

Parallel Visualization on Large Clusters using MapReduce

Huy T. Vo*
Polytechnic Institute
of New York University

Jonathan Bronson
SCI Institute
University of Utah

Bill Howe
eScience Institute
University of Washington

Brian Summa
SCI Institute
University of Utah
Valerio Pascucci
SCI Institute
University of Utah

João L.D. Comba
Instituto de Informática
UFRGS, Brazil

Cláudio T. Silva
Polytechnic Institute
of New York University

Juliana Freire
Polytechnic Institute
of New York University

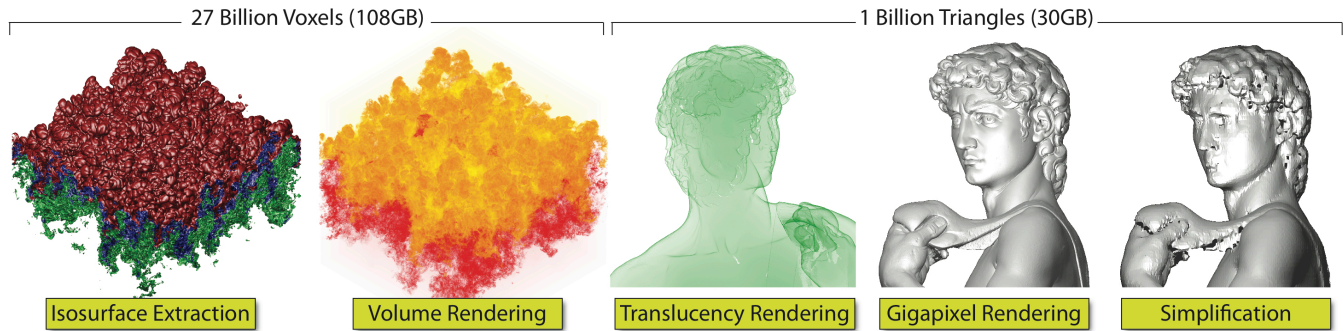


Figure 1: A representative suite of visualization tasks being evaluated with MapReduce: isosurface extraction, volume and mesh rendering, and mesh simplification. Our MapReduce-based renderer can produce a giga pixel rendering of a 1 billion triangle mesh in just under two minutes. With the capability of sustaining high I/O rate with fault-tolerance, MapReduce methods can be used as tools for quickly exploring large datasets with isosurfacing and rendering in a batch-oriented manner.

ABSTRACT

Large-scale visualization systems are typically designed to efficiently “push” datasets through the graphics hardware. However, exploratory visualization systems are increasingly expected to support scalable data manipulation, restructuring, and querying capabilities in addition to core visualization algorithms. We posit that new emerging abstractions for parallel data processing, in particular computing clouds, can be leveraged to support large-scale data exploration through visualization. In this paper, we take a first step in evaluating the suitability of the MapReduce framework to implement large-scale visualization techniques. MapReduce is a lightweight, scalable, general-purpose parallel data processing framework increasingly popular in the context of cloud computing. Specifically, we implement and evaluate a representative suite of visualization tasks (mesh rendering, isosurface extraction, and mesh simplification) as MapReduce programs, and report quantitative performance results applying these algorithms to realistic datasets. For example, we perform isosurface extraction of up to 16 isovalues for volumes composed of 27 billion voxels, simplification of meshes with 30GBs of data and subsequent rendering with image resolutions up to 80000^2 pixels. Our results indicate that the parallel scalability, ease of use, ease of access to computing resources, and fault-tolerance of MapReduce offer a promising foundation for a combined data manipulation and data visualization system deployed in a public cloud or a local commodity cluster.

Keywords: MapReduce, Hadoop, cloud computing, large meshes, volume rendering, gigapixels.

*corresponding author: huy.vo@nyu.edu

Index Terms: I.3.3 [Computer Graphics]: Picture/Image Generation—Display Algorithms

1 INTRODUCTION

Cloud computing has emerged as a viable, low-cost alternative for large-scale computing and has recently motivated industry and academia to design new general-purpose parallel programming frameworks [5, 8, 30, 45]. In contrast, large-scale visualization has traditionally benefited from specialized couplings between hardware and algorithms, suggesting that migration to a general-purpose cloud platform might incur in development costs or scalability¹.

The MapReduce framework [8, 9] provides a simple programming model for expressing loosely-coupled parallel programs using two serial functions, *Map* and *Reduce*. The *Map* function processes a block of input producing a sequence of *(key, value)* pairs, while the *Reduce* function processes a set of values associated with a single key. The framework is responsible for “shuffling” the output of the *Map* tasks to the appropriate *Reduce* task using a distributed sort. The model is sufficiently expressive to capture a variety of algorithms and high-level programming models, while allowing programmers to largely ignore the challenges of distributed computing and focus instead on the semantics of their task. Additionally, as implemented in the open-source platform Hadoop [14], the MapReduce model has been shown to scale to hundreds or thousands of nodes [8, 33]. MapReduce clusters can be constructed inexpensively from commodity computers connected in a shared-nothing configuration (*i.e.*, neither memory nor storage are shared across nodes). Such advantages motivated cloud providers to host Hadoop and similar frameworks for processing data at scale [1, 7].

These platforms have been largely unexplored by the visualization community, even though these trends make it apparent that our

¹ Scalability refers to the relative performance increase by allocating additional resources.

community must inquire into their viability for use in large-scale visualization tasks. The conventional modus operandi of “throwing datasets” through a (parallel) graphics pipeline relegates data manipulation, conditioning, and restructuring tasks to an offline system and ignores their cost. As data volumes grow, these costs — especially the cost of transferring data between a storage cluster and a visualization cluster — begin to dominate. Cloud computing platforms thus open new opportunities in that they afford both general-purpose data processing as well as large-scale visualization.

In this paper, we take a step towards investigating the suitability of the cloud-based infrastructure for large-scale visualization. We observed that common visualization algorithms can be naturally expressed using the MapReduce abstraction with simple implementations that are highly scalable. We designed MapReduce-based algorithms for memory-intensive visualization techniques, and evaluated them with several experiments. Results indicate that MapReduce offers a foundation for a combined storage, processing, analysis, and visualization system that is capable of keeping pace with growth in data volume (attributable to scalability and fault-tolerance) as well as growth in application diversity (attributable to extensibility and ease of use). Figure 1 illustrates results for isosurface extraction, volume and mesh rendering, and simplification.

In summary, the main contributions of the paper are:

- The design of scalable MapReduce-based algorithms for core, memory-intensive visualization techniques: mesh and volume rendering, isosurface extraction, and mesh simplification;
- An experimental evaluation of these algorithms using both a multi-tenant cloud environment and a local cluster;
- A discussion on the benefits and challenges of developing visualization algorithms for the MapReduce model.

