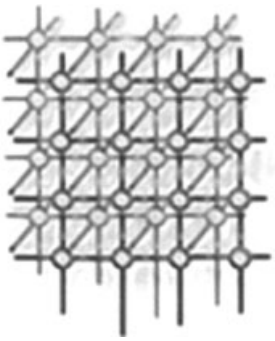
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2010; **22**:2266–2281
Published online 23 April 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.1590



Global-scale distributed I/O with ParaMEDIC

P. Balaji[1,*,†], W. Feng [2], H. Lin[2], J. Archuleta[2], S. Matsuoka[3], A. Warren[4], J. Setubal[4], Ewing Lusk[1], R. Thakur[1], I. Foster[1,5], D. S. Katz[1,5,6], S. Jha[6], K. Shinpaugh[2], S. Coghlan[1] and D. Reed[7]

[1]*Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, U.S.A.*
[2]*Department of Computer Science, Virginia Tech, Blacksburg, VA, U.S.A.*
[3]*Department of Mathematical and Computing Science, Tokyo Institute of Technology, Tokyo, Japan*
[4]*Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, U.S.A.*
[5]*Computation Institute, University of Chicago, Chicago, IL, U.S.A.*
[6]*Center for Computation and Technology, Louisiana State University, Baton Rogue, LA, U.S.A.*
[7]*Scalable Computing and Multicore Division, Extreme Computing Group, Microsoft Research, Redmond, WA, U.S.A.*

SUMMARY

Achieving high performance for distributed I/O on a wide-area network continues to be an elusive holy grail. Despite enhancements in network hardware as well as software stacks, achieving high-performance remains a challenge. In this paper, our worldwide team took a completely new and non-traditional approach to distributed I/O, called *ParaMEDIC*: *Parallel Metadata Environment for Distributed I/O and Computing*, by utilizing application-specific *transformation* of data to orders of magnitude smaller metadata before performing the actual I/O. Specifically, this paper details our experiences in deploying a large-scale system to facilitate the discovery of missing genes and constructing a genome similarity tree by encapsulating the mpiBLAST sequence-search algorithm into ParaMEDIC. The overall project involved nine computational sites spread across the U.S. and generated more than a petabyte of data that was 'teleported' to a large-scale facility in Tokyo for storage. Copyright © 2010 John Wiley & Sons, Ltd.

Received 11 February 2010; Revised 21 February 2010; Accepted 24 February 2010

KEY WORDS: distributed I/O; Bioinformatics; BLAST; grid computing; cluster computing

*Correspondence to: P. Balaji, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, U.S.A.
†E-mail: balaji@mcs.anl.gov

1. INTRODUCTION

With the rapid growth in the scale and complexity of scientific applications over the past few decades, the requirements for compute, memory, and storage resources are now greater than ever before. With the onset of petascale and exascale computing, issues related to managing such grand-scale resources, particularly related to data I/O, need to be carefully studied. For example, applications including genomic sequence search and the emergent field of metagenomics, large-scale data mining, data visual analytics, and communication profiling on ultrascale parallel computing platforms generate massive amounts of data that needs to be managed for later processing or archival.

Adding to the complexity of this problem is the issue of resource locality. While system sizes have certainly grown over the past few years, most researchers do not have local access to systems of the scale required by their applications. Therefore, researchers access such large systems remotely to perform the required computations and move the generated data to their local systems after the computation is complete. Similarly, many applications tend to require multiple resources simultaneously for efficient execution. For example, applications that perform large computations and generate massive amounts of output data are becoming increasingly common. While several large-scale supercomputers provide either the required compute power or the storage resources, very few provide both. Thus, data generated at one site often has to be moved to a different site for storage and/or analysis.

In order to alleviate the issues related to moving such massive data across sites, considerable monetary and intellectual investments have been put into high-speed distributed network connectivity [1–3]. However, the utility of these investments is limited in the light of three primary observations: (1) such infrastructure is scarce and does not provide end-to-end connectivity to a very high percentage of the scientific community, (2) the amount of data generated by many applications is so large that even at 100% network efficiency, the I/O time for these applications can significantly dominate their overall execution time, and (3) based on recent trends and published results, existing distributed I/O mechanisms have not been able to achieve a very high network utilization for 'real data' on high-speed distributed networks, particularly for single-stream data transfers [4,5].

To resolve such issues on a *global* scale, we proposed a new, non-traditional approach for distributed I/O known as *ParaMEDIC* (Parallel Metadata Environment for Distributed I/O and Computing) [6–8]. ParaMEDIC uses application-specific semantic information to process the data generated by treating it as a collection of high-level abstract objects, rather than as a generic byte-stream. It uses such information to transform the data into orders of magnitude smaller metadata before transporting it over the distributed environment and regenerating it at the target site. All data transformation, movement, and regeneration are done while the application is executing, giving the illusion of an ultrafast teleportation device for large-scale data over distributed environments.

At a high level, ParaMEDIC is similar to standard compression algorithms. However, the term 'compression' typically has the connotation that the data is processed as a generic byte-stream. As ParaMEDIC uses a more abstract application-specific representation of the data to achieve a *much* larger reduction in the data size, we use the terminology of 'metadata transformation' in this case.

Because ParaMEDIC utilizes application semantics to generate metadata, it loses some portability compared to traditional byte-stream-based distributed I/O. For example, an instance of

DOI: 10.1002/cpe

ParaMEDIC's metadata transformation in the context of the mpiBLAST sequence search application is described in Section 3.2. By giving up some portability, however, ParaMEDIC can potentially attain tremendous savings in the amount of actual distributed I/O performed, consequently resulting in substantial performance gains. Further, through the use of a generic framework with an application plug-in model, different applications can use the overall framework in an easy and flexible manner.

