptIPortal will enable the Rohwer Lab to analyze this data in ways not possible on desktop visualization systems. A single GigE optical fiber connection was established between Calit2 and SDSU in June 2007 over CENIC fiber. The connection is distributed to the Frost CSL 102B visualization and Rohwer GMCS 429A ACE laboratories. This dedicated GigE between SDSU and Calit2, makes it possible to perform distributed processing of satellite imagery, and metagenomic data.

The OptIPuter dedicated fiber linkage connecting SDSU to Calit2 also provides a pathway to spectrally ''contextualize'' the biologic study by being able to bring in and rapidly process MODIS [31] imagery from the NASA Goddard group, which is also connected by OptIPuter to Calit2. The NASA MODIS instruments fly in formation on the Aqua and Terra satellites and acquire an image of most of the world's oceans twice a day with spectral window that can be very effectively tuned to image plankton and microbial life. Even though the pixel size of MODIS imagery is 250 m at the best, it effectively images life in the oceans, as well as numerous chemical components, because of the concentrations in the water. Spin patterns, upwelling patterns, and changes from severe weather events are all directly imageable. Linking to ground measurements of the DNA, even by going back to the historic imagery of the day

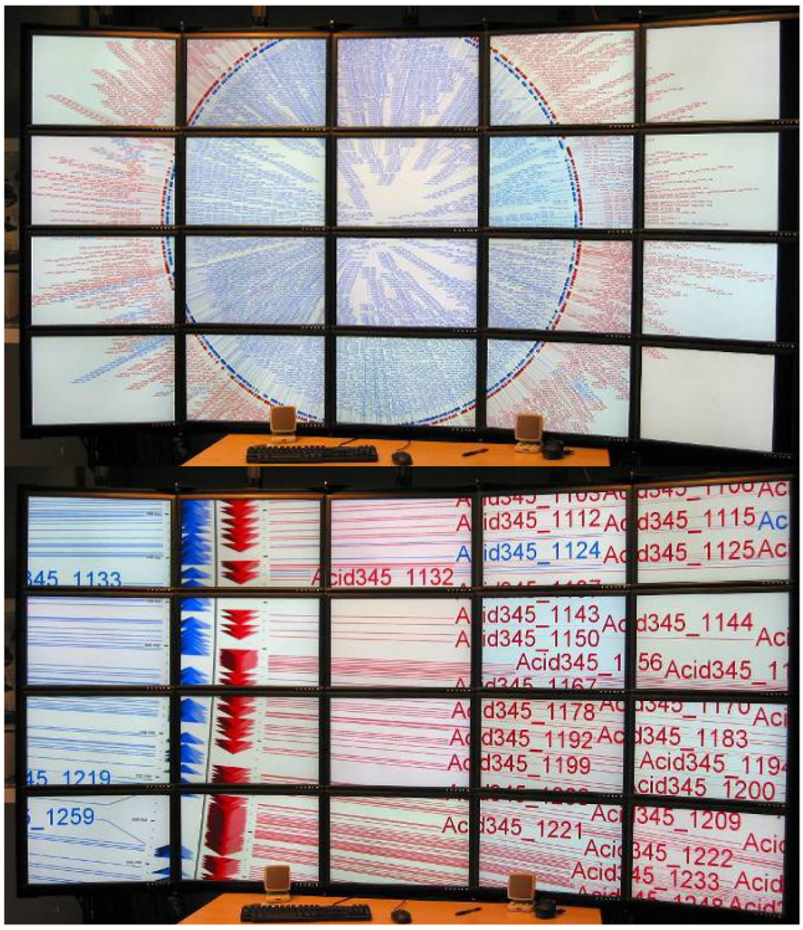


Fig. 2. Use of tiled display wall OptIPortal to interactively view a 5.6 Mb microbial genome (a soil bacterium acidobacteria bacterium Ellin345).

when past samples were collected, is providing a calibration of the ocean so that imagery can more effectively predict the biologic content, diversity, and transport within the ocean.

