Semantics-based Distributed I/O with the ParaMEDIC Framework

P. Balaji, W. Feng, and H. Lin

Preprint
Argonne National Laboratory #ANL/MCS-P1481-0108

Semantics-based Distributed I/O with the ParaMEDIC Framework[*]

P. Balaji
Math. and Comp. Science,
Argonne National Laboratory
balaji@mcs.anl.gov

W. Feng
Dept. of Computer Science,
Virginia Tech
feng@cs.vt.edu

H. Lin
Dept. of Computer Science,
North Carolina State University
hlin2@ncsu.edu

Abstract

Many large-scale applications simultaneously rely on multiple resources for efficient execution. For example, such applications may require both large compute and storage resources; however, very few supercomputing centers can provide large quantities of both. Thus, data generated at the compute site often has to be moved to a remote storage site for either storage or visualization and analysis. Clearly, this is not an efficient model, especially when the two sites are distributed over a wide-area network.

We present a framework called "ParaMEDIC: Parallel Metadata Environment for Distributed I/O and Computing" that uses application-specific semantic information to convert the generated data to orders-of-magnitude smaller metadata at the compute site, transfer the metadata to the storage site, and reprocess the metadata at the storage site to regenerate the output. Specifically, ParaMEDIC trades a small amount of additional computation (in the form of data post-processing) for a potentially significant reduction in data that needs to be transferred in distributed environments.

Keywords: distributed I/O, framework, metadata, profiling, performance evaluation

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. While system sizes have also grown over the past few years, most researchers do *not* have local access to systems of the scale that they need for their applications. Therefore, they access such large systems remotely to perform the required computations. Further, many of these applications require multiple resources simultaneously for efficient execution. For example, applications that perform large computations to 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 a remote compute site often has to be moved to a local storage site for either storage or visualization and analysis. Clearly, this is not an efficient model, especially when the two sites are distributed over a wide-area network.

These issues have become particularly onerous with the advent of petascale computing and the petabyte datasets that they produce. For instance, transferring a petabyte dataset, even over a dedicated 10-Gbps network such as National LambdaRail, requires on the order of ten days to complete, assuming that the network does not fail and that the transfer is uninterrupted and persistent. If these assumptions do not hold, then the transfer takes even longer. As a result, scientists and engineers continue to use the convenience of Federal Express for their data-transfer needs because next-day arrival of a petabyte data (stored on tapes or disk drives) is preferred over the ten-day arrival via a high-speed network [6]. Furthermore, we argue that the integrity of the dataset sent via Federal Express will be better than the network because the relatively weak 16-bit checksum of TCP/IP over a fast wide-area network will result in undetected bit errors ("silent errors") approximately every 500 GB [13].

Many of these issues are artificially self-imposed by our current set of computational tools or by our misuse of cyberinfrastructure resources, e.g., not using local compute resources intelligently to accelerate large-scale data transfers. Thus, we present a novel framework called "ParaMEDIC: Parallel Metadata Environment for Distributed I/O and Computing" to effectively use both large-scale remote computational resources and local compute resources to perform efficient data movement in distributed environments. ParaMEDIC uses application-specific semantic information and converts the generated data from a given application to orders-of-magnitude smaller metadata at the compute site. It then transfers the metadata to the storage site and reprocesses the metadata at the storage site to regenerate the output. Specifically, ParaMEDIC trades a small amount of additional computation (in the form of data post-processing) for a potentially significant reduction in data that needs to be transferred in distributed environments. Furthermore, this trade-off is tunable with respect to desired performance and available computational resources.

Together with a detailed description of the ParaMEDIC framework, this paper presents a performance evaluation of ParaMEDIC on three distributed systems. The first system used TeraGrid nodes at the University of Chicago and San Diego Supercomputer Center (30-Gbps network connection). The second system consists of a secure filesystem hosted between Argonne National Laboratory and Virginia Tech over an Internet2 connection (1-Gbps connection). The third system is a local cluster connected with a dedicated 10-Gbps network that also has the ability to vary both bandwidth and latency between any two nodes. We use this last system to emulate various

[*]This research is funded in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357 and the Computer Science Department at Virginia Tech.

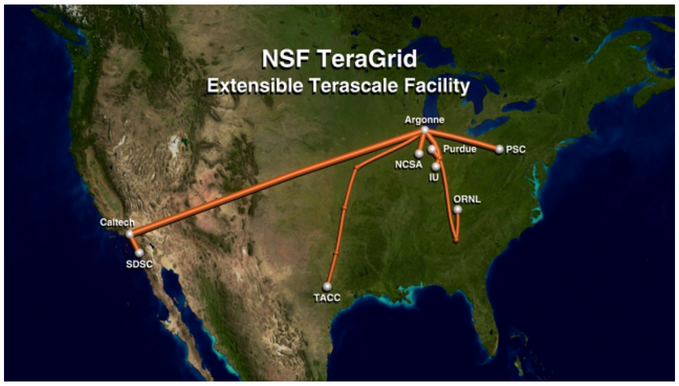


Figure 1: NSF TeraGrid Infrastructure

forms of high-latency and high-bandwidth distributed computing infrastructures. Our experimental results demonstrate that ParaMEDIC achieves *order-of-magnitude improvements* in performance compared to the traditional approach of blindly moving data in distributed environments.

The rest of the paper is organized as follows. Section 2 describes two sample distributed computing environments and their properties. Section 3 presents the design details of the ParaMEDIC framework. Two case studies that use ParaMEDIC to enhance the performance of real-world applications in distributed environments are discussed in Section 4. Section 5 presents our experimental evaluation of ParaMEDIC, and Section 6 addresses work related to ours. Finally, we conclude the paper in Section 7 with a discussion of future work.

2 *Sample Distributed Environments*

A wide variety of distributed environments exist today, ranging from high-latency, high-bandwidth environments such as *LambdaGrids*, to low-bandwidth environments such as those connected over the Internet or Internet2, to unsecure environments that require data encryption before being transmitted. Here we present two such distributed environments.

2.1 *NSF TeraGrid*

NSF TeraGrid, as shown in Figure 1, is a distributed computing facility combining leadership-class resources at eleven partner sites within the U.S. to create an integrated, persistent computational resource. Using high-performance network connections, the TeraGrid integrates high-performance computers, data resources and tools, and high-end experimental facilities around the country. Currently, TeraGrid resources include more than 750 teraflops of computing capability and more than 30 petabytes of online and archival data storage, with rapid access and retrieval over high-performance networks. Researchers can also access more than 100 discipline-specific databases. With this combination of resources, the TeraGrid is the world's largest, most comprehensive distributed cyberinfrastructure for open scientific research.

The TeraGrid comprises several sites including the University of Chicago/Argonne National Laboratory (Illinois), San Diego Supercomputer Center (California), Purdue University (Indiana), Texas Advanced Computing Center (Texas), and others. The San Diego Supercomputer Center (SDSC) also doubles as a host for a global parallel filesystem (GPFS) that is visible and usable by all TeraGrid compute servers. All sites are connected via high-bandwidth optical links. However, the large physical distance between the sites forces the latency to be high (up to tens of milliseconds).