In this paper, we demonstrate how we used ParaMEDIC to tackle two large-scale computational biology problems—discovering missing genes and adding structure to genetic sequence databases—on a worldwide supercomputer [9]. The overall worldwide supercomputer comprised nine different supercomputers distributed at seven sites across the U.S. and one large-scale storage facility located in Japan. The overall experiment consisted of sequence searching the entire microbial genome database against itself, generating approximately a petabyte of data that was transported to Tokyo for storage. We present several insights gained from this large-scale run, which will be valuable to other researchers performing such large, global-scale distributed computation and I/O.

2. LARGE-SCALE COMPUTATIONAL BIOLOGY: A PEEK AT COMPUTE AND STORAGE REQUIREMENTS

In this section we discuss different aspects of computational biology with a focus on the compute and storage requirements of large-scale applications in this domain.

2.1. Sequence searching

With the advent of rapid DNA sequencing, the amount of genetic sequence data available to researchers has increased exponentially [10]. The GenBank database, a comprehensive database that contains genetic sequence data for more than 2 60 000 named organisms, has exhibited exponential growth since its inception over 25 years ago [11]. This information is available for researchers to search new sequences against and infer homologous relationships between sequences or organisms. This helps a wide range of projects, from assembling the Tree of Life [12] to pathogen detection [13] and metagenomics [14].

Unfortunately, the exponential growth of sequence databases necessitates faster search algorithms to sustain reasonable search times. The Basic Local Alignment Search Tool (BLAST), the *de facto* standard for sequence searching, uses heuristics to prune the search space and decrease search time with an accepted loss in accuracy [15,16]. mpiBLAST parallelizes BLAST using several techniques including database fragmentation, query segmentation [17], parallel input–output [18], and advanced scheduling [19]. As shown in Figure 1, mpiBLAST uses a *master–worker* model and performs a *scatter–search–gather–output* execution flow. During the scatter, the master splits the database and query into multiple pieces and distributes them among worker nodes. Each worker then searches the query segment against the database fragment that it was assigned. The results are gathered by the master, formatted, and output to the user. Depending on the size of the query and that of the database, such output generated can be large. For example, as shown in Table I, an all-to-all search of the nucleotide database can generate as much as 30 TB of data. Thus, for environments with limited I/O capabilities, such as distributed systems, the output step can cause significant overheads.

DOI: 10.1002/cpe

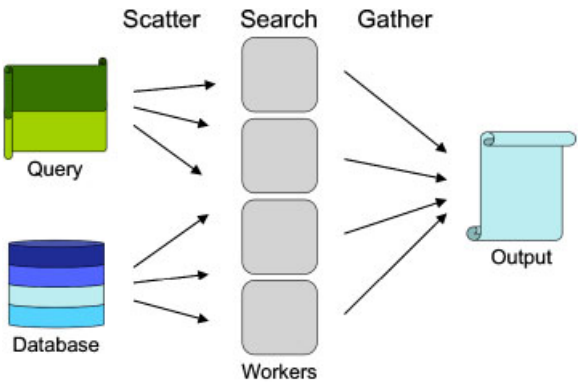


Figure 1. High-level algorithm of mpiBLAST.

Table I. Estimated output of an all-to-all NT search.

Query size (KB)	Number of queries	Estimated output (GB)
0–5	3 305 170	1139
5–50	87 506	593
50–150	25 960	23 555
150–200	26 524	3995
>200	9 840	*Not run*
Total	3 455 000	>29 282

2.2. Discovering missing genes

Genome annotation is the process of associating information with a genomic sequence. Part of this process includes determining (by computational analysis) the location and structure of protein-encoding and RNA-encoding genes, also known as making gene calls. It is important to be as accurate and as sensitive as possible in making gene calls: avoiding false positives and missing real genes. Gene prediction in prokaryotes (bacteria and archaea) typically involves evaluating the coding potential of genomic segments that are delimited by conserved nucleotide motifs. The most widely used gene-finding programs [20,21] build a sequence model based on statistical properties of genes known to be (very likely) real genes in a given genome. This model is then used to evaluate the likelihood that an individual segment codes for a gene; using this method, some genes with anomalous composition are almost always missed. Another popular method for locating genes is to compare genomic segments with a database of gene sequences found in similar organisms. If the sequence is conserved, the segment being evaluated is likely to be a coding gene (this is the 'similarity method'). Genes that do not fit a given genomic pattern and do not have similar sequences in current annotation databases may be systemically missed.

One way to detect missed genes is to use the similarity method and compare raw genomes against each other, as opposed to comparing a raw genome to a database of known genes [22]. If gene a in genome A and gene b in genome B have been missed and a is similar to b, then this method will find both. However, this involves performing an all-to-all comparison of the entire database

DOI: 10.1002/cpe

against itself (in our case study, the entire microbial genome database against itself). This task is heavily compute and data intensive, requiring thousands of compute processors and generating on the order of a petabyte of output data that needs to be stored for processing.

2.3. Adding structure to genetic sequence databases

One of the major issues with sequence search is the structure of the sequence database itself. Currently, these databases are unstructured and stored as a flat file, and each new sequence that is discovered is simply appended to the end of the file. Without more intelligent structuring, a query sequence has to be compared to every sequence in the database, forcing the best case to take just as long as the worst case. By organizing and providing structure to the database, searches can be performed more efficiently by discarding irrelevant portions entirely.

One way to structure the sequence database is to create a sequence similarity tree where 'similar' sequences are closer together in the tree than dissimilar sequences. The connections in the tree are created by determining how 'similar' the sequences are to each other through sequence searches. To create *every* connection, however, the entire database has to be searched against itself, resulting in an output size of N^2 values (where N is the number of sequences in the database).