The biological oceanography research team of E. Virginia Armbrust at the University of Washington uses molecular approaches and combines lab-based and field-based studies to address basic questions about the function of marine ecosystems. Phytoplankton are the main focus of research in the Armbrust lab. These organisms are responsible for about 40% of the total amount of photosynthesis that occurs on our planet. They play a critical role in the global carbon cycle and ultimately in global climate. Because much of the organic carbon generated by phytoplankton is used by bacteria, the Lab also studies bacterial/phytoplankton interactions.

The Armbrust research lab was in the process of designing a lab in the new UW Benjamin D. Hall Interdisciplinary Research building during 2007, when the opportunity to install an OptIPortal arose. Development of a special visualization room with glassed walls on two sides was a central feature of the lab design. A physical wall separates the 12,800 by 4800 pixel (∼60Mpixel) tiled wall from the server room in which the OptIPortal cluster of nine computers is located. The group then worked with the campus to install fiber from the Pacific Northwest Gigapop to the Ben Hall Interdisciplinary Research Building, enabling a 10GigE link from Pacific Wave over CaveWave to Calit2's CAMERA server [32].

The first day the wall was available, the group carried out two experiments which immediately provided them with new scientific insights. First, they put up single nucleotide and sequence coverage maps of the 24 chromosomes of the marine diatom Thalassiosira pseudonanna (see Fig. 3), which have been studied extensively in the lab. They declared "for the first time we could see the the fine detail of the genomic map in the context of all 24 chromosomes which allowed us to begin to see patterns in the distributions of genomic features."

At SC07, held in Reno Nevada during November 2007, Larry Smarr stood in the University of Washington's ResearchChannel's booth talking with Professor Armbrust in Seattle. High definition teleconferencing was accomplished using ResearchChannel's iHDTV™, which streams uncompressed 1080i high-definition video, and is now being integrated into SAGE. As seen in Fig. 4, Professor Armbrust was using her new OptIPortal to compare simulations of summer and winter circulation patterns in Puget Sound. This is a first step in comparing the environmental metadata that influence the distribution patterns of different specimens of diatoms her lab has analyzed genomically. The next step is to superimpose the genomic information on the environmental data, something that will be possible to see only because of the high resolution provided by the OptIPortal.

6. The future–connecting the CAMERA OptIPlanet collaboratory

In 2008, we will complete the addition of two more laboratories across the US using OptIPortals, located at UC-Davis (Jonathan Eisen) and University of Michigan (Tom Finholt). Additional OptIPortals are being deployed internationally using local funds in Canada, Mexico, Korea, Japan, Taiwan, Australia, and the

Fig. 3. UWashington's E. Virginia Armbrust displays the 24 chromosomes of the marine diatom Thallasiosira Pseudonanna.

Fig. 4. Smarr conversing with Professor Armbrust using uncompressed iHDTV integrated with University of Washington CAMERA OptIPortal.

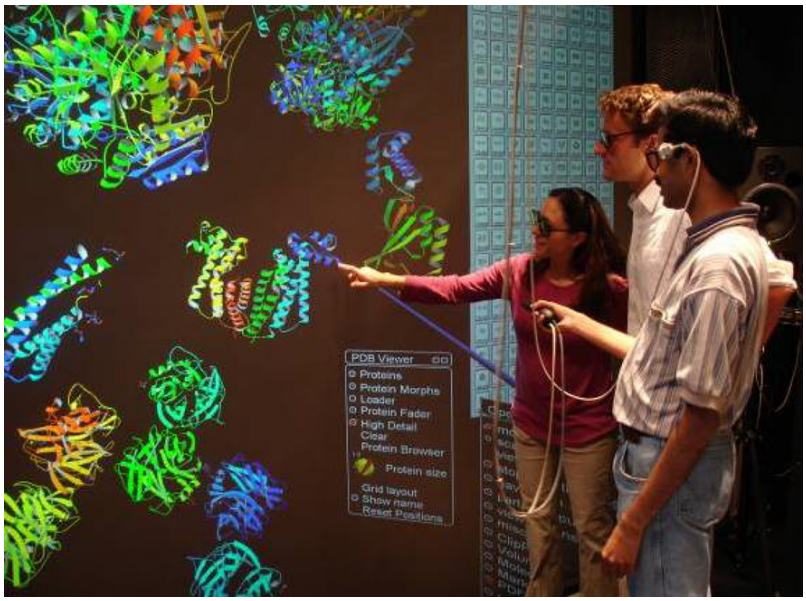


Fig. 5. Use of the StarCAVE OptIPortal to interactively visualize the proteins of the marine microbe *Thermotoga maritima.*