Scientists use the computational power available at different locations to run computations and then write the final output to the globally shared filesystem to be viewed or post-processed at a later time. While such a system provides good computational capability for I/O-rich applications, the distributed filesystem can form a significant bottleneck.

2.2 *Argonne-VT Distributed System*

The Argonne-VT distributed supercomputing system is a small-scale research infrastructure built to share application output data for post-processing and visualization. The system comprises 200 processors of shared compute resources and about 200 gigabytes (GB) of main memory spread across the two sites. The Argonne site also provides 10 terabytes of storage resources for use by the compute systems.

The storage resources available at Argonne National Laboratory are shared across the two sites using a distributed file-system that uses the shared Internet2 connectivity (1 Gbps). This network connectivity is much slower than the NSF TeraGrid infrastructure. Furthermore, though this network is mostly a dedicated infrastructure, it occasionally experiences large traffic bursts, causing further performance degradation in the network. In addition, because the connection between the two sites is over the regular wide-area network, it is considered to be insecure, thus, the data transmission might require encrytion, which can add substantial overhead as well.

3 *ParaMEDIC Design Overview*

In this section, we present a detailed description of the ParaMEDIC framework (short for *Para*llel *M*etadata *E*nvironment for *D*istributed *I*/O and *C*omputation). We describe the overall framework in section 3.1, followed by details of trade-offs between computation and I/O in Section 3.2. In Section 3.3, we describe the management of the computational resources available and its impact.

3.1 *The ParaMEDIC Framework*

ParaMEDIC provides a framework for *decoupling* computation and I/O in applications that rely on large quantities of both. As shown in Figure 2, the framework provides several capabilities including support for data encryption and data integrity as well as data transfer in distributed environments (either directly via

TCP/IP communication or through global filesystems). However, the primary semantics-based metadata creation is done by the *application plugins*.

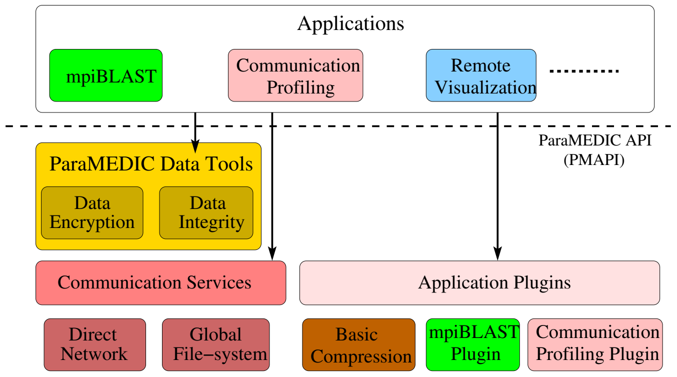


Figure 2: ParaMEDIC Architecture

Most application plugins are specific to each application and thus rely on knowledge of application semantics. These plugins provide two primary functionalities:

1. Functionality to perform post-processing on the data generated by the application in order to create metadata. This metadata can either be *intermediate data* generated during the application execution or data that is separately extracted by processing the generated output.

2. Functionality to convert the metadata back to the final output. This conversion can be 100% accurate or lossy, depending on the application requirements. However, in this paper, we deal only with lossless data conversion.

Though the development of application-specific plugins is mostly straightforward, they require application knowledge to be developed. Thus, as an intermediate solution, ParaMEDIC also provides a standard compression plugin that views the output data as a byte stream and performs non-application-specific compression on it. While this compression plugin is straightforward to use and does not require any application knowledge (completely application transparent), the performance benefits it can achieve are minimal as compared to application-specific plugins. Thus, we do not discuss this approach further in this paper and concentrate only on application-specific plugins.

3.2 *Trading Computation with I/O cost*

The amount of computation required in the ParaMEDIC framework is higher than what is required by the original application. For example, after the output is generated by the application processes, it has to be further processed to generate the metadata, sent to the destination nodes, 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, the ParaMEDIC framework aims at trading (a small amount of) additional computation for reduced I/O cost.

The ParaMEDIC framework 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 that is performed on the output data, with the general trend being, the more the post-processing computation, the better the reduction in the meta-data size. That is, an application plugin can perform more processing of the output data to reduce the I/O cost at the cost of the additional processing overhead. Similarly, it can perform very little processing of the output data to save on the processing overhead, but at the cost of a larger amount of data that might need to be transferred. In the rest of this section, we formalize this capability of ParaMEDIC to trade additional computational cost for larger savings in I/O.

Consider a distributed environment connecting a large-scale supercomputer that is performing the computation and generating the data and a small-scale system where the data has to be stored or visualized. In this context, we use the following definitions:

B: Network bandwidth in the above distributed environment.

T: Overall application execution time.

D: Total data generated by the application.

f(x): Time taken to convert the output data to x units of metadata.

g(y): Time taken to convert y units of metadata to final output.

Based on the above definitions, the actual time (A) taken for the application to be executed and the data to be moved to the target system is given by the sum of four components: (a) computation time of the application, (b) time to convert the generated output to metadata, (c) time to transfer the metadata over the network, and (d) time to post-process the metadata to generate back the required output. This is formalized by the following equation:

$$A = T + f(x) + \frac{x}{B} + g(x) \tag{1}$$

The actual time taken by the application without using ParaMEDIC, on the other hand, would be given by the summation of the application computation time and the time to transfer the overall data:

$$A' = T + \frac{D}{B} \tag{2}$$

Clearly, ParaMEDIC is beneficial only when A (Equation 1) is small than A' (Equation 2), that is,

$$T + f(x) + \frac{x}{B} + g(x) < T + \frac{D}{B}$$

Simplifying the above equation, we get

$$D - x > B \times (f(x) + g(x)) \tag{3}$$

Thus, as illustrated by Equation 3, ParaMEDIC would be beneficial only when the difference between the actual data size and the metadata size is larger than the network bandwidth times the amount of time taken for the post-processing (including converting the output to metadata and the metadata back to the final output). Accordingly, ParaMEDIC is expected to significantly outperform the native implementations of the applications when either the network bandwidth of the distributed environment is low or the post-processing cost is low. For example, even in distributed environments such as the NSF TeraGrid which have a 40-Gbps high-speed network connectivity between the sites, ParaMEDIC can improve performance if the application data post-processing time is low enough.

Figure 3 illustrates the implications of Equation 3. As shown in the figure, applications in *Quadrant 1* have a large difference in size between the actual data and the processed metadata, while requiring only a small amount of metadata post-processing. Such applications can fully benefit from the ParaMEDIC framework. Applications in *Quadrant 2* have a large difference in size between the actual data and the processed metadata, but they require a large amount of metadata post-processing as well. Hence, it is not clear whether such applications can fully benefit from ParaMEDIC. Applications in *Quadrant 3* almost never benefit from ParaMEDIC because they use a lot of post-processing but do not reduce the size of the data output too much. For applications in *Quadrant 4*, it is not clear whether they would benefit from ParaMEDIC, because although they require only a small amount of post-processing, they do not reduce the amount of data significantly.