3. OVERVIEW OF PARAMEDIC-ENHANCED MPIBLAST

In our previous work [7,8], we provided a detailed description of ParaMEDIC. Here we present a brief summary of that work.

3.1. The ParaMEDIC framework

ParaMEDIC provides a framework for *decoupling* computation and I/O in applications relying on large quantities of both. Specifically, it does not hinder application computation. As the output data is generated, however, the framework differs from traditional distributed I/O in that it uses application-semantic information to process the data generated by treating it as a collection of high-level application-specific objects rather than as a generic byte-stream. It uses such information to transform the data into orders of magnitude smaller metadata before transporting it over the distributed environment and regenerating the original data at the target site.

As shown in Figure 2, ParaMEDIC provides several capabilities, including support for data encryption and integrity as well as data transfer in distributed environments (either directly via Transmission Control Protocol/Internet Protocol (TCP/IP) communication or through global file-systems). However, the primary semantics-based metadata creation is done by the *application plug-ins*. Most application plug-ins are specific to each application and thus rely on knowledge of application semantics. These plug-ins provide two functionalities: (1) processing output data generated by the application to create metadata and (2) converting metadata back to the final output. Together with application-specific plug-ins, ParaMEDIC also provides application-independent components, such as data compression, data integrity, and data encryption. These can be used in conjunction with the application-specific plug-ins or independently.

DOI: 10.1002/cpe

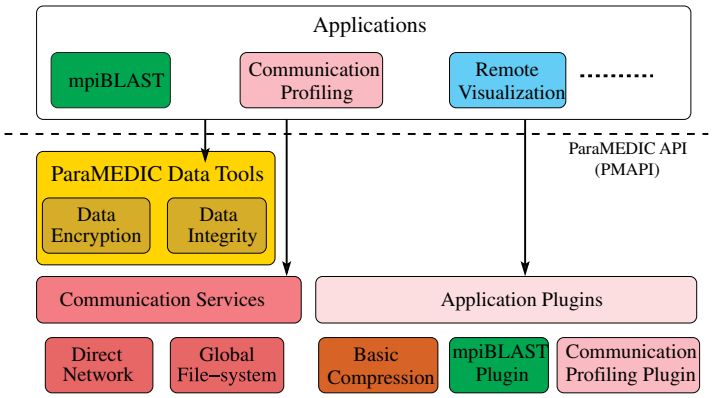


Figure 2. ParaMEDIC architecture.

Trading Computation with I/O: The amount of computation required in ParaMEDIC is higher than what is required by the original application. After the output is generated by the application processes, it has to be further processed to generate the metadata, sent to the storage site, and processed yet again to regenerate the final output. However, the I/O cost achieved can potentially be significantly reduced by using this framework. In other words, ParaMEDIC trades (a small amount of) additional computation for (potentially large) reduction in I/O cost. With respect to the additional computational cost incurred, ParaMEDIC is quite generic with respect to the metadata processing required by the different processes. For many applications, it is possible to tune the amount of post-processing performed on the output data, with the trend being, the more the post-processing computation, the better the reduction in the metadata size. That is, an application plug-in can perform more processing of the output data to reduce the I/O cost.

3.2. Integration with mpiBLAST

In a *cluster* environment, most of the mpiBLAST execution time is spent on the search itself, since the BLAST string-matching algorithm is computationally intensive. In comparison, the cost of formatting and writing the results is minimal, especially when many advanced clusters are configured with high-performance parallel file-systems. In a *distributed* environment, however, the output typically needs to be written over a wide-area network to a remote file-system. Hence, the cost of writing the results can easily dominate the execution profile of mpiBLAST and become a severe performance bottleneck. By replacing the traditional distributed I/O framework with ParaMEDIC (as shown at the top of Figure 3), we can provide a large reduction in the amount of data communication performed. For example, as we will see in Section 5, an mpiBLAST-specific instance of ParaMEDIC reduces the volume of data written across a wide-area network by *more than 2 orders of magnitude*.

Figure 3 depicts how mpiBLAST is integrated with ParaMEDIC. First, on the compute site (the left cloud in Figure 3), once the output is generated by mpiBLAST, the mpiBLAST application plug-in for ParaMEDIC processes this output to generate orders of magnitude lesser metadata. Specifically, the output of mpiBLAST consists of alignment information and scores corresponding