2 RELATED WORK

Recently, a new generation of systems have been introduced for data management in the cloud, such as file systems [3, 23], storage systems [6, 10], and hosted DBMSs [29, 42]. MapReduce [8, 44] and similar massively parallel processing systems (*e.g.*, Clustera [11], Dryad [20], and Hadoop [14]) along with their specialized languages [5, 30, 45] are having a great impact on data processing in the cloud. Despite their benefits to other fields, these systems have not yet been applied to scientific visualization.

One of the first remote visualization applications [39] uses the X Window System’s transport mechanism in combination with Virtual Network Computing (VNC) [34] to allow remote visualization across different platforms. IBM’s Deep Computing Visualization (DCV) system [18], SGI’s OpenGL Vizserver [37] and the Chromium RenderServer (CRRS) [32] perform hardware accelerated rendering for OpenGL applications. A data management and visualization system for managing finite element simulations in materials science, which uses Microsoft’s SQL Server database product coupled to IBM’s OpenDX visualization platform is described in [15]. Indexes provide efficient access to data subsets, and OpenDX renders the results into a manipulable scene allowing inspection of non-trivial simulation features such as crack propagation. However, this architecture is unlikely to scale beyond a few nodes due to its dependency on a conventional database system.

Another approach to distributed visualization is to provide access to the virtual desktop on a remote computing system [18, 24, 32, 37], such that data remains on the server and only images or graphics primitives are transmitted to the client. Systems like VisIt [24] and ParaView [31] provide a scalable visualization and rendering backend that sends images to a remote client. Many scientific communities are creating shared repositories with increasingly large, curated datasets [19, 27, 38]. To illustrate the scale of these projects, the LSST [27] is predicted to generate 30 terabytes of raw data per

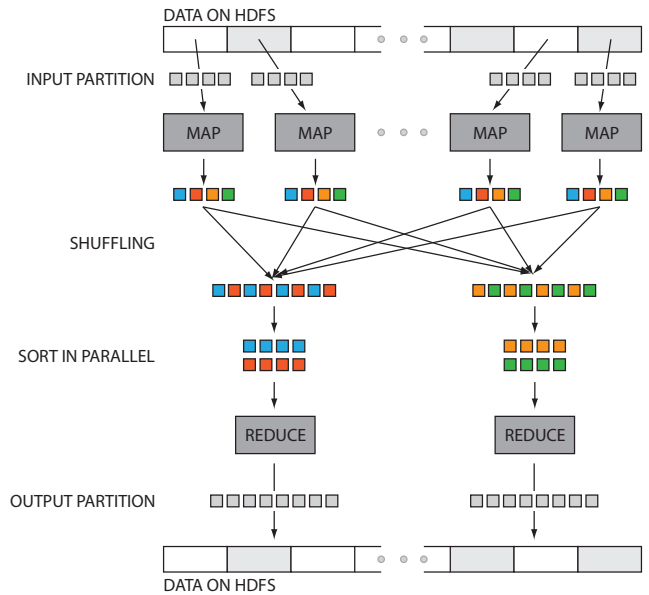


Figure 2: Data transfer and communication of a MapReduce job in Hadoop. Data blocks are assigned to several Maps, which emit key/value pairs that are shuffled and sorted in parallel. The Reduce step emits one or more pairs, with results stored on the HDFS.

night for a total of 6 petabytes per year. Systems associated with these repositories support only simple retrieval queries, leaving the user to perform analysis and visualization independently.

3 MAPREDUCE OVERVIEW

MapReduce is a framework to process massive data on distributed systems. It provides an abstraction that relies on two operations:

- Map: Given input, emit one or more (*key, value*) pairs.
- Reduce: Process all values of a given key and emit one or more (*key, value*) pairs.

A MapReduce job is composed of three phases: map, shuffle and reduce. Each dataset to be processed is partitioned into fixed-size blocks. In the map phase, each task processes a single block and emits zero or more (*key, value*) pairs. In the shuffle phase, the system sorts the output of the map phase in parallel, grouping all values associated with a particular key. In Hadoop, the shuffle phase occurs as the data is processed by the mapper (*i.e.*, the two phases overlap). During execution, each mapper hashes the key of each key/value pair into bins, where each bin is associated with a reducer task and each mapper writes its output to disk to ensure fault tolerance. In the reduce phase, each reducer processes all values associated with a given key and emits one or more new key/value pairs. Since Hadoop assumes that any mapper is equally likely to produce any key, each reducer may potentially receive data from any mapper. Figure 2 illustrates a typical MapReduce job.

MapReduce offers an abstraction that allows developers to ignore the complications of distributed programming — data partitioning and distribution, load balancing, fault-recovery and inter-process communication. Hadoop is primarily run on a distributed file system, and the Hadoop File System (HDFS) is the default choice for deployment. Hadoop has become a popular runtime environment for expressing workflows, SQL queries, and more [16, 30]. These systems are becoming viable options for general purpose large-scale data processing, and leveraging their computational power to new fields can be a very promising prospect. For example, MapReduce systems are well-suited for *in situ visualization*,

which means that data visualization happens while the simulation is running, thus avoiding costly storage and post-processing computation. There are several issues in implementing *in situ* visualization systems as discussed by Ma [28]. We posit that the simplicity of the implementation, inherent fault-tolerance, and scalability of MapReduce systems make it a very appealing solution.

4 VISUALIZATION ALGORITHMS USING MAPREDUCE

We describe MapReduce algorithms for widely-used and memory-intensive visualization techniques: mesh rendering using volumetric and surface data, isosurface extraction, and mesh simplification.

4.1 Rendering

Out-of-core methods have been developed to render datasets that are too large to fit in memory. These methods are based in a streaming paradigm [12], and for this purpose the rasterization technique is preferred due to its robustness, high parallelism and graphics hardware implementation. We have designed a MapReduce algorithm for a rasterization renderer for massive triangular and tetrahedral meshes. The algorithm exploits the inherent properties of the Hadoop framework and allows the rasterization of meshes composed of gigabytes in size and images with billions of pixels.

Our implementation first asks Hadoop to partition the input “triangle soup” among mappers with the constraint that each partition must be a multiple of 36 bytes, *i.e.*, the size of each triangle on disk, to avoid block boundary splitting a triangle. For each triangle, the mapper computes its projection onto the image plane and its corresponding pixels. For each pixel, the mapper outputs a (key,value) pair, with the key being the pixel location in the image plane (x, y), and the value being the depth and color of the pixel. The MapReduce framework sorts pixels into the proper order (row-major) to construct the final image. Pixel colors emitted by the mapper that share the same image plane location are grouped by this sorting. The reducer emits the smallest depth value for each pixel location, therefore accomplishing the z-buffering algorithm automatically. In Figure 3, we give an overview of this algorithm. Mappers and reducers are viewed as geometry and multi-fragment shaders, respectively, in two distinct phases. This parallels a graphics hardware pipeline and can be similarly extended to handle more advanced visualizations by custom “geometry shaders” and “fragment shaders.” For example, in a volume renderer, each reducer sorts its fragments and composite them, instead of a selection based on depth.