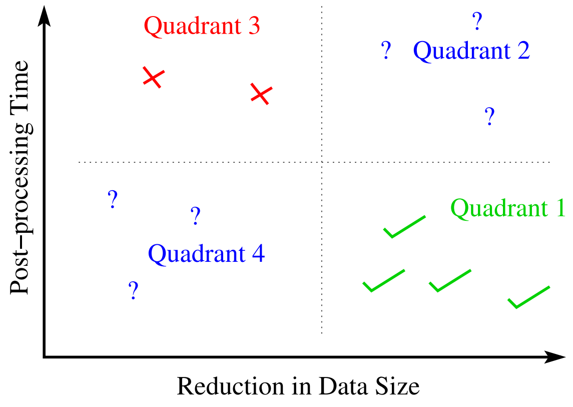


Figure 3: Application Output Data Patterns

3.3 *Managing Computational Resources*

In ParaMEDIC, the conversion of the actual output to metadata and the conversion of metadata back to the actual output does not come free of cost. Computational resources have to be expended for such post-processing as well. Thus, if the total number of compute resources is fixed, managing how many of these resources are provided for the actual application computation and how many are provided for post-processing essentially determines the tradeoff in the amount of time spent in computation versus the amount of time saved in I/O.

Specifically, if too many compute resources are given for application processing and too little for the post-processing, the time taken for post-processing either increases significantly or only superficial post-processing will be possible, resulting in larger amounts of metadata that needs to be communicated. On the other hand, if too few compute resources are given for the application processing and too many for the post-processing, the application execution time increases rapidly, resulting in loss of performance as well.

In general, since the post-processing resources are restricted to the client system (i.e., the system where the final data has to be moved to), these resources are limited. For example, a remote large-scale cluster where the actual application is executed might have about 10,000 processors, while the local small-scale cluster used by the scientist might have only about 1,000 processors. In this case, it is ideal to maintain a 10:1 ratio between the number of compute resources allocated to the actual application and the number of compute resources allocated to the post-processing. Deviating from this ratio will result in degradation in performance.

Hence ParaMEDIC must use algorithms that can generate metadata from the result and re-generate the result from the metadata with minimal processing requirements. That is, for most applications, the framework ensures that the post-processing cost required to process the metadata and generate the final output is significantly less than the actual computation needed to search for the query sequence within the database.

Evaluation describing the trade-offs associated with managing the number of computation resources provided to application computation vs. post-processing are presented in Section 5.2.2.

4 *Case Studies in Scientific Computing*

As described in Section 3, different applications can plug into the ParaMEDIC framework by providing application-specific plugins using the ParaMEDIC API. In this section, we discuss two sample applications that can take advantage of this framework, namely mpiBLAST (in Section 4.1) and the MPI Parallel Environment (MPE) profiling library (in Section 4.2).

4.1 *mpiBLAST Sequence Search Application*

Here we present an overview of the mpiBLAST application and discuss how we integrate it with the ParaMEDIC framework.

4.1.1 *Application Overview*

mpiBLAST [4] parallelizes the NCBI BLAST toolkit [1], the de facto “gold standard” for sequential pairwise sequence-search that is ubiquitously used in biomedical research. The BLAST tool searches one or multiple input query sequences against a database of known nucleotide (DNA) or amino acid sequences. A score (termed `evalue`) is calculated for each close match based on a statistical model. A database sequence can have many matches with the query sequences. The similarity of the comparison is measured by the match with the highest score. The database sequences that are most similar to the query sequence are reported, along with their matches scored beyond a certain threshold. In other words, the BLAST process is essentially a top-k search, where k can be specified by the user, with a default value of 500.

The core of the mpiBLAST algorithm is based on *database segmentation*. Before the search, the raw sequence database is formatted, partitioned into fragments, and stored in a shared storage space. mpiBLAST then organizes parallel processes into one master and many workers. The master breaks down the search job in a Cartesian-product manner and maintains a list of un-searched tasks, each represented as a pair of a query sequence and a database fragment. Whenever a worker becomes idle, it asks the master for a un-searched task, copies the needed fragment to its local disk (if the fragment has not been cached locally), and performs a BLAST search on its assignment. Upon finishing a task, a worker reports its local results to the master for centralized result merging. Once the results of searching a query sequence against all database fragments have been collected, the master calls the standard NCBI BLAST output function to format and print results of this query to the output file. By default, those results contain the top 500 database sequences with highest similarity to the query sequence along with their matches. This process repeats until all tasks have been searched. With database segmentation, mpiBLAST can deliver super-linear speedup when searching sequence databases larger than the memory of a single node. In recent developments, mpiBLAST has evolved to use a more scalable parallel approach that allows different queries to be searched concurrently as well [9].

4.1.2 *Integration with ParaMEDIC*

In a cluster environment, most of the mpiBLAST execution time is spent on the search itself, comparing input query sequences to the database fragments, because the search requires a full scan of the database fragment and the BLAST string-matching algorithm is computationally intensive (quadratic complexity in the worst case). In contrast, the cost of formatting and writing the results that are generated from the search is much less significant, especially when many advanced clusters are configured with high-performance parallel filesystems.

However, in a distributed environment such as those presented in Section 2, the execution profile of mpiBLAST differs significantly from what is found in a cluster environment because mpiBLAST output typically needs to be written over a wide-area network to a remote filesystem. Hence, the cost of writing the results can easily dominate the execution profile of mpiBLAST and, thus, become a severe performance bottleneck.

This is where "ParaMEDIC comes to the rescue" for mpiBLAST. By replacing the traditional global parallel filesystem over a wide-area network with the ParaMEDIC framework (as shown at the top of Figure 4), we can still support basic functionality of a global parallel filesystem, e.g., transferring large volumes of data to a distant filesystem (if needed). But more importantly we can also provide advanced functionality that trades a small amount of additional computation for a potentially significant reduction in data that needs to be transferred in distributed environments. For example, as we show in Section 5, an mpiBLAST-specific instance of ParaMEDIC reduces the volume of data that needs to written across a wide-area network by *more than two orders of magnitude*.

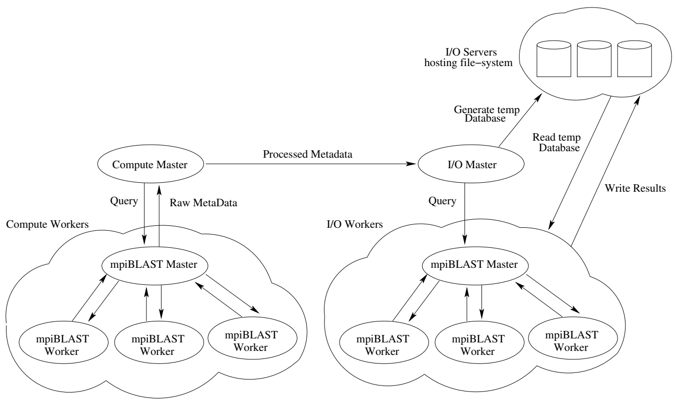


Figure 4: ParaMEDIC and mpiBLAST Integration

Specifically, Figure 4 depicts how mpiBLAST can be integrated with the ParaMEDIC framework. First, on the compute site (the left cloud in Figure 4), instead of having mpiBLAST collect and write all the result sequences and their matches of a query sequence, the mpiBLAST application plugin in ParaMEDIC, as shown in Figure 2, generates semantics-based metadata based on the mpiBLAST output at the compute site. ParaMEDIC then transfers this metadata to the I/O site (the right cloud in Figure 4), metadata that is orders of magnitude smaller than the actual data output that would have been transferred in a traditional global parallel filesystem. Next, at the I/O site, a small amount of additional computation is performed on the metadata in order to regenerate the actual output data. That is, a temporary (and much smaller) database that contains only the result sequences is created by extracting the corresponding sequence data from a local database replica. ParaMEDIC then reruns mpiBLAST at the I/O site by taking as input the same query sequence and the temporary database to generate and write output to the local filesystem. The overhead in rerunning mpiBLAST at the I/O site is quite small because the temporary database that is searched is substantially smaller, with only 500 sequences in it by default, as opposed to the several millions of sequences in large DNA databases.