DOI: 10.1002/cpe

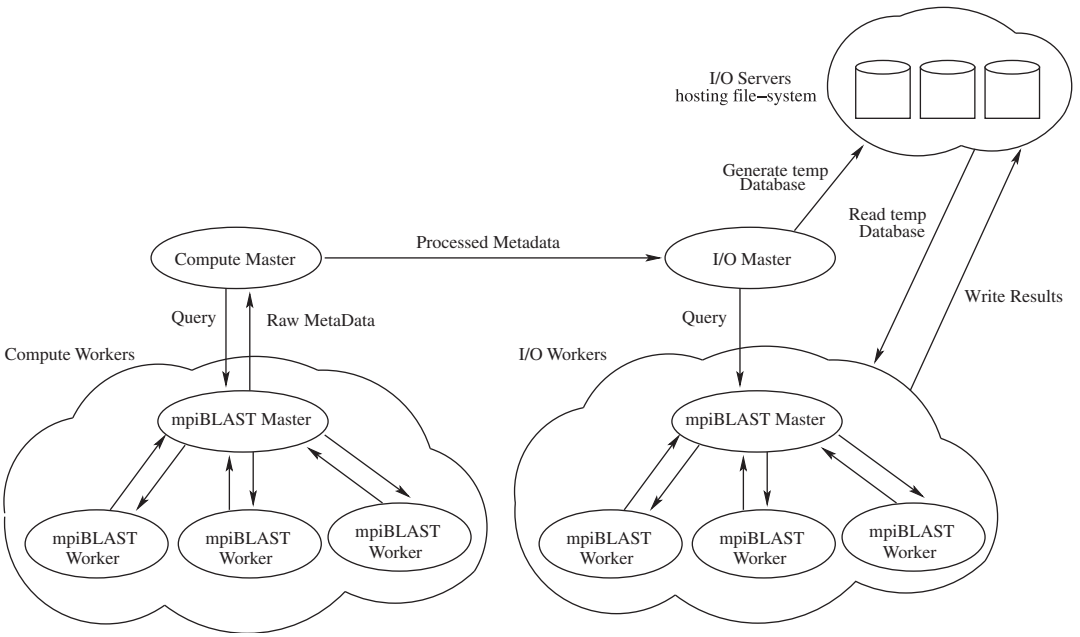


Figure 3. ParaMEDIC and mpiBLAST integration.

to the top matches it found for each sequence in the entire database. Thus, while the search time largely depends on the size of the database, once the search is complete, the output only depends on how closely the input query sequence matches the top matching sequences in the database. Based on this observation, the metadata basically contains information identifying the top matching sequences in the database, and not the other alignment or score information.

ParaMEDIC transfers this metadata to the storage site. At the storage site, a temporary (and much smaller) database that contains only the top matching sequences is created by extracting the corresponding sequence data from a local database replica. ParaMEDIC then reruns mpiBLAST at the storage site by taking as input the same query sequence and the temporary database to generate and write the complete output (including the alignments and scores) to the storage system. The overhead in rerunning mpiBLAST at the storage site is small, since the temporary database that is searched is substantially smaller, with only about 250 sequences in it, compared to the several millions of sequences in large DNA databases.

4. PARAMEDIC ON A WORLDWIDE SUPERCOMPUTER

To accommodate the compute and storage requirements of the computational biology applications discussed in Section 2, we utilize a worldwide supercomputer that, in aggregate, provides the required compute power and storage resources. The worldwide supercomputer comprises nine high-performance computing systems at seven different sites across the U.S. and a large-scale storage facility in Japan, to create a single high-performance *distributed* computing system. The specifics of

DOI: 10.1002/cpe

each individual system are in Table II. In the following sections, we address the issues with working on such a large-scale distributed system that are not immediately apparent on smaller-scale systems.

4.1. Dealing with dynamic availability of compute clients and other faults

Several systems in our worldwide supercomputer operate in batch mode. Users submit jobs to system queues and are scheduled to execute on the available resources. That is, compute resources are not available in a dedicated manner but become available when our job is scheduled for execution and become unavailable when our job terminates.

To handle this issue, we segment the overall work to be performed into small tasks that can be performed independently (i.e. sequentially, concurrently, or out of order). The management of tasks is done by a centralized server running on a dedicated resource. As each job is executed, it contacts this server for the next task, computes the task, transforms the output to metadata, and transmits the metadata to the storage site. This approach has two benefits. First, the clients are completely stateless. That is, if a client job terminates before it has finished its computation or metadata transmission to the storage site, the servers handle this failure by reassigning the task to a different compute client. The second advantage is if the metadata corresponds to a task that is either not received completely or is corrupted, the server just discards the data and reassigns the task to another compute node. Thus, I/O failures are never catastrophic.

4.2. Architectural heterogeneity

One of the key impediments to large-scale distributed systems is system heterogeneity. Many distributed systems, such as the one used in this paper, cannot obtain a homogeneous environment in either hardware or software, and efficient use of the system requires overcoming this obstacle. The worldwide supercomputer used in this paper contains six different processor architectures (IBM PowerPC 970FX, IBM PowerPC 440, AMD Opteron, SiCortex MIPS64, Intel Xeon, and Intel Itanium2), five different network interconnects (Gigabit Ethernet, 10-Gigabit Ethernet, InfiniBand, IBM proprietary 3D toroidal network, and SiCortex Kautz graph), and eight variations of the Linux operating system.

In order to deal with this issue, all data being transferred over the network has to be converted to an architecture-independent format. As the total amount of data that is generated and must be moved to the storage site is enormous, this can have a significant impact on traditional distributed I/O. However, with ParaMEDIC, only metadata generated by processing the actual output is transferred across the wire. As this metadata is orders of magnitude smaller as compared to the actual output, such byte manipulation to deal with heterogeneity has minimal impact on the overall performance.

4.3. Utilizing the parallelism in compute nodes

In traditional file I/O, there are two levels of parallelism. First, multiple I/O servers are aggregated into a parallel file-system to take advantage of the aggregate storage bandwidth of these servers. Second, multiple compute clients, that process different tasks, write data to such file-systems in parallel as well. Most parallel file-systems are optimized for such access to give the best performance.

DOI: 10.1002/cpe

Table II. Specification of the different systems that formed our worldwide supercomputer.

Name	Location	Cores	Architecture	Memory (GB)	Network	Storage (TB)	Distance from storage (km)
SystemX Compute Cluster	Virginia Tech	2200	IBM PowerPC 970FX	4	InfiniBand (8 Gbps)	NFS (30)	11 000
Breadboard Compute Cluster	Argonne National Lab	128	AMD Opteron	4	10G Ethernet (10 Gbps)	NFS (5)	10 000
BlueGene/L Supercomputer	Argonne National Lab	2048	IBM PowerPC 440	1	Proprietary (1.4 Gbps)	PVFS (14)	10 000
SiCortex Supercomputer	Argonne National Lab	5832	SiCortex SC5832	3	Proprietary (11.2 Gbps)	NFS (4)	10 000
Jazz Compute Cluster	Argonne National Lab	700	Intel Xeon	1–2	1G Ethernet (1 Gbps)	GFS (10) PVFS (10)	10 000
NSF TeraGrid Cluster	University of Chicago	320	Intel Itanium2	4	Myrinet 2000 (2 Gbps)	NFS (4)	10 000
NSF TeraGrid Cluster	San Diego Supercomputer Center	60	Intel Itanium2	4	Myrinet 2000 (2 Gbps)	GPFS (50)	9000
Oliver Compute Cluster	Louisiana State University, Lafayette	512	Intel Xeon	4	InfiniBand (8 Gbps)	Lustre (12)	11 000
Open Science Grid	United States	200	AMD Opteron Intel Xeon	1–2	1G Ethernet (1 Gbps)	—	11 000
TSUBAME Storage Cluster	Tokyo Institute of Technology	72	AMD Opteron	16	1G Ethernet (1 Gbps)	Lustre (350)	0