4.2 Isosurface Extraction

Isosurfaces are instrumental in visual data exploration, allowing scientists to study function behavior in static or time-varying scenarios. Given an input scalar volume, the core of extraction is the computation of the isosurface as a collection of simplicial primitives that can be rendered using common graphical methods. Our MapReduce-based algorithm for isosurface extraction is based on the Marching Cubes algorithm [26], which is the de-facto standard for isosurface extraction due to its efficiency and robustness.

Partitioning relies on the natural representation of a scalar volume as a collection of 2D slices. The Hadoop distributed file system uses this strategy to partition data into blocks for each mapper, but imposes some constraints. First, each partition must contain complete slices. Second, it allows the overlap by one slice in only one direction to account for triangles spanning across partitions. Although it may result in duplication of input data, there is no duplication of output triangles since this overlap only occurs in one dimension. In practice, the duplication of input data is small and has no significant effect on the performance of the system. Each mapper computes the triangles of several isosurfaces using the Marching Cubes algorithm and emits a (key,value) pair for each isovalue. The key is the isovalue and the value is the triangle data for the each cube in binary format. The reducer receives the data sorted

and binned by isovalue, thus, the reduce stage only needs to act as a pass-through, writing the isosurface as a triangle soup to file.

4.3 Mesh Simplification

Despite advances in out-of-core methods for rendering structured or unstructured meshes, it may still not be feasible to use the full resolution mesh. Several mesh simplification techniques have been proposed [13, 35, 43]. Memory usage is a key aspect of this problem, since techniques often require storage proportional to the size of the input or output mesh. An alternative is given by the OoCSx (improved Out-of-Core Simplification) algorithm [25], which decouples this relationship and allows the simplification of meshes of arbitrary sizes. This is accomplished by superimposing a regular grid over the underlying mesh with associations between grid cells and vertices: every grid cell that contains a vertex of the input mesh must also contain a vertex on the output mesh, and every cell must have only one representative vertex. The problem is broken into finding all triangles that span three unique cells, and then finding an optimum representative vertex for each cell. Only a linear pass through the triangle mesh to hash each vertex is needed to find its representative bin before the output of all triangle indices.

The linear pass of OoCSx is not suitable for a parallel implementation, due to the minimal error criteria for optimal representative vertices. We use two MapReduce jobs to implement the algorithm since it requires two sorting phases. The first Map phase bins each vertex into a regular grid to ensure that all triangles contributing vertices to a particular bin arrive on the same node in the Reduce phase. It also computes the quadric measure vector associated with the contributing triangle. For each triangle, three (key, value) pairs are emitted, one for each vertex. The key is the bin coordinate that contains the vertex, and the value is a concatenation of the quadric measure vector with the three indices of the triangle.

The first Reduce phase receives the same (key, value) pair from the Map phase, but sorted and grouped by key. It reads each unique key (bin), and uses the quadric measures of all triangles falling into that bin to compute the representative vertex. If the indices of all vertices of a triangle contributing to a representative vertex are unique, the Reduce phase emits the indexed triangle as key, and the current grid cell and vertex. Thus, across all reducers, there is exactly three (key,value) pairs with the same key (triangle), each storing a different representative vertex and corresponding bin as its value. Since multiple Reduce phases are currently not supported, we use a second MapReduce job to complete the dereference. The second Map reads and emits the data output from the first Reduce job. Keyed on triangle index, the second Reduce receives the exact three bin-vertex pairs, and emit as final output the simplified mesh.

5 EXPERIMENTAL ANALYSIS

An in-depth analysis of the algorithms was presented in the previous section. We designed our experiments to evaluate, for each algorithm, its ability to scale up and down, as well as the overhead introduced by Hadoop. The first series of tests shows the cost of data transfer through a MapReduce job without any computation, followed by a detailed evaluation of each individual algorithm.

By default, the number of mappers that Hadoop launches for a job is a multiple of the number of data blocks, without exceeding the actual number of blocks on its HDFS (counting all replications). On the other hand, the number of reduce tasks can be specified. To simplify comparison, in our tests we maximize the number of reducers to the system capacity while keeping its ratio to the number of mappers equal to 1. The number of mappers and reducers is always equal whenever the number of input data blocks permits.

Tests were performed on two Hadoop-based systems: a local cluster and the NSF CLuE cluster managed by IBM [7]. The local cluster consists of 60 nodes, each with two quad-core Intel Xeon Nehalem 2.6GHz processors, 24GB of memory and a 320GB disk.

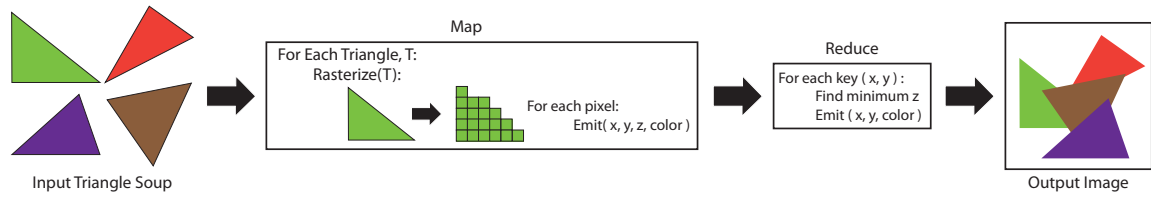


Figure 3: MapReduce rasterization. The map phase rasterize each triangle and emits rasterized fragments, i.e. the pixel coordinates as key, and its color and depth as value. The reducer composites fragments for each location, e.g. pick the smallest depth one for surface rendering.

The CLuE cluster consists of 410 nodes each with two single-core Intel Xeon 2.8GHz processors, 4GB of memory and a 400GB disk. While still a valuable resource for research, the CLuE hardware is outdated if compared to modern clusters, since it was originally built in 2004. Thus, we mostly utilize the performance numbers from the CLuE cluster as a way to validate and/or compare with our results on the local cluster. Since the CLuE cluster is a shared resource among multiple universities, there is currently no way to run experiments in isolation. We made sure to run all of our experiments at *dead* hours to minimize the interference from other jobs. HDFS files were stored in 64MB blocks with 3 replications.