4.2 *MPE Communication Profiler*

The MPI Profiling Environment (MPE) is a suite of performance analysis tools comprising profiling libraries, utility programs, graphical tools, and checking libraries. MPE is a part of the MPICH2 distribution of the Message Passing Interface

(MPI) communication standard but can be used by any MPI implementation that provides the MPI profiling interface. The MPE profiler generates logs viewable by the integrated Jumpshot visualization tool. The datatype and collective verification library finds argument inconsistencies in MPI calls. The tracing library records all MPI calls and the animation and X-graphics libraries provide a real-time program animation of the trace using X-window routines.

In this paper, we are interested primarily in the profiling capabilities of MPE. MPE performs postmortem performance analysis based on trace file generated during parallel program execution. It uses CLOG2 as a low-overhead logging format, a simple collection of single timestamp events.

4.2.1 *Output Data Management in MPE*

Depending on the communication pattern, number of processors used, and length of execution, the output data generated by MPE can be enormous. For example, a large-scale application such as FLASH when executed on a 16,384-processor cluster can easily generate 40GB of data every second. That is, for an hour-long run, the amount of data generated is close to 150-terabytes. Thus, the output data has to be managed carefully in order to avoid performance overheads.

MPE allocates a default 8MB memory buffer in each process during initialization time. During the run, MPE profiles its communication pattern and stores this information in this memory buffer. As the memory buffer fills up, the content of the buffer is written to the local storage. At the end of the run, all processes form a binary tree. Each process reads in its local trace data from the disk and performs a three-way merge of its own buffer with the buffers sent from its children. When a merged buffer is filled up, it will then send to its parent. At the root process, the overall merged buffers are written to local storage.

Once the profiled data is obtained, it can viewed through visualization tools such as Jumpshot. For this process, the data has to be available locally for the scientists to visualize. Thus, if the application was executed and profiled at a remote location, the profile data has to be transferred (often over a distributed environment) to be processed and viewed.

4.2.2 *Integration with ParaMEDIC*

As described in Section 4.2.1, the amount of data generated by the MPE profiling tool is enormous. However, most scientific applications follow a periodic pattern where they repeatedly perform the same task to refine a data set, or to process different portions of the data set. Accordingly, the communication pattern for such applications is periodic as well. Figure 5 shows this pattern for the GROMACS molecular dynamics code.

While most application communication patterns are periodic, the periodicity for each application is different. Thus, the primary idea for integrating MPE with ParaMEDIC is two-fold:

1. Find the periodicity of the profiled data using a Fast Fourier Transform (FFT) of the data.
2. Use this periodicity information to convert the actual output into metadata that breaks the output into periods and stores only differences between the periods, rather than the entire communication profiling information.

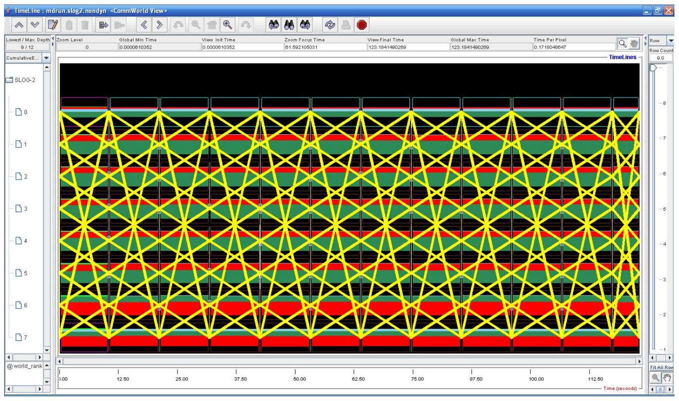


Figure 5: Periodic Communication Pattern in Applications

Finding Output Periodicity with FFT: Though FFT is a popular approach for finding the periodicity of data sets, it can be quite compute intensive to evaluate, especially for large datasets. Specifically, FFT is an `O(N.log(N))` time operation, where `N` is the number of points in the dataset. Further, several applications also have recursive periodicity. That is, within each recurring communication pattern, these applications might have a separate recurring pattern embedded. While identifying such patterns recursively can allow us to improve the I/O savings even further, this can significantly increase the amount of computation required as well. Thus, in our approach, we consider only the first level of periodicity.

Given the importance of FFT in scientific computing, several parallel implementations of FFT have been designed and developed. FFTW [8], Parallel 3D FFT (P3DFFT) [3], and BG/L 3D FFT (BGL3DFFT) are some popular implementations of parallel FFT. Such implementations can reduce the time taken to compute the periodicity of the data set by using multiple compute nodes in parallel. Also, since the periodicity detection has to take place only on the computational site, this process can be heavily parallelized by using the large number of compute resources available on the site, thus further reducing the total time taken. The post-processing required to convert the metadata back to the actual output does not require another FFT calculation and can be done much more efficiently even with the lesser compute resources on the storage site. In our approach, we used the Parallel 3D FFT (P3DFFT) implementation of FFT.

Metadata Storing Only Differences between Periods: Once the periodicity of the dataset is calculated, the repeating information in each period does not have to be stored and can be regenerated. In the metadata format for this application, we store the complete information for the first period and store only the differences for the remaining periods.

For most applications, the differences between each period is minimal even when storing the timing information to up to two decimal places. For systems such as IBM Blue Gene, where the system noise level is extremely low, the accuracy is even higher. Thus, the size of the differences is also very small, allowing for larger savings in the amount of data that needs to be communicated.

Our experimental results demonstrate about 3-5X improvement in performance for the MPE profiling library by using ParaMEDIC. However, because of space constraints, we do not present detailed performance results for this library in this paper. Rather we restrict ourselves to the detailed performance results with the mpiBLAST application.

5 *Experimental Evaluation and Analysis*

This section presents a performance evaluation of mpiBLAST in its native form, as compared to ParaMEDIC-enhanced mpiBLAST (hereafter referred to as simply ParaMEDIC). Specifically, we evaluate the two schemes in a local cluster emulating various distributed computing infrastructures in Section 5.2. Evaluations on a distributed testbed between Argonne National Laboratory and Virginia Tech over the Internet2 network connection are presented in Section 5.3. Evaluations on the NSF TeraGrid infrastructure with nodes from University of Chicago and San Diego Supercomputing Center participating in the experiment are presented in Section 5.4. Performance results for the MPE profiling library are not presented because of space constraints.

All experiments have been performed with the Nucleotide (NT) database downloaded from the NCBI website. NT contains the GenBank, EMB L, D, and PDB sequences. At the time our experiments were performed, it contained over 5 million sequences with a total raw size of about 20GB. All queries were synthesized by randomly sampling sequences from NT itself.