DOI: 10.1002/cpe

With ParaMEDIC, there are three I/O components: (1) compute clients that perform I/O, (2) post-processing servers that process the metadata to regenerate the original output, and (3) I/O servers that host the file-system. Similar to the traditional I/O model, the first and third components are already parallelized. That is, multiple streams of data being written in parallel by different compute clients and the I/O servers parallelize each stream of data that is being written to them. However, in order to achieve the best performance, it is important that the second component, post-processing servers, be parallelized as well.

Parallelizing the post-processing servers adds its own overhead and complexity mainly with respect to synchronization between the different parallel processes. To avoid this, we use an embarrassingly parallel approach for these servers. Each incoming stream of data is allocated to a separate process till a maximum number of processes is reached, after which the incoming data requests are queued till a process becomes available again. Thus, different processes do not have to share any information and can proceed independently. The advantage of this approach is its simplicity and the lack of synchronization required between different parallel post-processing servers. The disadvantage, however, is that the number of data streams generated from the post-processing servers is equal to the number of incoming data streams. That is, if only two tasks are active at one time, only two streams of data are written to the actual storage system. Thus, the performance might not be optimal. However, in most cases, we expect the number of incoming streams to be sufficiently large so as to not face such performance issues.

4.4. Handling large communication latencies with disconnected I/O

As seen in Table II, the computational sites are between 9000 and 11 000 km away from the storage site. At these distance, communication latencies are in tens of milliseconds. Such large latencies can severely limit the effectiveness of a synchronous remote file-systems that can be used for distributed I/O, since each synchronization operation has to make round-trip hops on the network. To overcome the bottleneck incurred by such high-latency, our worldwide supercomputer utilizes a lazy asynchronous I/O approach. By caching the output data locally before performing the actual output, clients can perform their computations while disconnected from the remote file-system. After a substantial amount of data is generated a bulk transfer of the metadata occurs, thereby maximally utilizing the bandwidth available between the sites and mitigating the effect of high latency.

An issue with this approach of disconnected computation is fault tolerance. Once a task is assigned to a compute client, the server is completely disconnected from this client. After the computation is complete, the client reconnects and sends the generated metadata. Although, this two-phase synchronization model is more error prone and requires additional state information in the server, it makes the compute clients *truly stateless* even when the actual computation is going on.

5. EXPERIMENTS, MEASUREMENTS, AND ANALYSIS

In this paper, we use ParaMEDIC to search the entire microbial genome database against itself. Several supercomputing centers within the U.S. perform the computation, while the data generated is stored at a large storage resource in Tokyo. This section compares the performance of ParaMEDIC

DOI: 10.1002/cpe

with vanilla mpiBLAST with respect to the amount of data communicated and the I/O time taken, as well as the storage bandwidth utilization.

5.1. Data I/O overheads

Figure 4(left) illustrates the amount of data transmitted between the compute and the storage sites for different number of post-processing threads, and Figure 4(right) shows the factor of reduction in the amount of data. Each post-processing thread processes one segment of data that has the output for 10 000 query sequences. Most segments have approximately similar output sizes, hence the amount of data communicated over the distributed network increases linearly with the number of segments, and hence the number of post-processing threads. ParaMEDIC, on the other hand, processes the generated data and converts it into metadata before performing the actual transfer. Thus, the actual data that is transferred over the network is much smaller. For example, with 288 threads, mpiBLAST communicates about 100 GB of data, whereas ParaMEDIC only communicates about 108 MB—a *900-fold* reduction.

We also illustrate the I/O time in Figure 5. As shown, ParaMEDIC outperforms mpiBLAST by more than 2 orders of magnitude. This result is attributed to multiple aspects. First, given the shared network connection between the two sites, the achievable network performance is usually much lower than within the cluster. Thus, with mpiBLAST transferring the entire output over this network, its performance would be heavily impacted by the network performance. Second, the distance between the two sites causes the communication latency to be high. Thus, file-system control messages tend to take a long time to be exchanged, resulting in further loss in performance. Third, for mpiBLAST, since the wide-area network is a bottleneck, the number of simultaneously transmitted data streams does not matter; communication is serialized in the network. However, with ParaMEDIC, since the wide-area network is no longer a bottleneck, it can more effectively utilize the parallelism in the data streams to better take advantage of the storage bandwidth available at the storage site, as described in more detail in Section 5.2.