5.1 MapReduce Baseline

To evaluate the cost incurred solely from streaming data through the system, several baseline tests were performed. For our scaling tests, we have evaluated our algorithms' performance only for weak-scaling. (*i.e.*, scaling the number of processors with a fixed data size per processor). This was chosen over strong scaling (*i.e.*, scaling the number of processors with a fixed total data size) since the latter would require changing a data blocksize to adjust the number of mappers appropriately. The Hadoop/HDFS is known for degraded performance for data with too large or small block sizes depending on job complexity [40], therefore strong scaling is currently not a good indicator of performance in Hadoop. The weak-scaling experiments vary data size against task capacity and proportionally change the number of mappers and reducers. An algorithm that has proper weak scaling should maintain a constant runtime. To avoid biasing results by our optimization schemes, we use the default MapReduce job, with a trivial record reader and writer. Data is stored in binary format and split into 64-byte records, with 16 bytes reserved for the key. Map and reduce functions pass the input directly to the output, and are the simplest possible jobs such that the performance is disk I/O and network transfer bounded.

The top table in Figure 4 shows the average cost for map, shuffle and reduce tasks respectively in the local cluster. The last two columns depict the overall disk I/O throughput and data throughput for a particular job. I/O rates were computed by dividing the total number of disk reads and writes including temporary files over the total time, while data rates represent how much input data pass through the job in a second. For map tasks, Hadoop was able to keep the runtime constant, since input files are read in sequence on each node and directed to appropriate mappers. In the reducing step, even though the amount of data is the same as to the map phase and each node write data to its local disk, there is also a local external sorting that incurs in overhead. Nevertheless, both runtimes are still considerably constant, except for the jump from 64GB to 128GB. At this point, the number of reducers guarantees each node has to host at least two reduce tasks if distributed properly, therefore each disk now has double the I/O and seek operations. This can be seen in the disk I/O rates, where the throughput is optimal at 64 tasks on the local cluster with 60 disks and drops while maintaining a relatively high speed for the larger number of tasks.

The shuffle phase of Hadoop is where weak scaling is not linear. This accounts for the data transfer between map and reduce phases

WEAK-SCALING OF DATASIZE VS. THE NUMBER OF TASKS (on Cluster)

Datasize	#Maps	#Reducers	Map Time	Shuffle Time	Reduce Time	Total Time	I/O Rate	Data Rate
1GB	16	1	7s	18s	27s	63s	84 MB/s	16 MB/s
2GB	32	2	8s	18s	27s	66s	161 MB/s	31 MB/s
4GB	64	4	9s	24s	30s	75s	283 MB/s	55 MB/s
8GB	128	8	10s	26s	29s	78s	545 MB/s	105 MB/s
16GB	256	16	10s	32s	29s	90s	944 MB/s	182 MB/s
32GB	512	32	12s	56s	32s	130s	1308 MB/s	252 MB/s
64GB	1024	64	11s	69s	30s	153s	2222 MB/s	428 MB/s
128GB	2048	128	13s	146s	57s	320s	2125 MB/s	410 MB/s

HADOOP OVERHEAD TIME (on Cluster)

#Maps	#Reducers	Map Only	Total
16	1	15s	30s
32	2	15s	30s
64	4	15s	30s
128	8	15s	30s
256	16	15s	30s
512	32	15s	33s
1024	64	15s	35s
2048	128	15s	36s

WEAK-SCALING (on CLuE)

Datasize	Total Time	I/O Rate	Data Rate
1GB	971s	5 MB/s	1 MB/s
2GB	946s	11 MB/s	2 MB/s
4GB	986s	22 MB/s	4 MB/s
8GB	976s	44 MB/s	8 MB/s
16GB	1059s	80 MB/s	15 MB/s

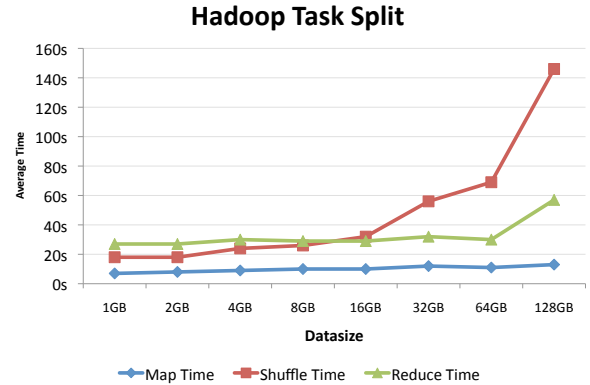
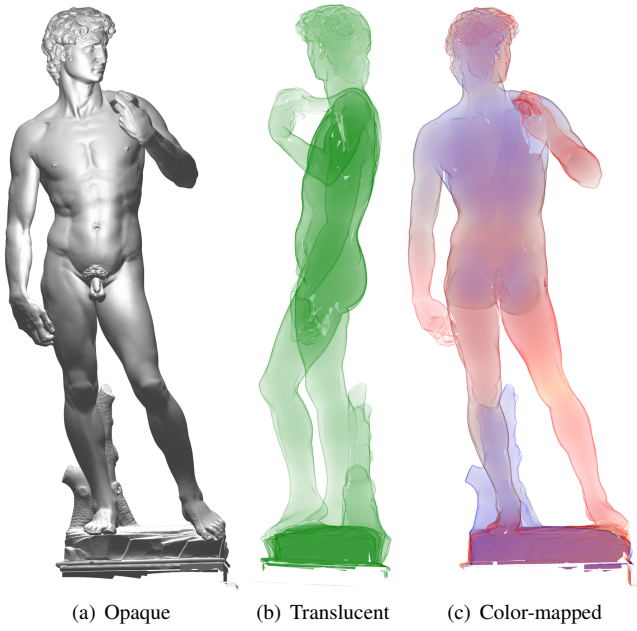


Figure 4: Hadoop baseline test evaluates data transfer costs. In the local cluster we achieve data transfer rates up to 428MB/s

of a job along with sorting. In Hadoop each mapper is likely to contribute a portion of data required by each reducer, and therefore is not expected to scale well. The plot in Figure 4 illustrates the breakdown of the three phases. The *Hadoop Overhead Time* table shows (only) the overhead of communication across the map and reduce phases. Each phase takes about 15 seconds to start on our local cluster. We also include weak-scaling results for the CLuE cluster for comparison. As expected, the I/O rates are considerably lower than the local performance, due to the age of the system and shared utilization. From the weak scaling tests we conclude that the MapReduce model can be robust when the number of nodes scales with the data size. Little cost is incurred for using more input data, and the effective transfer rates scale proportionally to the input data size. However, in order to ensure fault tolerance, disk I/O is heavily involved and could bound the overall performance.