5.1 *Experimental Testbeds*

We used three testbeds for our evaluation, as described below.

Testbed 1 (Local Cluster): This testbed consists of 24 nodes, each equipped with two dual-core Opteron 2220 2.8GHz processors (totally 4 cores). Each processor has 2MB of L2-cache (1MB local L2 cache for each core). The nodes were equipped with 4GB of 667MHz DDR2 SDRAM and four SATA disks using a software RAID0. The network used to connect these machines was the NetEffect NE010 10-Gigabit Ethernet iWARP adapters. For the experiments, they were used as regular 10-Gigabit Ethernet adapters with host-based TCP/IP without using the iWARP support. The nodes were installed with Fedora Core 6 operating system.

To emulate the various distributed computing infrastructures, we used the NetEm software package, which allows data from a network to be delayed without hampering the overall throughput. This package enabled us to emulate high-latency, high-bandwidth distributed computing environments by separating the nodes in the system into two sub-clusters, where communication within the sub-cluster is fast, while within sub-clusters is slow.

Testbed 2 (Argonne-VT Distributed System): This testbed consists of two clusters (one at Argonne and one at VT) connected over Internet2. The Argonne cluster is the same as the one given in *Testbed 1*, while the VT cluster consists of a 24-node Orion D-12 system. The Orion Multisystems DT-12 is a personal *desktop* cluster workstation that contains 12 individual x86 compute nodes in a 24” x 18” x 4” (or one cubic foot) pizza-box enclosure. Each compute node contains a Transmeta Efficeon processor running the Linux operating system and Transmeta’s power-aware LongRun2 software, its own memory and Gigabit Ethernet interface. The nodes share a power supply, cooling system, and external 10-Gigabit Ethernet network connection. At load, it achieves 14 Gflops on Linpack while drawing only 185 watts of power.

Testbed 3 (NSF TeraGrid): This testbed consists of a subset of the TeraGrid, using nodes at the University of Chicago and the San Diego Supercomputer Center. The University of Chicago site comprises two sets of nodes. The first set has 96 2.4GHz Intel Xeon 32-bit dual-processor systems each with 3GB memory and 512KB L2 cache, while the second set has 64 1.5GHz Intel Itanium II 64-bit dual-processor systems each with 4GB memory. The San Diego Supercomputer Center site comprises 64 1.5GHz Intel Itanium II 64-bit dual-processor systems each with 4GB memory. The two sites are connected with a 40-Gbps high-bandwidth network, and the end-to-end delay between the two sites is approximately 10ms.

All experiments in Section 5.2 were performed on *Testbed 1*, while experiments in Sections 5.3 and 5.4 were performed on *Testbed 2* and *Testbed 3* respectively.

5.2 *Local Cluster Evaluation*

Here we compare the performance of the ParaMEDIC scheme with that of the existing mpiBLAST implementation on a local cluster, which emulates various distributed computing infrastructures.

5.2.1 *Impact of High-Latency, High-Bandwidth Networks*

In this section, we analyze the impact of distributed environments connected with high-latency, high-bandwidth networks on the performance of ParaMEDIC and basic mpiBLAST. For this experiment, we divide the local cluster into two logical sub-clusters. While all the nodes are connected with a 10Gbps network, we artificially delay the communication between nodes belonging to different sub-clusters. The actual amount of artificial delay is varied in the experiment, and the performance of the application for different network delays is measured. The socket buffer sizes are set to be equal to the bandwidth-delay product of the network so as to maximize the performance that the network subsystem can provide. Four nodes (each node with 4 SATA disks connected with a software RAID0) in the second sub-cluster host a PVFS2 filesystem that is visible to all nodes in both sub-clusters.

For the evaluation, 80 processors are used for the computation. For native mpiBLAST, all the processors were used for performing the actual application computation. However, for ParaMEDIC, in order to keep the overall computational resources constant, 76 processors hosted on the first sub-cluster were used for performing the actual computation, while 4 processors on the second sub-cluster were used for the post-processing.

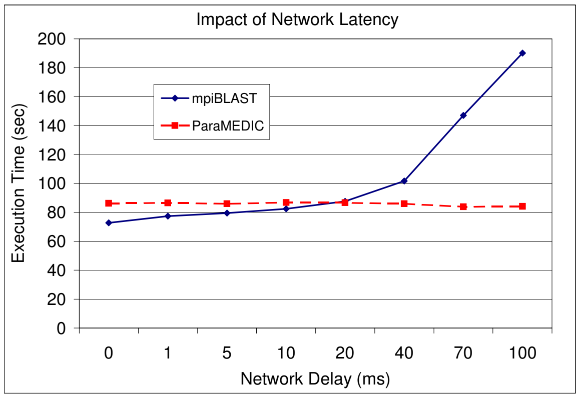


Figure 6: Impact of High Latency Networks

As shown in Figure 6, when the network delay between the two sub-clusters is low, the original mpiBLAST outperforms ParaMEDIC. This result is expected because ParaMEDIC performance additional computation for converting the search results to metadata and converting the metadata back to the final output. As the network delay increases, however, ParaMEDIC starts to outperform mpiBLAST. In fact, for a network delay of 100ms, ParaMEDIC outperforms mpiBLAST by a factor of 2.26. This improvement in performance is attributed to two reasons. First, the high network latency causes degradation in the file-system communication and synchronization operations that are required whenever data needs to be written or read from the server. Second, the total amount of data written in mpiBLAST over the I/O subsystem is much higher as compared to ParaMEDIC, since ParaMEDIC writes only metadata that is significantly smaller than the final results to the filesystem.

To further clarify these results, we show the performance breakdown of the time taken by mpiBLAST and ParaMEDIC in Figure 7. As shown in Figure 7(a), for mpiBLAST, as the network delay increases, the I/O time increases very quickly. Thus, though the computation time does not change much, the overall execution time suffers. On the other hand, for ParaMEDIC (Figure 7(b)), the computation time, the I/O time, and the post-processing time required to handle the metadata are nearly constant for all values of network delays. This result is expected because the only component in ParaMEDIC that would be affected by the network latency is the post-processing, since it requires moving the metadata from the compute workers to the I/O workers. And because the metadata amount is very small (few kilobytes), this time typically does not make any difference to either the post-processing time or the overall execution time of the application.

We have also performed experiments where ParaMEDIC does not use reduced computational resources for the actual application processing. That is, it is allowed to use all 80 processors instead of just 76 for the application processing, together with four *additional* processors for post-processing. In this case, we saw a further improvement in performance. However, we feel that the results are quite similar to those shown above and, to avoid repetition, are not shown in this paper.