5.2. Storage bandwidth utilization

Figure 6(left) illustrates the storage bandwidth utilization achieved by mpiBLAST, ParaMEDIC, and the MPI-IO-Test benchmark, which is used as an indication of the peak performance capability

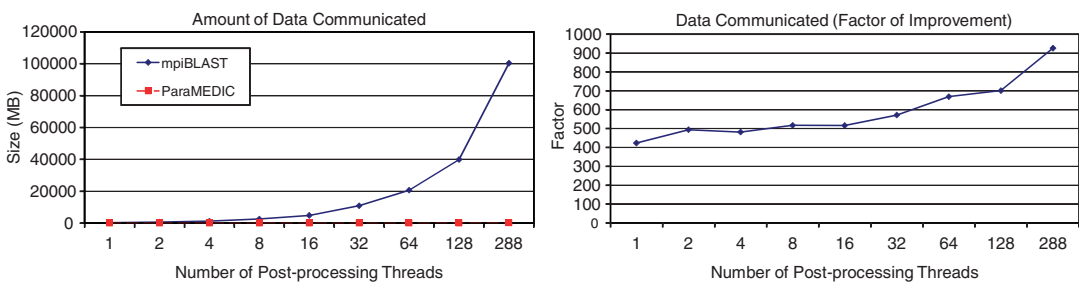


Figure 4. Data I/O overheads: (left) total amount of data communicated and (right) factor of improvement.

DOI: 10.1002/cpe

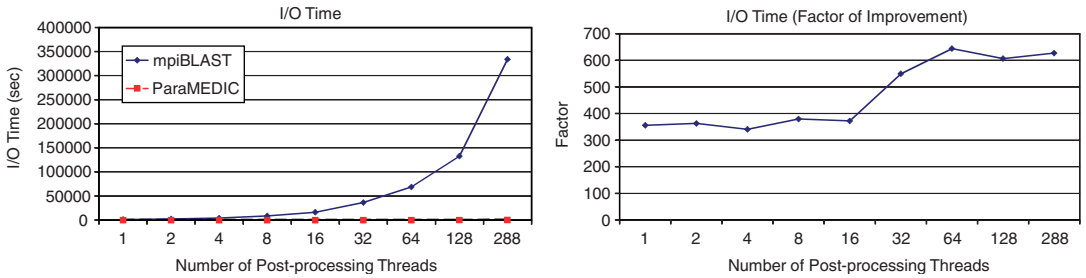


Figure 5. I/O time measurements: (left) total I/O time and (right) factor of improvement.

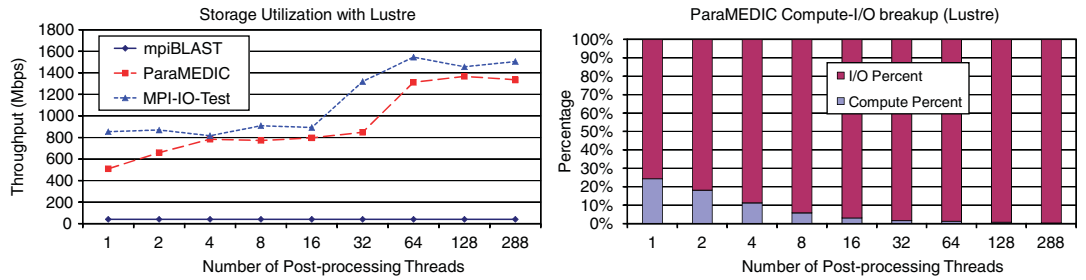


Figure 6. Storage bandwidth utilization using Lustre: (left) storage utilization improvement and (right) computation and I/O time.

of the I/O subsystem. We notice that the storage utilization of mpiBLAST is very poor compared to ParaMEDIC. The reason is that, for mpiBLAST, the I/O is limited by the wide-area network bandwidth. Thus, although more than 10 000 processors are performing the compute part of the task, the network connecting the compute servers in the U.S. and the storage system in Tokyo becomes the bottleneck.

On the other hand, ParaMEDIC uses more than 90% of the storage system capability (shown by MPI-IO-test). When the number of processing threads is low (x-axis in the figure), ParaMEDIC uses about half the storage capability. However, as the number of processing threads increases, the I/O utilization of ParaMEDIC increases as well.

Figure 6(right) illustrates the percentage breakup of the time spent in ParaMEDIC's post-processing phase. A significant portion of the time spent is in the I/O part. This shows that in spite of using a fast parallel file-system like Lustre, ParaMEDIC is still bottlenecked by the I/O subsystem. In fact, our analysis has shown that in this case the bottleneck lies in the 1-Gb Ethernet network subsystem connecting the storage nodes. Thus, we expect that even for systems with faster I/O subsystems, ParaMEDIC will further scale up and continue to use a significant portion of the I/O bandwidth provided.

In Figure 7(left), we remove the file-system network bottleneck and directly perform I/O on the local nodes. Storage utilization achieved in this case is significantly higher than going over the network. Even in this case, ParaMEDIC completely uses the storage capability with more than

DOI: 10.1002/cpe

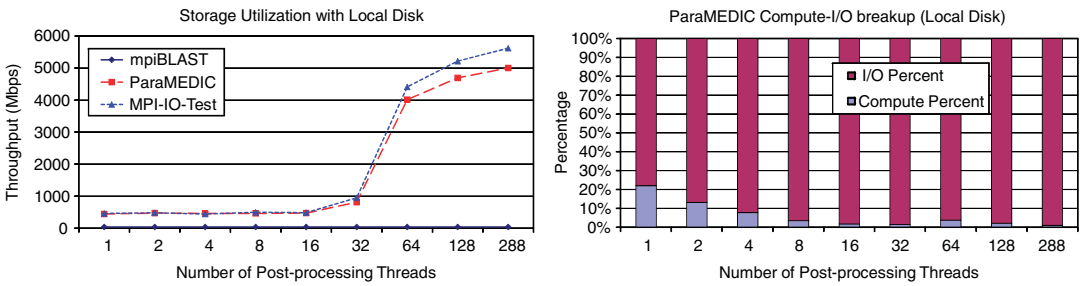


Figure 7. Storage utilization using local file-system: (left) storage utilization improvement and (right) computation and I/O time.