WEAK SCALING (RESOLUTION)

St. MATTHEW (13 GB)					ATLAS (18 GB)				
Resolution	#M/R	CLuE time	Cluster time	File Written	#M/R	CLuE time	Cluster time	File Written	
1.5 MP	256/256	1min 54s	46s	33MB	273/273	1min 55s	46s	41MB	
6 MP	256/256	1min 42s	46s	147MB	273/273	2min 11s	46s	104MB	
25 MP	256/256	1min 47s	46s	583MB	273/273	2min 12s	46s	412MB	
100 MP	256/256	1min 40s	46s	2.3GB	273/273	2min 12s	46s	1.6GB	
400 MP	256/256	2min 04s	46s	10.9GB	273/273	2min 27s	47s	5.5GB	
1.6 GP	256/256	3min 12s	1min08s	53.14GB	273/273	3min 55s	55s	37.8GB	
6.4 GP	256/256	9min50s	2min55s	213GB	273/273	10min30s	1min58s	151.8GB	

WEAK SCALING (RESOLUTION AND REDUCE)

St. MATTHEW (13 GB)					ATLAS (18 GB)				
Resolution	CLuE #R	256M time	Cluster #R	480M time	CLuE #R	256M time	Cluster #R	480M time	
1.5 MP	4	1min 13s	8	46s	4	1min 18s	8	46s	
6 MP	8	1min 18s	15	46s	8	1min 19s	15	45s	
25 MP	16	1min 18s	30	46s	16	1min 51s	30	46s	
100 MP	32	2min 04s	60	47s	32	1min 52s	60	47s	
400 MP	64	2min 04s	120	49s	64	2min 34s	120	46s	
1.6 GP	128	4min 45s	240	1min06s	128	5min 06s	240	55s	
6.4 GP	256	9min 50s	480	2min14s	256	10min 30s	480	1min41s	

DAVID (1 Billion Triangles, 30GB)

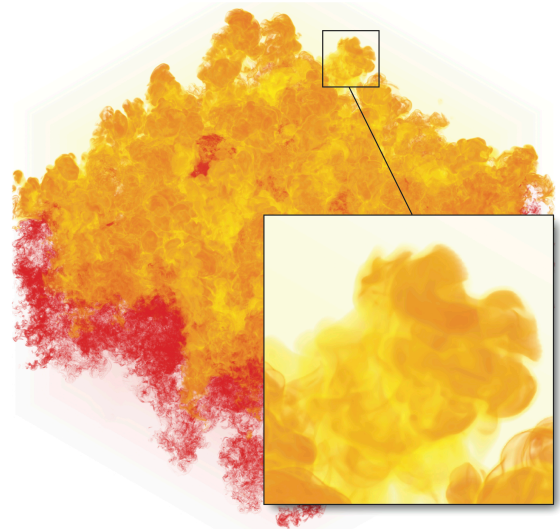
	1.5 MP	6 MP	25 MP	100 MP	400 MP	1.6 GP	6.4 GP
Time	59s	59s	59s	59s	1m 1s	1m 40s	1m 47s

Figure 5: Rendering results for the David statue with different composition stage (a) nearest depth, (b) glass multi-fragment effects [2] and (c) color-mapped glass effect. The model with 1 billion triangles was rendered to a 1 gigapixel canvas in 1 minute and 40 seconds.

5.2 Rendering

Performance of the rasterization algorithm depends on the output resolution, camera parameters and geometry size. The impact of geometry and resolution is proportional to the number of triangles to be rasterized and fragments generated. The impact of camera parameters is hard to estimate since pixels may receive none or multiple fragments depending on the camera position. Hence, reducers may receive none or several fragments to compute depth ordering.

Figure 5(a,b,c) shows rendering results for the Digital Michelangelo Project's David [4] model consisting of 933 million triangles with different composition algorithms: picking the nearest depth fragment, glass multi-fragment effects [2] and color-mapped glass effects. To reduce aliasing effects, our renderer also performs supersampling with an 4×4 kernel, *i.e.*, actually rendering to a 16 times larger canvas.



TETRAHEDRAL MESH VOLUME RENDERING (on Cluster)

Model	#Tetrahedra	#Triangles	Time	#Fragments	Read	Write
Spx	0.8 millions	1.6 millions	3m 29s	9.8 billions	320 GB	473 GB
Fighter	1.4 millions	2.8 millions	2m 20s	5.3 billions	172 GB	254 GB
Sf1	14 millions	28 millions	6m 53s	16.8 billions	545 GB	807 GB
Bullet	36 millions	73 millions	4m 19s	12.7 billions	412 GB	610 GB

STRUCTURED GRID VOLUME RENDERING (on Cluster)

Model	Grid Size	#Triangles	Time	#Fragments	Read	Write
RT27	3072 ³ floats	161 billions	19m 20s	22.2 billions	1.2 TB	1.6 TB

Figure 6: Volume rendering of the Rayleigh-Taylor instability dataset consisting of 27 billion voxels using a 100MP image. Table shows volume rendering statistics for tetrahedral meshes as well.

The tables in Figure 5 reports weak scaling tests with run times and temporary disk usage. For the CLuE cluster, the cost for rendering images of 100MP or less is insignificant compared to the Hadoop overhead. For our local cluster, this threshold is more than 1GP. For images of this size, the cluster is stretched to its limit and performance is limited by the amount of data written to disk. There is a significant increase in the data size due to the large amount of temporary caching by the system due to insufficient buffer space for the shuffling phase. The last table shows the timings to render the David model up to billions of pixels. Compared to [22], a ray tracer, which took 30 hours to complete a frame [36], our MapReduce approach is considerably faster with just under 2 minutes.

Figure 6 shows results of a volume rendering pipeline modified from our surface rendering pipeline. This volume renderer works with both tetrahedral and structured data. Input meshes are broken down into a triangle soup composed of element faces. The reduce phase is modified to perform color, opacity mapping and compositing of the fragments. The accompanying table shows results for a variety of additional meshes. As also shown in the table, for tetrahedral meshes, the most time-consuming image to render at 100MP is not the largest dataset (Bullet) but the earthquake dataset (SF1). This is due to the many large (and flat) tetrahedra that define empty regions at the bottom of the model. Scalar values of these triangles rarely contribute to the final image, but generate a large number of fragments which causes a more expensive shuffle phase.

Figure 6 also shows the rendering of the Rayleigh-Taylor instability dataset consisting of 108GB of structured data. The volume renderer proposed here is a general implementation that is modified from our surface renderer. Its main purpose is to demonstrate the flexibility of our approach. It simply splits each voxel into 12 triangles to be rasterized and composited independently. However, this results in a large amount of primitives to be rasterized for struc-

tured grid. Comparing to a structured-grid specific approach [17] that can volume render a similar dataset in 22 seconds using 1728 cores, ours is slower with roughly 20 minutes on 256 cores. However, we are in fact rendering 161 billion triangles in this case.