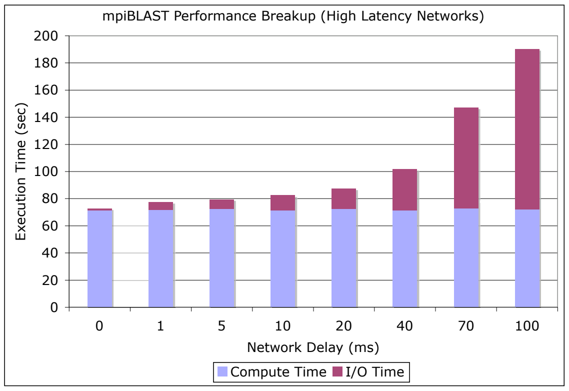
5.2.2 *Trading Computation to I/O*

This section analyzes the performance of mpiBLAST and ParaMEDIC by varying the ratio of computational resources allocated to the actual application processing vs. the resources allocated for post-processing. For all experiments in this section, for ParaMEDIC, we allocate four processes for performing the post-processing. The number of processes allocated for the actual application computation is varied from 16 to 80. That is, the ratio of resources allocated for actual application computation to post-processing is varied from 4:1 to 20:1. For mpiBLAST, on the other hand, all of the processes are allocated for the actual application computation. That is, the native mpiBLAST implementation gets four extra processes for application computation as compared to ParaMEDIC.

The ratio of resources allocated to application processing and to post-processing essentially determines the trade-off in computation time to I/O time. Specifically, a large ratio means that more resources are given for application processing; thus the resources given for metadata generation and management are minimal. This implies that the application execution time will be lesser, while the amount of data that needs to be moved over the distributed environment will be large. Similarly, a small ratio means that fewer resources are given for application processing; thus the resources given for metadata generation and management is high. This implies that the application execution time will be greater, while the amount of data that needs to be moved over the distributed environment will be small.

Figure 8 shows the performance of the two schemes as the ratio of resources allocated to application processing and to post-processing is varied. As shown in the figure, when this ratio is low, mpiBLAST outperforms ParaMEDIC. As the ratio increases, however, the performance of mpiBLAST degrades faster than ParaMEDIC, and it is eventually outperformed by ParaMEDIC.

This behavior is related to the actual number of compute resources the application has for execution. Specifically, when the total number of compute resources is *N*, ParaMEDIC uses only (N-4) of them for the application computation. The remain-



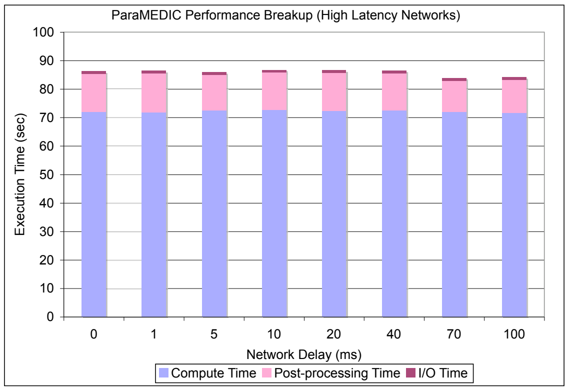


Figure 7: Breakup of Performance with Network Delay: (i) mpiBLAST and (ii) ParaMEDIC

ing 4 processes are used for metadata post-processing. Thus, when N is very large, the increase in computation time caused by using resources (approximately N / (N - 4)) is not very high. When N is small, however, the increase in computation time can be substantial. For example, when N is 8 processes, ParaMEDIC uses only 4 processes for the application processing while mpiBLAST uses 8. Thus, the computation time taken by ParaMEDIC would be nearly twice that of mpiBLAST. This overshadows any benefit in the I/O time ParaMEDIC can bring, causing it to deliver worse performance than mpiBLAST. In summary, ParaMEDIC is most effective only when the number of resources used for application processing is sufficiently large as compared to the number of resources used for post-processing.

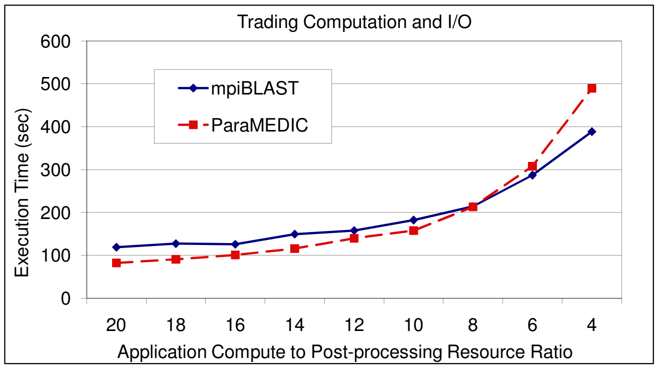


Figure 8: Varying the Number of Worker Processes

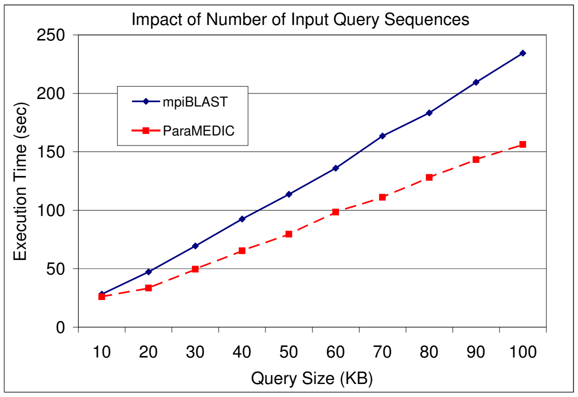
5.2.3 *Impact of Output Data Size*

In this section, we vary two parameters that affect the output data size: the number of input query sequences provided by the user and the number of output query sequences requested by the user. We study the impact of each parameter on the performance of mpiBLAST and ParaMEDIC. Varying the number of sequences in the input query increases the search time for both mpiBLAST and ParaMEDIC. However, it is also expected to impact the post-processing time for ParaMEDIC. Thus, increasing the query size is expected to affect the computation time of ParaMEDIC more than that of mpiBLAST. At the same time, an increase in the query size also typically results in more output and can potentially impact mpiBLAST more than ParaMEDIC. Figure 9(a) shows the performance of the two schemes with increasing number of input query sequences (depicted by input query size). We see that while the increase in the input query size increases the execution time of ParaMEDIC, it has a more drastic effect on mpiBLAST. Thus, as the query size increases, we notice that the performance difference between the two schemes increases, with ParaMEDIC outperforming mpiBLAST by about 66% for a 100KB query file size.

Figure 9(b) shows the impact of increasing the number of requested output result sequences. Increasing the number of output result sequences does not increase the computation much, while it can increase the amount of I/O. Thus, because the I/O cost for ParaMEDIC is very low, increasing the number of output result sequences does not vary its performance too much. On the other hand, since the I/O cost for mpiBLAST is very high, increasing the number of output result sequences significantly hurts its performance.

5.2.4 *Impact of Encrypted Filesystems*

For distributed filesystems that span unsecure network connections (such as the Internet), using data encryption to protect transmitted data is a common occurrence in environments such as government national laboratories and other secure facilities. Figure 10 shows the impact of such data encryption on the performance of the two schemes. As shown in the figure, the performance of the two schemes is similar to the case where there is no file encryption (except that the performance of mpiBLAST degrades faster). This is attributed to the data encryption overhead. That is, since all the data that is being transmitted has to be encrypted and the amount of data transmitted by mpiBLAST over an unsecure network is significantly larger than the amount transmitted by ParaMEDIC, encryption affects mpiBLAST more significantly than it does ParaMEDIC.