90% efficiency. Figure 7(right) shows the percentage breakup of the time spent. Similar to the case with the Lustre file-system, a significant portion of the time is still spent on I/O. Thus, again, ParaMEDIC can be expected to scale and fully use even faster storage resources.

6. DISCUSSION

Although this paper deals only with enhancing the mpiBLAST application through ParaMEDIC, the idea is relevant for many other applications as well. For example, applications that have natively been built for distributed environments such as SETI@home [23] and other BOINC applications [24] can easily use similar ideas and can benefit aspects in which such techniques are possible. In the field of communication profiling with MPE [25], we have also done some preliminary work that uses metadata transformation of profiled data through ParaMEDIC. Specifically, based on the observation that most scientific applications have a very uniform and periodic communication pattern, we perform a Fourier transform on the data to identify this periodicity and use this as an abstract block. The metadata comprises one complete abstract block and just the differences for all other blocks. Our preliminary numbers in this field have demonstrated between two- and fivefold reduction in the I/O time using ParaMEDIC. Work on other application fields, including earthquake modeling and remote visualization, is ongoing as well with promising preliminary results.

7. RELATED WORK

Efficient I/O access for scientific applications in distributed environments has been an ongoing subject of research for various parallel and distributed file-systems [26–29]. There has also been work on explicit data transfer protocols like GridFTP [30]. Other efforts include providing remote data access through MPI-IO [31]. RIO [32] introduced a proof-of-concept implementation that allows applications to access remote files though ROMIO [33]. RFS [34] enhanced the remote write performance with *active buffering*, by optimizing overlap between applications computation and I/O. Studies have also been done in translating MPI-IO calls into operations of lower level data protocols like Logistic Network [35]. However, all these approaches deal with data as a

DOI: 10.1002/cpe

byte-stream. Conversely, our approach focuses on aggressively reducing the amount of I/O data to be communicated by taking advantage of application semantics and dealing with data as high-level abstract units, rather than a stream of bytes.

Semantic-based data transformation is not new. Several semantic compression algorithms have been investigated in compressing relational databases [36–38]. Leveraging the table semantics, these algorithms first build descriptive models of the database using data mining techniques like clustering and then strip out data that can be regenerated. In the multimedia field, context-based coding techniques (similar to semantics-based approaches) have been widely used in various video compression standards [39–41]. With the aid of context modeling, these techniques efficiently identify redundant information in the media. Although sharing the same goal of reducing data to store or transfer with ParaMEDIC, these data compression studies do not address the remote I/O issue.

Thus, ParaMEDIC utilizes ideas from different fields to provide a novel approach for distributed I/O.

8. CONCLUSION

Rapid growth of computational power is enabling computational biology to tackle increasingly large problems, such as discovering missing genes and providing structure to genetic sequence databases. As the problems grow larger, however, so does the data consumed and produced by the applications. For many applications, the required compute power and storage resources cannot be found at a single location, precipitating the transfer of large amounts of data across the wide-area network. ParaMEDIC mitigates this issue by pursuing a non-traditional approach to distributed I/O. By trading computation for I/O, ParaMEDIC utilizes application semantics information to transform the output to orders of magnitude smaller metadata. In this paper, we presented our experiences in solving large-scale computational biology problems by utilizing nine different high-performance compute sites within the U.S. to generate a petabyte of data that then was transferred to a large-scale storage facility in Tokyo using ParaMEDIC's distributed I/O capability. We demonstrated that ParaMEDIC can achieve a performance improvement of several orders of magnitude compared with traditional I/O. In the future, we plan to evaluate semantic-based compression for other applications.

ACKNOWLEDGEMENTS

We thank the following people for their technical help in managing the large-scale run and other experiments associated with this article: R. Kettimuthu, M. Papka, and J. Insley from the University of Chicago; N. Desai and R. Bradshaw from Argonne National Laboratory; G. Zelenka, J. Lockhart, N. Ramakrishnan, L. Zhang, L. Heath, and C. Ribbens from Virginia Tech; M. Rynge and J. McGee from Renaissance Computing Institute; R. Fukushima, T. Nishikawa, T. Kujiraoka, and S. Ihara from Tokyo Institute of Technology; S. Vail, S. Cochrane, C. Kingwood, B. Cauthen, S. See, J. Fragalla, J. Bates, R. Cagle, R. Gaines, and C. Bohm from Sun Microsystems; and H. Liu from Louisiana State University/LONI.