5.3 Isosurface Extraction

We tested the isosurface MapReduce algorithm on two datasets: a small ppm Richtmyer-Meshkov instability simulation volume (7.6GB) and another larger simulation, Rayleigh-Taylor dataset (108GB). Since our baseline testing has shown that the amount data produced can effect Hadoop’s performance, we performed tests that varied the number of isosurfaces generated in a single job, since this can have a drastic effect on the amount of data being produced. For few isosurfaces, we expect a small number of triangles to be produced. Conversely, for many isosurfaces, more data will be output to disk than was used for input. We keep the number of maps constant at 256, as this is the largest power-of-two we can use without pigeon holing more than 1 mapper to a node of the CLuE cluster.

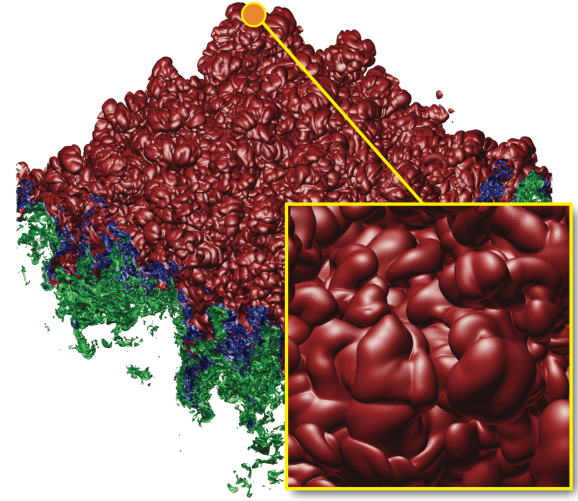
In the extraction of dozens of isosurfaces (as part of parameter exploration) we observed that data output increases proportionally to runtime. Jobs are relatively fast for the standard case of fewer isosurfaces, since input and output data are partitioned evenly to mappers and reducers, thus the amount of disk I/O for input and output is relatively small (*e.g.*, approximately 32MB per Mapper). The runtime is mostly affected by the shuffling phase, where triangles emitted from mappers exceed the available buffer and are sorted out-of-core before being transferred to the reducers. The performance of this phase depends on the amount of temporary disk I/O used when the mapper runs out of in-core buffer space.

The *Written* column in the table of Figure 7 denotes the amount of temporary data produced (not the HDFS output). For both datasets, the algorithm runs quite fast up to 8 isosurfaces, close to Hadoop’s overhead. For 16 isosurfaces, the disk I/O starts to increase abruptly causing the algorithm to slow down. This increase reveals the amount of temporary disk storage needed for the Mappers to sort the data. Figure 7 also shows isosurfaces for the Rayleigh-Taylor dataset. The rendering combined the map task of our isosurface extraction with the map and reduce job of our surface rendering described in Section 5.2. This implementation yields comparable performance to a hand-tuned implementation [21] for extracting isosurfaces from the same datasets: 250 seconds with 64 cores [21] vs. 39 seconds with 256 cores in ours.

5.4 Mesh Simplification

To analyze the out-of-core simplification algorithm in the MapReduce model, we use two large triangle meshes as input: the Atlas statue (18GB) and the St Matthew statue (13GB) from the Digital Michelangelo Project at Stanford University. In these tests, we are interested in seeing the effects of scaling the simplifying grid size. The amount of work done in the Map phase should be very consistent, as each triangle must always compute a quadric measure vector and bin its three vertices. Smaller grid sizes force more vertices to coincide in any particular bin, thus changing the grouping and potentially reducing the parallelism in the Reduce phase.

However, the decrease in performance should be amortized by the decreased output of the Reduce phase, as fewer triangles are generated. In the tables of Figure 8 we observe that this is exactly what occurs. Since our method must be a two pass algorithm in Hadoop, we have included the runtimes for both jobs (Job 1 and Job 2). Rendered images of simplified models of the St Matthew statue are also shown in Figure 9 with the grid sizes varying from 8^3 to 1024^3 . Decimation rates for these results are all under 5% and they were all rendered using the renderer proposed in Section 5.2.



#Iso	Richtmyer-Meshkov (7.6GB)		Rayleigh-Taylor (108GB)	
	Total Time	Written	Total Time	Written
1	30s	1.78GB	39s	8.4GB
2	31s	5.9GB	39s	11.1GB
4	45s	22.5GB	1m 5s	62.0GB
8	45s	52.7GB	1m 25s	155.9GB
16	1m 26s	112.4GB	2m 50s	336.6GB

Figure 7: Isosurface results for varying isovalues using the MapReduce framework for the Richtmyer-Meshkov and Rayleigh-Taylor instability datasets in a local cluster with 256 cores enabled.

Size	St MATTHEW (13 GB)					ATLAS (18 GB)				
	CLuE Time		Cluster Time		Output	CLuE Time		Cluster Time		Output
	Job 1	Job 2	Job 1	Job 2	Size	Job 1	Job 2	Job 1	Job 2	Size
8^3	5m 45s	52s	58s	56s	22 KB	5m 45s	52s	54s	55s	23 KB
16^3	3m 54s	49s	58s	55s	98 KB	3m 54s	49s	54s	54s	105 KB
32^3	3m 51s	49s	55s	54s	392 KB	3m 51s	49s	51s	52s	450 KB
64^3	3m 40s	49s	57s	54s	1.6 MB	3m 40s	49s	55s	55s	1.9 MB
128^3	4m 12s	49s	55s	58s	6.4 MB	4m 12s	49s	52s	52s	7.5 MB
256^3	3m 50s	49s	55s	55s	26 MB	3m 50s	49s	55s	55s	30 MB

Figure 8: Simplification algorithm uses 2 map-reduce jobs. The local cluster processes both datasets in roughly 55s per job.

6 DISCUSSION

In this section, we discuss some of the “lessons learned” from our experience with MapReduce and Hadoop. For users of visualization techniques, it is difficult to know when the results or workload will push beyond the cluster limits and severely increase runtimes. In these cases, increasing the number of maps and reduces may result in a lower memory foot print, yet more balancing, task distribution across all computing nodes. While nodes can run multiple tasks, we find that increasing the number of nodes in proportion to data size provides the most reliable and consistent scalability, suggesting that the overhead to manage additional nodes is not prohibitively expensive.