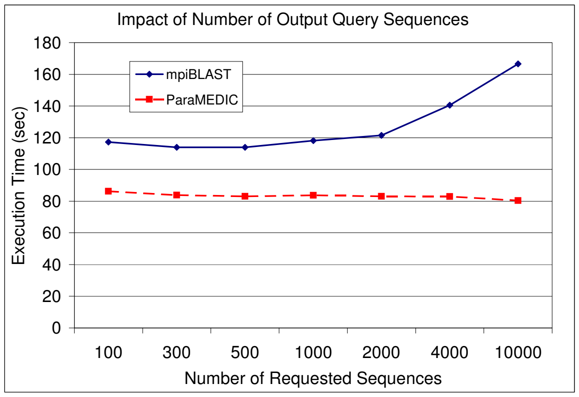


Figure 9: Varying the Number of Requested Sequences: (i) Input Query Sequences and (ii) Output Result Sequences

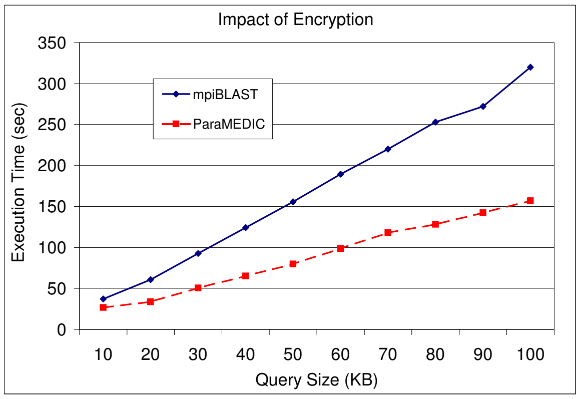


Figure 10: Impacted of Encrypted Filesystems

5.3 *Distributed Setup from Argonne and VT*

In this section, we evaluate the performance of mpiBLAST and ParaMEDIC on a distributed system between Argonne National Laboratory and Virginia Tech connected over Internet2. Since the network connecting the two clusters is *not* secure, data encryption is used to protect the data transmitted over this network.

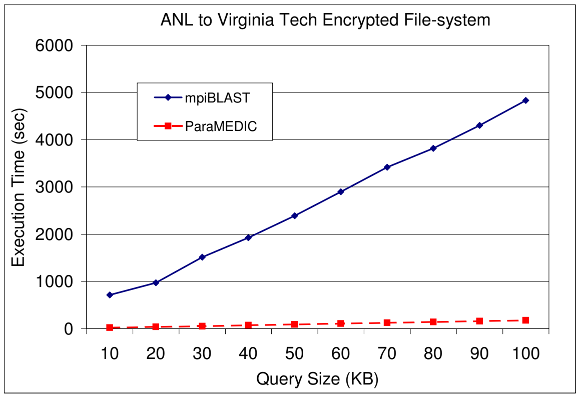


Figure 11: Argonne to Virginia Tech Encrypted Filesystem

As shown in Figure 11, ParaMEDIC significantly outperforms mpiBLAST in this environment. Further, as the query size increases, the performance difference between the two schemes increases. For a query size of 100KB, we observe more than a *25-fold improvement* in performance for ParaMEDIC as compared to mpiBLAST. This difference is attributed to multiple aspects. First, given that the network connection between the two sites is shared by other users, the effective network performance achievable is usually much lower than within the cluster. Thus, with mpiBLAST transferring the entire output result over this network, its performance would be heavily impacted by the network performance. Second, since data communicated is encrypted, mpiBLAST also has to pay the penalty for such encryption. Though ParaMEDIC also pays such data encryption penalty, the amount of data it transfers is significantly less, and hence the penalty is less as well. Third, the distance between the two sites causes the communication latency to be high. Thus, file-system communication and synchronization messages tend to take a long time to be exchanged resulting in further loss of performance.

5.4 *TeraGrid Infrastructure*

The TeraGrid infrastructure represents a widely used environment for several compute- and I/O-intensive applications. As described in Section 2, a GPFS-based distributed filesystem is hosted at San Diego Supercomputer Center (SDSC), which can be accessed from all facilities, and forms a part of the TeraGrid facility. For the experiments in this section, we used the nodes at the University of Chicago and SDSC.

Both mpiBLAST and ParaMEDIC perform their application computation on the University of Chicago nodes. However, mpiBLAST directly writes the output data to the global GPFS file-system. ParaMEDIC, on the other hand, converts the output data to metadata, transfers the metadata to the SDSC site, and reconverts the metadata to the final output at the SDSC site.

Figure 12 illustrates the performance of mpiBLAST and ParaMEDIC on the TeraGrid infrastructure. While the final output is written to the same global filesystem in both cases, mpiBLAST suffers from the fact that the application process-

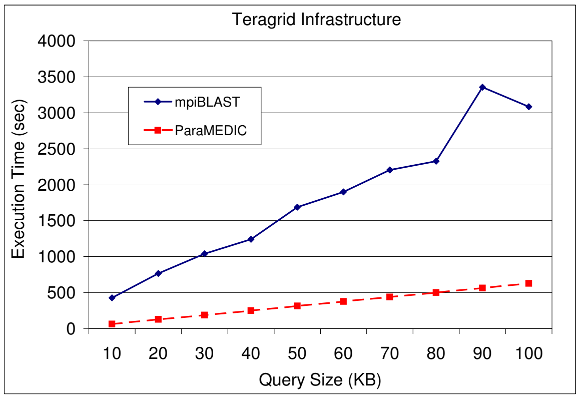


Figure 12: TeraGrid Infrastructure using University of Chicago and the San Diego Supercomputing Center

ing nodes at the University of Chicago are performing the I/O for the output results. Since these nodes reside on a remote cluster as compared to a physical file-system, their I/O performance is limited resulting in an overall degradation in execution time. For ParaMEDIC, on the other hand, since the post-processing nodes are performing the I/O for the output results, the amount of time taken is significantly smaller. For a query file size of 100KB, ParaMEDIC outperforms mpiBLAST by a factor of about five times.

Figure 13 shows the performance breakdown of the two schemes. As shown in the figure, as the query size increases, the computation time for both mpiBLAST and ParaMEDIC increases. However, for mpiBLAST, the I/O time also increases quickly. On the other hand, for ParaMEDIC, there is practically no difference in the I/O time with increasing query sizes. ParaMEDIC is only minimally impacted by the limited I/O of the subsystem, and it efficiently distributes its computational resources across the system to achieve high performance.

