CompStor Analytics™

Scientific Computing on Large Life Science Datasets using a Tiered-Memory Architecture with Commodity Servers

A tiered memory architecture analyzes multi-terabyte datasets, performing singular value decomposition (SVD), principal component analysis (PCA) and least-squares (LSQR) solutions to terabyte-scale matrices at a fraction of the cost of supercomputing.

Executive summary

The fields of genomics and proteomics produce data at exponential rates and in increasingly distributed settings, posing a computational challenge. Local compute resources are often inadequate and supercomputing facilities inaccessible. Cloud based solutions suffer from long data ingress and egress times. Finally, hardware accelerators can be difficult to deploy and lack scalability. CompStor Analytics™ is a scientific computing solution utilizing standard commercial-off-the-shelf (COTS) servers with a novel tiered-memory configuration. Performance greatly exceeds state-of-the-art, out-of-memory solutions for important analytical techniques such as SVD, PCA and LSQR. The datasets analyzed here include DNA sequence variants, epigenetic methylation, imaging mass spectrometry, and computed tomography.

CompStor Analytics™ is designed for terabyte plus datasets—those considered too large for local, non-supercomputing analyses. Its key innovation is optimal data distribution and access across non-volatile memory express (NVMe) solid state drives (SSD) and dynamic random-access memory (DRAM) in a multi-threaded configuration, driving down run time. For example, a single CompStor® node outperforms a four-node Spark configuration by a

factor of at least five in singular value decomposition (SVD) analysis (Figure 1).

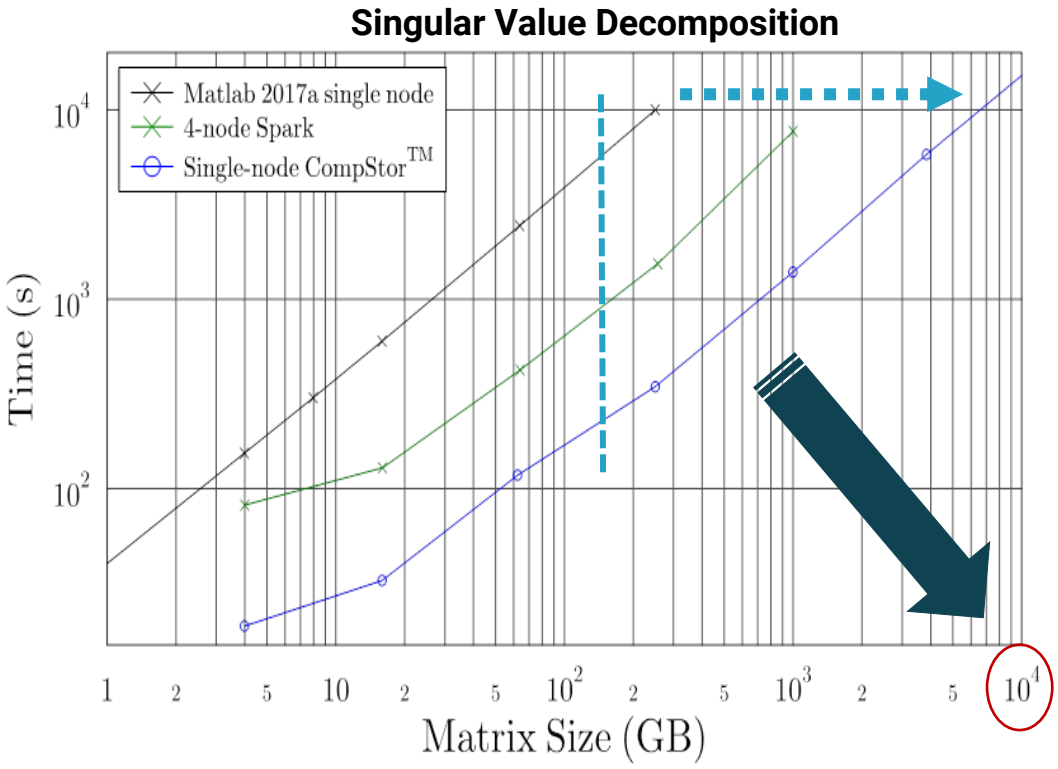


Figure 1. *Single-node CompStor Analytics™ outperforms both MatLab and 4-node Spark configurations on singular value decompositions of large data. MatLab failed to complete analysis over 200GB of data and Spark failed with over one terabyte of data. The down arrow indicates the effect of adding multiple nodes to a CompStor® system; a multi-noded CompStor® cluster solves multi-terabyte problems in less than one hour.*

Life science datasets

This paper features analysis of four *big data* life sciences applications: genome sequencing variants; epigenomic methylation; 3D imaging mass spectrometry; and, 3D tomography. CompStor Analytics performs PCA on the first three of the datasets and LSQR solutions to linear equations on a simulated tomography dataset. We demonstrate CompStor Analytics on data too large for analysis with conventional solutions, which are often restricted to the size of their installed DRAM. Where necessary, we extend the input data into the TB size range to show that much larger analysis is possible.