Netherlands to access the CAMERA server complex. As these international OptIPortals are linked up with HD flows over GLIF in 2008, an OptIPlanet Collaboratory will emerge. This lambda-enabled cyberinfrastructure will be devoted both to fundamental biological research, as well as to applications in emerging areas, such as studying the health of coral reefs under increasing environmental stress or optimizing genetically engineered microbial bio-fuel generators.

This lambda-connected collaboratory will allow the researchers to "feel" as if the two research centers, separated by thousands of miles, were next door to each other! It will be quite straightforward to build on this Calit2 prototype to create the cyberinfrastructure linking the collaborating science teams at any of the CAMERA user sites that require this higher performance capability.

Furthermore, we will expand our experiments in how to use 3D OptIPortals, such as the StarCAVE and Varrier [33] to explore the proteomics of microbial genomes. Here Calit2's affiliation with the Joint Center for Structural Genomics [34] has enabled us to carry out early evaluations of the use of the StarCAVE as a "3D web browser" (see Fig. 5) for the first 100% complete coverage of a metabolic network, which was derived using the roughly 1200 proteins that have structural coverage (over 60% of the 1859 genes) from the genome of the high temperature microbe *Thermotoga maritima*. We will continue this research in 2008 and explore using Covise to move 3D visualization capabilities to our remote walls at SDSU and UW.

In conclusion, these are early days in the adoption of OptIPuter technologies by scientific disciplines, but indications are that the great enhancement of bandwidth and pixel real estate the OptIPuter makes possible are quite useful to scientists studying the enormous complexity of living creatures in their environments.

References

[1] For a comprehensive recent review see The New Science of Metagenomics—Revealing the Secrets of Our Microbial Planet, National Research Council, National Academy Press, 2007.

[2] See www.optIPuter.net for a list of participants and publications. The authors acknowledge the National Science Foundation in providing funding for the OptIPuter project under the auspices of grant # OCI-0225642.

[3] The Gordon and Betty Moore Foundation (www.moore.org), established in September 2000, works in collaboration with grantees and other partners to achieve significant and measurable outcomes in three areas: environmental conservation, science, and the San Francisco Bay Area. In April 2004, the foundation launched its 10-year Marine Microbiology Initiative with the goal of attaining new knowledge regarding the composition, function, and ecological role of microbial communities in the world's oceans. The authors acknowledge the Gordon & Betty Moore Foundation in providing funding for the CAMERA project under the auspices of grant #951.

[4] The California Institute for Telecommunications and Information Technology (www.calit2.net), a partnership between UC San Diego and UC Irvine, houses more than 1000 researchers organized around more than 50 projects on the future of telecommunications and information technology and how these technologies will transform a range of applications important to the economy and citizens' quality of life.

[5] J. Craig Venter Institute (www.venterinstitute.org) is a not-for-profit research institute dedicated to the advancement of the science of genomics; the understanding of its implications for society; and communication of those results to the scientific community, the public, and policymakers. Founded by J. Craig Venter, Ph.D., Venter Institute is home to approximately 200 staff and scientists with expertise in human and evolutionary biology, genetics, bioinformatics/informatics, information technology, high-throughput DNA sequencing, genomic and environmental policy research, and public education in science and science policy. J. Craig Venter Institute is a 501(c)(3) organization.

[6] The Center for Earth Observations and Applications (http://ceoa.ucsd.edu) was established in November 2005 by UCSD to stimulate support and coordinate sustained research and applications in Earth observations. Led by Scripps Institution of Oceanography in partnership with Calit2 and other campus organizations, CEOA provides an integrating vision for work across the spectrum of natural, physical, and social sciences, engineering, and information technology related to Earth observations and applications. Working through CEOA is Terry Gasterland's Scripps Genome Center (see http://ucsdnews.ucsd.edu/newsrel/science/GaasterlandGenomeCenter.asp), which is producing annotations of the microbial metagenomics data.