6 *Related Work*

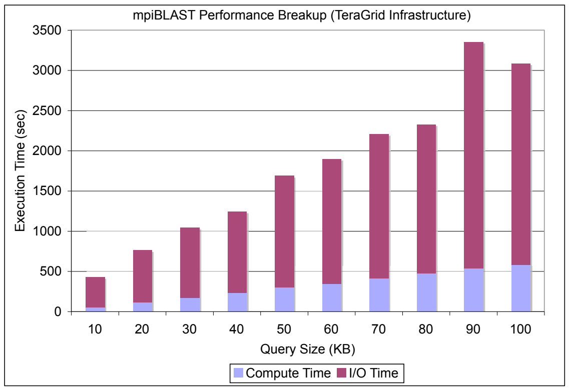
Providing portable, efficient remote I/O capabilities for scientific applications has been an ongoing subject of research. RIO [7] introduced a proof-of-concept library that allowed application to access remote files though the ROMIO [15] interface. It addressed the portability issue by using a client-server architecture and hiding the platform dependent details within the ADIO layer of ROMIO. However, RIO relied on a legacy communication protocol and required setting up extra nodes dedicated for message processing. Those limitations were addressed in a more recent study RFS [10]. RFS also adopted the active buffering technique to reduce the visible remote write cost, which essentially optimized the overlaps between application I/O and computation. Other approaches of translating remote I/O requests into operations of general data transferring protocols such as Grid FTP [2] and Logistic Network [11] have also been investigated. Different from existing remote I/O studies, ParaMEDIC focuses on aggressively reducing the amount of I/O data that needs to be shipped across the wide area network with efficient utilization of application semantics knowledge.

A few efforts have been made to alleviate the I/O bottleneck in parallel BLAST. The version of mpiBLAST used in this paper overlaps the gather phase (i.e., I/O phase) and search phase (i.e., compute phase) by searching only a few sequences at a time, instead of the entire query and immediately sending the results to the master [9]. Although decreasing the total runtime over previous versions of mpiBLAST, as seen in Section 5.2.1, the small data transfers and many smaller filesystem messages are impacted by the larger latency across a wide-area network. A second approach for improving I/O performance uses parallel I/O to write the output file. pioBLAST offloads I/O from the master and distributes the work among the workers [12]. By having each worker write a portion of the output file in parallel, the time to write the output file is reduced. In turn, this decrease in processing time for the gather phase significantly improves the scalability of parallel BLAST to hundreds of processors. This scalability, however, is realizable only on advanced parallel filesystems. Furthermore, the small metadata synchronization messages of a parallel filesystem will hinder performance when the filesystem is connected across a wide-area network.

Example frameworks of decoupling computation and I/O include MapReduce and TCP Linda. MapReduce is a programming model and an associated implementation for processing and generating large datasets [5]. TCP Linda is a virtual shared-memory system whereby parallel processes execute simultaneously and exchange data be generating, reading, and consuming data objects [14]. Both MapReduce and TCP Linda decouple the different phases of an algorithm and transfer intermediate objects to the appropriate processes. In ParaMEDIC, these intermediate objects are the metadata consisting of the GI numbers corresponding to the subset of sequences to be searched against on the I/O nodes. That MapReduce has been shown to achieve high performance when working with large datasets in [5] is indicative that ParaMEDIC, a similar model, should also be able to substantially improve performance, and indeed it has.

7 *Conclusions and Future Work*

In this paper, we presented a novel framework called ParaMEDIC (Parallel Metadata Environment for Distributed I/O and Computing) that uses application-specific semantic information to convert the generated data to orders-of-magnitude smaller metadata at the compute site, transfer the metadata to the storage site, and reprocess the metadata at the storage site to regenerate the output. In other words, ParaMEDIC trades a small amount of additional computation (in the form of data post-processing) for a potentially significant reduction in data that needs to be transferred in distributed environments. We presented the detailed design of the framework and presented experimental evaluations on different experimental as



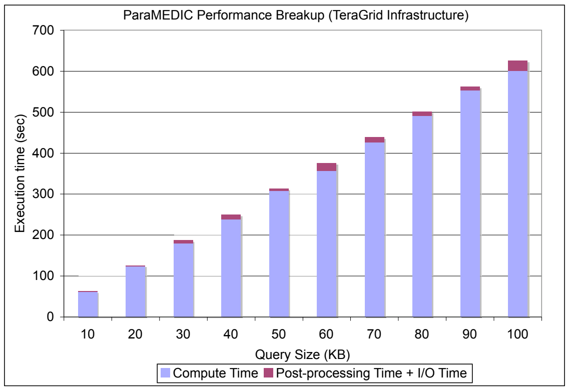


Figure 13: TeraGrid Infrastructure Performance Breakup: (i) mpiBLAST and (ii) ParaMEDIC

well as real distributed systems. Our results show an *order-of-magnitude* improvement in performance with ParaMEDIC in some cases.

Currently, ParaMEDIC provides two sets of functionalities—one set that is completely application transparent (such as data integrity, data encryption) and a second set that is completely based on application semantics. As future work, we plan to abstract certain functionalities that are neither completely application-dependent nor application-transparent but are common for some important *classes* of applications. For example, some of the metadata generation aspects described for mpiBLAST are common for many bioinformatics applications and can be abstracted out as utility functionality within the ParaMEDIC framework itself. This will allow other application developers to use such functionality to create their respective plugins with more ease.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSIBLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] T. Baer and P. Wyckoff. A Parallel I/O Mechanism for Distributed Systems. In *Proceedings of the International Conference on Cluster Computing*, 2004.

[3] San Diego Supercomputing Center. Parallel 3D FFT Library. http://www.sdsc.edu/us/% linebreak[0]resources/p3dfft.php.

[4] A. Darling, L. Carey, and W. Feng. The Design, Implementation, and Evaluation of mpiBLAST. In *International Conference on Linux Clusters: The HPC Revolution 2003*, 2003.

[5] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *6th Symposium on Operating System Design and Implementation (OSDI)*, 2004.

[6] A. Snavely (ed.). NSF Workshop Report: Petascale Computing in the Biological Sciences. Technical report, National Science Foundation.

[7] I. Foster, D. Kohr, R. Krishnaiyer, and J. Mogill. Remote I/O: Fast access to distant storage. In *Proceedings of the Fifth Workshop on I/O in Parallel and Distributed Systems*, pages 14–25, San Jose, CA, November 1997.

[8] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. special issue on "Program Generation, Optimization, and Platform Adaptation".

[9] M. Gardner, W. Feng, J. Archuleta, H. Lin, and X. Ma. Parallel genomic sequence-searching on an ad-hoc grid: Experiences, lessons learned, and implications. In *ACM/IEEE SC2006: The International Conference on High-Performance Computing, Networking, and Storage*, 2006.

[10] J. Lee, X. Ma, R. Ross, R. Thakur, and M. Winslett. RFS: Efficient and flexible remote file access for MPI-IO. In *Proceedings of the IEEE International Conference on Cluster Computing*, 2004.

[11] J. Lee, R. Ross, S. Atchley, M. Beck, and R. Thakur. MPI-IO/L: efficient remote i/o for mpi-io via logistical networking. In *Parallel and Distributed Processing Symposium, 2006.*, 2006.

[12] H. Lin, X. Ma, P. Chandramohan, A. Geist, and N. Samatova. Efficient Data Access for Parallel BLAST. In *International Parallel and Distributed Processing Symposium*, Apr 2005.

[13] Vern Paxson. End-to-end Internet packet dynamics. In *Proceedings of the ACM SIGCOMM '97 conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, volume 27 of *Computer Communication Review*, pages 139–154, Cannes, France, September 1997. ACM Press.

[14] TCP Linda. `http://www.lindaspaces.com/products/linda_overview.html`.

[15] R. Thakur, W. Gropp, and E. Lusk. Data sieving and collective I/O in ROMIO. In *Proceedings of the 7th Symposium on the Frontiers of Massively Parallel Computation*, February 1999.