The results from our exploratory implementations are encouraging and match the scalability we expected, up to a limit. When the size of the output data is unpredictable, as in the case of isosurface extraction, memory requirements can quickly exhaust available resources, leading to disk buffering and ultimately increasing runtime. Scalability, in our experience, is only achieved for data reductive tasks — tasks for which the output is smaller than the input. Most visualization tasks satisfy this property, since they typically render (or generate) data that is smaller than the input mesh or

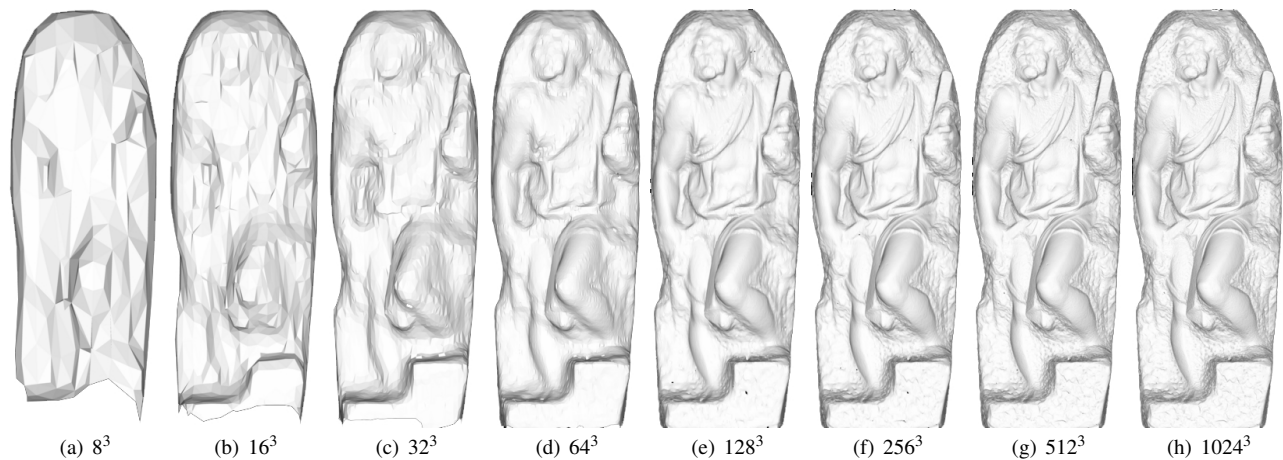


Figure 9: Simplified meshes of the St. Matthew statue using volumes from 8^3 to 1024^3 .

volume. It should also be pointed out that this cost is insignificant when compared to today’s standard practice of transferring data to a client, and running a local serial or parallel algorithm. Indeed, the cost of transferring the data to a local server alone dwarfs the cost of any such MapReduce job.

For those interested in developing visualization algorithms for MapReduce systems, our experience has shown that even naïve implementations can lead to acceptable results. Implementing MapReduce algorithms was relatively simple. However, as with any highly-parallel system, optimization can be painstaking. In the case of MapReduce, we found that the setup and tuning of the cluster itself was just as important, if not more important, than using the right data format, compressor, or splitting scheme.

To analyze the suitability of existing algorithms to the MapReduce model, attention should be paid to where and how often sorting is required. As the model only allows a single sort phase per job, multi-pass algorithms can incur on significant overhead when translated naïvely into MapReduce. Specifically, a MapReduce implementation will rarely be competitive with state-of-the-art methods in terms of raw performance, but the simplicity and generality of the programming model is what delivers scalability and extensibility. Furthermore, the degree of parallelism in the Reduce phase is given by the intended output of the algorithm and data distribution from the Map phase. Also, the hashing method used might have a dramatic effect on the algorithm performance.

Below we summarize our conclusions using the Hadoop system:

- Results from our scaling tests show Hadoop alone scales well, even without introducing optimization techniques;
- Considerations about the visualization output size are very important. Visualization techniques should decrease or keep relatively constant the size of the data in the pipeline rather than increase it. MapReduce was not designed to handle large intermediate datasets, and performs poorly in this context;
- From a qualitative standpoint, we found the MapReduce model easy to work with and implement our solutions. Optimization, in terms of compression and data reader/writers required thought and experimentation. Configuring job parameters and cluster settings for optimal performance was challenging. We feel that this complexity is inherent to a large distributed environment, and therefore is acceptable. Also, it can potentially be performed once per cluster, and the cost can be amortized over many MapReduce jobs;

- The inability to chain jobs makes multi-job algorithms such as the mesh simplification slightly cumbersome to execute, and more difficult to analyze. Projects such as Pig [30] and Hive [41] that offer a high-level yet extensible language on top of MapReduce are promising in this regard;
- The Hadoop community could greatly benefit from better progress reporting. Uneven distribution of data across reducers may result in display of near completion (*e.g.*, 98%) when in fact the bulk of the work remains to be completed. This is problematic if the user does not know *a priori* what a good reducer number should be, and arbitrarily chooses a high value;
- While at any particular time job runtimes are fairly consistent, they vary as a whole from day to day. This is most likely due to the HDFS state and movement of replicated data. Being aware of these effects is important to make meaningful comparisons of performance results. On that note, all data within any one table was generated within a short time span.

7 CONCLUSIONS AND FUTURE WORK

The analysis performed in this paper has shown that the MapReduce model provides a suitable alternative to support large-scale exploratory visualization. The fact that data transfer alone is more expensive than running such a job *in-situ* is sufficient justification, and will become more evident as datasets grow in size. The availability of a core set of visualization tools for MapReduce systems will allow faster feedback and learning from new and large datasets. Additionally, as these systems continue to evolve, it is important for the visualization community to periodically re-evaluate their suitability. We provide a baseline for such a comparative analysis.

We have shown how three visualization techniques can be adapted to MapReduce. Clearly, many additional methods can be adapted in similar ways, in particular memory-insensitive techniques or inherently parallel techniques. What remains to be investigated is how to combine visualization primitives with conventional data management, query, and processing algorithms to construct a comprehensive scalable visual analytics platform.

8 ACKNOWLEDGEMENTS

We would like to thank Marc Levoy at Stanford University for the David model and Bill Cabot, Andy Cook, and Paul Miller at LLNL for the Rayleigh-Taylor dataset. This work was supported in part by the National Science Foundation (CCF-08560,

CCF-0702817, CNS-0751152, CNS-1153503, IIS-0844572, IIS-0904631, IIS-0906379, IIS-1153728, and NIH ITKv4), the Department of Energy, CNPq (processes 200498/2010-0, 569239/2008-7, and 491034/2008-3), IBM Faculty Awards and NVIDIA Fellowships. This work was also performed under the auspices of the U.S. Department of Energy by the University of Utah under contract DE-SC0001922 and DE-FC02-06ER25781 and by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344, LLNL-JRNL-453051.