DOI: 10.1002/cpe

REFERENCES

1. Reed DA. Grids, the TeraGrid, and beyond. *Computer* 2003; **36**(1):62–68.
2. SURFnet. Available at: http://www.surfnet.nl [February 2010].
3. CANARIE. Available at: http://www.canarie.ca [February 2010].
4. Ethernet everywhere: Price/performance as the key to cluster & grid interconnects. Available at: http://www.cse.scitech.ac.uk/disco/mew14-cd/Talks/Force10_VanCampen.pdf [February 2010].
5. Simms S, Pike G, Balog D. Wide area filesystem performance using Lustre on the TeraGrid. *TeraGrid Conference*, Madison, WI, 2007.
6. Balaji P, Feng W, Archuleta J, Lin H. ParaMEDIC: Parallel metadata environment for distributed I/O and computing. *IEEE/ACM International Conference for High Performance Computing*, *Networking*, *Storage and Analysis* (*SC Storage Challenge Award*), Reno, Nevada, November 2007.
7. Balaji P, Feng W, Archuleta J, Lin H, Kettimuthu R, Thakur R, Ma X. Semantics-based distributed I/O for mpiBLAST. *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, Salt Lake City, Utah, February 2008.
8. Balaji P, Feng W, Lin H. Semantics-based distributed I/O with the ParaMEDIC framework. *Proceedings of the IEEE/ACM International Conference on High Performance Distributed Computing* (*HPDC*), Boston, MA, June 2008.
9. Balaji P, Feng W, Lin H, Archuleta J, Matsuoka S, Warren A, Setubal J, Lusk E, Thakur R, Foster I, Katz DS, Jha S, Shinpaugh K, Coghlan S, Reed D. Distributed I/O with ParaMEDIC: Experiences with a worldwide supercomputer. *Proceedings of the International Supercomputing Conference* (*ISC Best Paper Award*), Dresden, Germany, 2008.
10. Altshul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. *Nature Genetics* 1994; **6**(2):119–129.
11. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. Genbank. *Nucleic Acids Research* 2008; **36**(Database issue).
12. Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara BC, Sanderson MJ. Prospects for building the tree of life from large sequence databases. *Science* 2004; **306**(5699):1172–1174.
13. Gans JD, Feng W, Wolinsky M. Whole genome, physics-based sequence alignment for pathogen signature design. *12th SIAM Conference on Parallel Processing for Scientific Computing*, San Francisco, CA, February 2006. (Electronic version unavailable.)
14. Havre SL, Webb-Robertson B-J, Shah A, Posse C, Gopalan B, Brockma FJ. Bioinformatic insights from metagenomics through visualization. *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, Palo Alto, CA, 8–11 August 2005; 341–350.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 1990; **215**(3):403–410.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI–BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 1997; **25**:3389–3402.
17. Darling AE, Carey L, Feng W. The design, implementation, and evaluation of mpiBLAST. *ClusterWorld Conference and Expo and the 4th International Conference on Linux Cluster*: *The HPC Revolution*, San Jose, CA, 2003.
18. Lin H, Ma X, Chandramohan P, Geist A, Samatova N. Efficient data access for parallel BLAST. *IPDPS*, Denver, CO, 2005.
19. Thorsen O, Smith B, Sosa CP, Jiang K, Lin H, Peters A, Feng W. Parallel genomic sequence-search on a massively parallel system. *ACM International Conference on Computing Frontiers*, Ischia, Italy, May 2007.
20. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics* 2007; **23**(6):673–679.
21. Lukashin A, Borodovsky M. GeneMark: New solutions for gene finding. *Nucleic Acids Research* 1998; **26**:1107–1115.
22. Warren A, Archuleta J, Feng WC, Setubal JC. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* 2010; **11**(1):131+.
23. Seti@home: The search for extraterrestrial intelligence. Available at: http://setiathome.ssl.berkeley.edu/ [February 2010].
24. BOINC: Berkeley Open Infrastructure for Network Computing. Available at: http://boinc.berkeley.edu [February 2010].
25. MPE: MPI Parallel Environment. Available at: http://www-unix.mcs.anl.gov/perfvis/download/index.htm [February 2010].
26. Howard. An overview of the Andrew file system. *USENIX Winter Technical Conference*, Dallas, TX, February 1988.
27. Nowicki B. *NFS*: *Network File System Protocol Specification*. Network Working Group *RFC1094*, 1989.
28. Carns P, Ligon II IW, Ross R, Thakur R. PVFS: A parallel file system for Linux clusters. *LSC*, Atlanta, GA, 2000.
29. Schmuck F, Haskin R. GPFS: A shared-disk file system for large computing clusters. *FAST*, Monterey, CA, 2002.
30. Allcock B, Bresnahan J, Kettimuthu R, Link M, Dumitrescu C, Raicu I, Foster I. The globus striped GridFTP framework and server. *SC*, Seattle, WA, 2005.
31. Message Passing Interface Forum. *MPI-2*: *Extensions to the Message-Passing Standard*, July 1997.
32. Foster I, Kohr D, Krishnaiyer R, Mogill J. Remote I/O: Fast access to distant storage. *Workshop on I/O in Parallel and Distributed Systems*, San Jose, CA, 1997.

DOI: 10.1002/cpe

33. Thakur R, Gropp W, Lusk E. Data sieving and collective I/O in ROMIO. *Frontiers of Massively Parallel Computation*, Los Alamitos, CA, 1999.
34. Lee J, Ma X, Ross R, Thakur R, Winslett M. RFS: Efficient and flexible remote file access for MPI-IO. *Cluster*, San Diego, CA, 2004.
35. Lee J, Ross R, Atchley S, Beck M, Thakur R. MPI-IO/L: Efficient remote I/O for MPI-IO via logistical networking. *IPDPS*, Rhodes Island, Greece, 2006.
36. Jagadish H, Madar J, Ng R. Semantic compression and pattern extraction with fascicles. *VLDB '99*: *Proceedings of the 25th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc.: San Francisco, California, 1999.
37. Babu S, Garofalakis M, Rastogi R. SPARTAN: A model-based semantic compression system for massive data tables. *SIGMOD Rec.*: *Special Interest Group on Management of Data*, vol. 30(2), 2001; 283–294.
38. Jagadish H, Ng R, Ooi B, Tung A. It compress: An iterative semantic compression algorithm. *Proceedings of International Conference on Data Engineering*, Boston, MA, 2004.
39. Generic Coding of Moving Pictures and Associated Audio Information—Part 2: Video. *ITU-T and ISO/IEC JTC1*, *ITU-T Recommendation H.262 ISO/IEC 13 818-2* (*MPEG-2*), 1994.
40. Coding of Audio-Visual Objects—Part 2: Visual. *ISO/IEC JTC1, ISO/IEC 14 496-2* (MPEG-4 Visual version 1), April 1999; Amendment1 (version 2), February 2000; Amendment 4 (streaming profile), January 2001.
41. Wiegand T, Marpe D, Schwarz H. Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard. *IEEE Transactions on Circuits and Systems for Video Technology* 2003; **13**(7):620–636.

DOI: 10.1002/cpe