[7] In 2005, the San Diego Supercomputer Center (SDSC; www.sdsc.edu) celebrated two decades of enabling international science and engineering discoveries through advances in computational science and high-performance computing. Continuing this legacy into the era of cyberinfrastructure, SDSC is a strategic resource to academia and industry, providing leadership in data cyberinfrastructure, particularly with respect to data curation, management, and preservation, data-oriented high-performance computing, and cyberinfrastructure-enabled science and engineering. SDSC is an organized research unit of the University of California, San Diego, and one of the founding sites of NSF's TeraGrid.

[8] For a history of Woese's microbial revolution, see Science 276 (May) (1997) 699–702. www.sciencemag.org/cgi/content/full/276/5313/699, ai.arc.nasa.gov/news_stories/news_detail.cfm?ID=274, or http://en.wikipedia.org/wiki/Carl_Woese.

[9] 16S rRNA is found in the small ribosomal subunit of microbial ribosomes and the mitochondria and chloroplasts ribosomes of eukaryotes. Since ribosomes are essential for DNA transcription to form proteins, most ribosomal RNA mutations are deleterious, resulting in a very slow evolution of 16S rRNA. This makes it a very good molecule to use to compare organisms that may have diverged as far back as 3 or 4 billion years ago. For more on 16S rRNA sequencing, see www.microbeworld.org/htm/aboutmicro/tools/genetic.htm.

[10] A nice diagram of the tree is at http://genome.jgi-psf.org/tre_home.html or http://en.wikipedia.org/wiki/Image:PhylogeneticTree.jpg. For more on the biology of creatures in the Tree of Life, See http://tolweb.org/tree.

[11] For the story of the whole of life's evolution, see Richard Fortey, Life: A Natural History of the First Four Billion Years of Life on Earth, Alfred A. Knopf, NY, 1998; For a very readable account of the Cambrian explosion of multi-cellular body plans, see Stephen Jay Gould, Wonderful Life: The Burgess Shale and the Nature of History, W.W. Norton and Co., 1989.

[12] Norman Pace, The large-scale structure of the tree of life, in: Jan Sapp (Ed.), Microbial Phylogeny and Evolution: Concepts and Controversies, Oxford Univ. Press, 2005.

[13] J. Craig Venter, et al., Environmental genome shotgun sequencing of the Sargasso Sea, Science 304 (April) (2004) 66;
Paul Falkowski, Colomban de Vargas, Shotgun sequencing in the sea: A blast from the past?, Science 304 (April) (2004) 58. (and commentary) See also news story linking to Sorcerer II expedition at www.genomenewsnetwork.org/articles/2004/03/04/sargasso.php

[14] See for more details http://en.wikipedia.org/wiki/Shotgun_sequencing.

[15] See www.ncbi.nih.gov. As of August 2008, GenBank held more than 100 billion bases from more than 165,000 species.

[16] See, www.sorcerer2expedition.org.

[17] For a recent review, see Kevin Chen, Lior Pachter, Bioinformatics for whole-genome shotgun sequencing of microbial communities, PLoS Computational Biology: http://compbiol.plosjournals.org/perlserv/?request=get-document%26doi=10.1371/journal.pcbi.0010024 or;
Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, Jay M. Short, Eric J. Mathur, John C. Detter, Peer Bork, Philip Hugenholtz, Edward M. Rubin, Comparative Metagenomics of Microbial Communities, Science 308 (April) (2005) 554.

[18] Science 307 (March) (2005) 1558.

[19] See, www.syntheticgenomics.com.

[20] S. Look, et al., PNAS 83 (1986) 6283.

[21] http://news-service.stanford.edu/news/2006/may24/criddle-052406.html.

[22] See, www.whoi.edu/mr/pr.do?id=919.

[23] www.nlr.net.

[24] www.glif.is.

[25] The usable server room was doubled to 2000sf, we added a 500kva transformer and doubled to twelve the 225A breakers, added a StarLine modular power grid power distribution system, and added 66 tons of cooling, ducts and registers.