REFERENCES

- [1] Amazon web services - elastic mapreduce. <http://aws.amazon.com/elasticmapreduce/>.
- [2] L. Bavoil, S. P. Callahan, A. Lefohn, J. a. L. D. Comba, and C. T. Silva. Multi-fragment effects on the gpu using the k-buffer. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, I3D '07, pages 97–104, New York, NY, USA, 2007. ACM.
- [3] D. Borthakur. The Hadoop distributed file system: Architecture and design. http://lucene.apache.org/hadoop/hdfs_design.pdf, 2007.
- [4] B. Brown and S. Rusinkiewicz. Global non-rigid alignment of 3-D scans. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 26(3), Aug. 2007.
- [5] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope: easy and efficient parallel processing of massive data sets. In *Proc. of the 34th Int. Conf. on Very Large DataBases (VLDB)*, pages 1265–1276, 2008.
- [6] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: a distributed storage system for structured data. In *Proc. of the 7th USENIX Symp. on Operating Systems Design & Implementation (OSDI)*, 2006.
- [7] Nsf cluster exploratory (nsf08560). <http://www.nsf.gov/pubs/2008/nsf08560/nsf08560.htm>.
- [8] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In *Proc. of the 6th USENIX Symp. on Operating Systems Design & Implementation (OSDI)*, 2004.
- [9] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *CACM*, 51(1):107–113, 2008.
- [10] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon’s highly available key-value store. In *Proc. of the 21st ACM Symp. on Operating Systems Principles (SOSP)*, pages 205–220, 2007.
- [11] D. J. DeWitt, E. Paulson, E. Robinson, J. Naughton, J. Royalty, S. Shankar, and A. Krioukov. Clustera: an integrated computation and data management system. In *Proc. of the 34th Int. Conf. on Very Large DataBases (VLDB)*, pages 28–41, 2008.
- [12] R. Farias and C. T. Silva. Out-of-core rendering of large, unstructured grids. *IEEE Comput. Graph. Appl.*, 21(4):42–50, 2001.
- [13] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.
- [14] Hadoop. <http://hadoop.apache.org/>.
- [15] G. Heber and J. Gray. Supporting finite element analysis with a relational database backend; part 1: There is life beyond files. Technical report, Microsoft MSR-TR-2005-49, April 2005.
- [16] Hive. <http://hadoop.apache.org/hive/>. Accessed 3/7/2010.
- [17] M. Howison, W. Bethel, and H. Childs. Mpi-hybrid parallelism for volume rendering on large, multi-core systems. In *EG Symposium on Parallel Graphics and Visualization (EGPGV'10)*, 2010.
- [18] IBM Systems and Technology Group. IBM Deep Computing. Technical report, IBM, 2005.
- [19] Incorporated Research Institutions for Seismology (IRIS). <http://www.iris.edu/>.
- [20] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: Distributed data-parallel programs from sequential building blocks. In *Proc. of the European Conference on Computer Systems (EuroSys)*, pages 59–72, 2007.
- [21] M. Isenburg, P. Lindstrom, and H. Childs. Parallel and streaming generation of ghost data for structured grids. *Computer Graphics and Applications, IEEE*, 30(3):32–44, 2010.
- [22] T. Ize, C. Brownlee, and C. Hansen. Real-time ray tracer for visualizing massive models on a cluster. In *EG Symposium on Parallel Graphics and Visualization (EGPGV'11)*, 2011.
- [23] Kosmix Corp. Kosmos distributed file system (kfs). <http://kosmosfs.sourceforge.net>, 2007.
- [24] Lawrence Livermore National Laboratory. VisIt: Visualize It in Parallel Visualization Application. <https://wci.llnl.gov/codes/visit> [29 March 2008].
- [25] P. Lindstrom and C. T. Silva. A memory insensitive technique for large model simplification. In *VIS '01: Proceedings of the conference on Visualization '01*, pages 121–126, Washington, DC, USA, 2001. IEEE Computer Society.
- [26] W. Lorensen and H. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(4):163–169, 1987.
- [27] Large Synoptic Survey Telescope. <http://www.lsst.org/>.
- [28] K.-L. Ma. In situ visualization at extreme scale: Challenges and opportunities. *Computer Graphics and Applications, IEEE*, 29(6):14 – 19, nov.-dec. 2009.
- [29] Azure Services Platform - SQL Data Services. <http://www.microsoft.com/azure/data.mspx>.
- [30] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *SIGMOD'08: Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 1099–1110, 2008.
- [31] Paraview. <http://www.paraview.org> [29 March 2008].
- [32] B. Paul, S. Ahern, E. W. Bethel, E. Brugger, R. Cook, J. Daniel, K. Lewis, J. Owen, and D. Southard. Chromium Renderserver: Scalable and Open Remote Rendering Infrastructure. *IEEE Transactions on Visualization and Computer Graphics*, 14(3), May/June 2008. LBNL-63693.
- [33] A. Pavlo, E. Paulson, A. Rasin, D. J. Abadi, D. J. DeWitt, S. R. Madden, and M. Stonebraker. A comparison of approaches to large scale data analysis. In *SIGMOD*, Providence, Rhode Island, USA, 2009.
- [34] T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper. Virtual network computing. *IEEE Internet Computing*, 2(1):33–38, 1998.
- [35] W. J. Schroeder, J. A. Zarge, and W. E. Lorensen. Decimation of triangle meshes. In *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 65–70, New York, NY, USA, 1992. ACM.
- [36] U. o. U. SCI Institute. One billion polygons to billions of pixels. <http://www.sci.utah.edu/news/60/431-visus.html>.
- [37] Silicon Graphics Inc. OpenGL vizserver. <http://www.sgi.com/products/software/vizserver>.
- [38] Sloan Digital Sky Survey. <http://cas.sdss.org>.
- [39] S. Stegmaier, M. Magallón, and T. Ertl. A generic solution for hardware-accelerated remote visualization. In *VISSYM '02: Proceedings of the symposium on Data Visualisation 2002*, pages 87–ff, Aire-la-Ville, Switzerland, Switzerland, 2002. Eurographics Association.
- [40] I. Technologies. Hadoop performance tuning - white paper.
- [41] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive - a warehousing solution over a map-reduce framework. *PVLDB*, 2(2):1626–1629, 2009.
- [42] Yahoo! Research. PNUTS - Platform for Nimble Universal Table Storage. <http://research.yahoo.com/node/212>.
- [43] J. Yan, P. Shi, and D. Zhang. Mesh simplification with hierarchical shape analysis and iterative edge contraction. *IEEE Transactions on Visualization and Computer Graphics*, 10(2):142–151, 2004.
- [44] H. Yang, A. Dasdan, R.-L. Hsiao, and D. S. Parker. Map-reduce-merge: simplified relational data processing on large clusters. In *SIGMOD'07: Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 1029–1040, 2007.
- [45] Y. Yu, M. Isard, D. Fetterly, M. Budiu, U. Erlingsson, P. K. Gunda, and J. Currey. DryadLINQ: A system for general-purpose distributed data-parallel computing using a high-level language. In *Proc. of the 8th USENIX Symp. on Operating Systems Design & Implementation (OSDI)*, 2008.