[26] Rocks Core Development is sponsored by the NSF as part of award #OCI-438741. www.rocksclusters.org.

[27] See, for instance, The OptIPuter, Quartzite, and Starlight Projects: A Campus to Global-Scale Testbed for Optical Technologies Enabling LambdaGrid Computing: http://www.optiputer.net/publications/articles/Smarr-OFC-2005.pdf.

[28] www.teragrid.org.

[29] See accompanying FGCS article on OptIPortals.

[30] The tiled wall uses 3 × 5 16 Dell UltraSharp 2407FP Wide Flat Panels, each with a native resolution of 1920 × 1200 pixels. The Linux cluster supporting the OptIPortals has 8 Dell Dimension 9200 with Dual Intel E6700 2.66 GHz processors and Dual EVGA GeForce 7950 GX2 GPUs, together with 4GB RAM with a 160GB Hard Drive per PC.

[31] The Moderate Resolution Imaging Spectro-radiometer (MODIS) instrument is on the NASA Aqua Earth-observing satellite. Goddard Space Flight Center hosts the MODIS Data Processing System (https://modaps.nascom.nasa.gov:8499).

[32] These network connections serve 3 clusters: (1) 32-node compute cluster. Each compute node is dual-socket quad-core Xeon E5345 2.33 GHz with 16 GB RAM, the whole system is 256 cores. Interconnect is GigE and infiniband. (2) 4-node cluster for testing code and administrative changes before they get put on the 32-node cluster. Each compute node is a dual-socket dual-core AMD 2218 2.66 GHz, 4 GB RAM. (3) Visualization cluster. 8 viz nodes are dual-socket dual-core AMD 2216 2.4 GHz, 8 GB RAM, Nvidia 8800GTX card. Interconnect is GigE. There are fifteen 30" monitors in the 5 row 3 column display wall.

[33] See accompanying FGCS article on StarCAVE and Varrier.

[34] The Joint Center for Structural Genomics (GM062411-05) is funded by the National Institute for General Medical Sciences, as part of the Protein Structure Initiative of the National Institutes of Health. It is a multi-

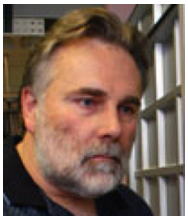
institutional consortium with major activities at the Scripps Research Institution; the Genomics Institute of the Novartis Research Foundation; the University of California, San Diego; and the Stanford Synchrotron Radiation Laboratory at Stanford University. See www.jcsg.org.

Larry Smarr became founding director in 2000 of the California Institute for Telecommunications and Information Technology (Calit2), a UCSD/UCI partnership. He is the Harry E. Gruber professor in the Jacobs School's Department of Computer Science and Engineering at UCSD. For the previous 15 years as director of the National Center for Supercomputing Applications and the National Computational Science Alliance, Smarr helped drive major developments in the planetary information infrastructure: the Internet, the Web, scientific visualization, virtual reality, and global telepresence. He was a member of the President's Information Technology Advisory Committee for President Clinton and served until 2005 on the Advisory Committee to the Director of the National Institutes of Health and the NASA Advisory Council. He was a member of the California Governor's Task Force on Broadband in 2007. He is a member of the National Academy of Engineering and is a Fellow of the American Physical Society and the American Academy of Arts and Sciences. In 2006 he was presented with the ESRI Lifetime Achievement Award and received the IEEE Computer Society Tsutomu Kanai Award for distributed computing systems achievements.

Paul Gilna is the former director of the Department of Energy's Joint Genome Institute (JGI) operations at Los Alamos National Laboratory (LANL) and group leader of Genomic Science and Computational Biology in LANL's Bioscience Division. Dr Gilna has worked with researchers in such diverse disciplines as cell biology, chemistry, computational and computer science, environmental science, genomic and proteomic science, measurement science, nanobiology, and space and theoretical physics. As CAMERA Executive Director, Gilna coordinates this complex project, including overseeing building the needed cyberinfrastructure, fostering partnerships within the emerging metagenomics scientific community and working with all stakeholders, including project staff, key users of the infrastructure, and advisors to the project.

Phil Papadopoulos recieved his Ph.D. in electrical engineering from UC Santa Barbara in 1993. He spent 5 years at Oak Ridge National Laboratory as part of the the Parallel Virtual Machine (PVM) development team. He is currently the Program Director of Grid and Cluster Computing at the San Diego Supercomputer Center. Dr Papadopoulos is deeply involved in key research projects including the Biomedical Informatics Research Network (BIRN), OptIPuter, the Geosciences Network (GEON), the NSF Middleware Initiative (NMI), The National Biomedical Computation Resource (NBCR), and the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA). He is also well known for the development of the open source Rocks Cluster toolkit, which has an installed base of 1000s of clusters. Dr Papadopoulos is a Co-PI of the CAMERA Project.

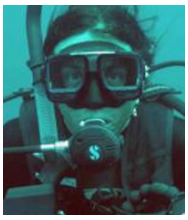
Thomas A. DeFanti at the University of California, San Diego, is a research scientist at the California Institute for Telecommunications and Information Technology (Calit2). At the University of Illinois at Chicago, DeFanti is director of the Electronic Visualization Laboratory (EVL), and a distinguished professor emeritus in the department of Computer Science. He has researched computer graphics since the early 1970s. His credits include: use of EVL hardware and software for the computer animation produced for the 1977 "Star Wars" movie; contributor and co-editor of the 1987 NSF-sponsored report "Visualization in Scientific Computing"; recipient of the 1988 ACM Outstanding Contribution Award; he became an ACM Fellow in 1994.



Greg Hidley received his Ph.D. from the University of California, San Diego, and is a technical lead for a number of Cyberinfrastructure projects including CAMERA (http://camera.calit2.net) and OptIPuter (http://www.optiputer.net). He was the first CIO of the UCSD Division of CALIT2 (http://www.calit.net) and has participated in campus IT infrastructure planning and implementation for UCSD and for the UC system for the past 25 years.

John Wooley is Associate Vice Chancellor for Research at the University of California San Diego, and an adjunct Professor of Pharmacology, Chemistry and Biochemistry. Dr Wooley also serves on The University of Chicago Board of Governors and the Federal Advisory Committee for the US DOE Office of Biological and Environmental Research (BERAC). He has previously held faculty appointments at Princeton University, the Marine Biological Labs and Searle Pharmaceuticals. At Calit2, Dr Wooley created and now directs the biology and biomedical layer or applications component, termed Digitally-enabled Genomic Medicine (DeGeM). His current research involves bioinformatics and structural biology, and he is co-Principle Investigator of the NIH-funded Joint Center for Structural Genomics.

E. Virginia Armbrust received her Ph.D from the Massachusetts Institute of Technology and Woods Hole Oceanographic Institution in 1990 and her B.A. in Human Biology from Stanford University in 1980. She currently serves as a Professor at the School of Oceanography at the University of Washington. The primary focus of her group's research is on phytoplankton. These organisms are responsible for about 40% of the total amount of photosynthesis that occurs on our planet. They play a critical role in the global carbon cycle and ultimately in global climate. Because much of the organic carbon generated by phytoplankton is used by bacteria, her lab also studies bacterial/phytoplankton interactions.



Forest Rohwer holds B.A.s in Biology, Chemistry, and History from the College of Idaho and a Ph.D in Molecular Biology from the Univerity of Californa, San Diego jointly with San Diego State University. The Rohwer group focuses their research on the dramatic increases in incidences of coral disease over the last two decades and which are believed to be instrumental in the global decline of coral reefs. Dr Rohwer's research is directed at the belief that many of these diseases are actually opportunistic infections caused by anthropogenic stressors.

Eric Frost received his Ph.D. from San Diego State University, and is currently Co-Director of the Visualization Center and also of the Center for Information Technology and Infrastructure that works very closely with Calit2, including across dedicated fiber linking servers, archives, imaging, and visualization infrastructure. Frost uses many of these tools for natural hazards response for humanitarian assistance disaster relief and also Homeland Security. He is also Co-Director of the Homeland Security Master's Program and the Center for Homeland Security Technology Assessment, which use the visualization and data fusion tools for public safety and international collaboration providing rapid imagery processing via fiber and delivery to the field, as the CAMERA project does.