1. **PCA on genomic variants**
   PCA is performed on the 1,000 Genomes dataset of 84 million genetic

variants (e.g. insertions, deletions, single nucleotide polymorphisms) over a global sample of 2,504 individuals.[1]

2. **PCA on genomic methylation**
   PCA is performed on The Cancer Genome Atlas (TCGA) Pan-Cancer dataset of 396,066 methylation site β-values over a population of 9,821 individuals with 33 different cancer types.[2]

3. **PCA on 3D MALDI imaging mass spectrometry**
   PCA is performed on three-dimensional matrix assisted laser desorption ionization imaging mass spectrometry (3D MALDI IMS) data generated from 75 mouse kidney sections. This data represents a mass-to-charge (m/z) spectrum of 7,680 m/z values over 1,362,830 points throughout the mouse kidney.[3]

4. **LSQR on simulated 3D tomographic data**
   A least-squares solution is performed on three-dimensional tomographic data generated from a digital 3D phantom.

	Genomic Variants	DNA Methylation	MALDI IMS	Tomography
Data size (TB)	3.9	2.8	3.8	18.0
Matrix elements (billion)	545	389	1047	1500
Principal components	25	10	40	N/A
Algorithms	PCA, t-SNE	PCA, t-SNE	PCA	LQSR
CompStor run time (minutes)	76	59	390	120 (8 nodes)

Table 1. *Dataset size and run time information for four life science examples. Data set sizes have been extended, where necessary, from the available datasets to achieve TB scale. All CompStor® computations are done on a single node except where noted.*

Architecture

CompStor Analytics successfully performs PCA and LSQR on datasets up to the capacity of installed SSDs and has been demonstrated on problems up to 24 TB in size. This is possible through a multi-threaded and tiered-memory architecture utilizing a proprietary memory management scheme. A configuration of DRAM and NVMe SSD storage optimizes both effective memory size and data transfer speeds in a multi-threaded environment. With this optimized architecture, big-data scientific computing problems requiring extreme amounts of effective in-line memory can be performed on COTS equipment at a fraction of the cost of equivalent super-computing. Furthermore, large data-transfer times for cloud-centric solutions can be avoided.

Conventional tools such as FlashPCA, MatLab, Python libraries, and Spark, are limited in their capacity to utilize multi-threaded and tiered-memory architectures for PCA and LSQR solutions. These methods cannot effectively process terabyte sized datasets.

CompStor's server architecture allows for both single-node and multi-node configurations. Using a high-speed network connection from a data source to the CompStor cluster, data can be quickly distributed with minimal impact on total-problem performance. The life-science datasets analyzed in this paper are summarized in Table 1 and include run times and memory footprints. These data were analyzed using pre-release CompStor nodes equipped with a NEC 2-socket Intel E5-2699 Xeon V4 (2.2 GHz), 22 hyper-threaded cores per socket, 512GB DRAM, and 6 x 3.2 TB NVMe SSDs. Commercial CompStor nodes will be more powerful yet, with a 4-socket Intel Xeon® Gold 6148, 20 cores per socket, 768 GB DRAM; and high-performance SSDs exhibiting 6GB/s sequential read performance.

PCA and t-SNE analysis

Principal component analysis (PCA) is a valuable tool to reduce the dimensionality of large datasets and investigate significant relationships in between related members of a sample population[4]. This numerical technique utilizes transformations on a set of possibly correlated features across a sample population into a set of orthogonal vectors, or "principal components." The components are calculated in descending order of their capacity to describe the variance in the data. It is often possible to describe much of the data with relatively few principal components. CompStor leverages the results of a distributed and multi-threaded singular value decomposition algorithm to produce its results.

In addition to PCA on CompStor, we can further reduce dimensionality using multi-core t-distributed stochastic neighbor embedding (t-SNE).[5] As data visualization is most accessible in two to three dimensions, t-SNE is useful for representing relationships in high-dimensional data. The multicore-TSNE Python library[6] is used to take advantage of the available memory, the number of cores, and the reduced dimensionality of the PCA-processed data.

For the genomic variants, cancer methylation, and 3D MALDI mass spectrometry data, the PCA is performed for varying number of components. The genomic variants and cancer methylation datasets most readily portray distinct clusters in t-SNE visualizations, because they represent known discrete groupings, continental populations, and cancer types.

Global-scale genetic variants

This demonstration of CompStor Analytics PCA analyzes autosomal genetic variants from the 1,000 Genomes Project. This dataset represents 84 million distinct genetic variants discovered from the autosomes (chromosomes 1-22) of 2,504 individuals subject to whole genome sequencing. These 2,504 individuals are sampled across 29 sub-populations of five continental populations. The dataset is read from twenty-two variant calling format (VCF) files in an 800 GB file converted to CompStor's sparse, binary matrix format. The sparsity and internal representation of this binary matrix reduces the footprint of this dataset to 48 GB.

For the above dataset, CompStor Analytics generates the vectors and weights for ten principal components within eight minutes. The results of twenty-five principal components are produced within sixteen minutes. The first ten components account for 12.7% of the variance in the data while an additional fifteen only describes 13.5% of the variance in the

data. After applying further dimensional reduction using the t-SNE algorithm, we observe distinct relationships between continental populations and their sub-populations (Figure 2).

Assuming appropriate hardware, the above data can be analyzed using available methods. To demonstrate the capability of CompStor Analytics, the dataset is replicated 50 times to create a matrix with 84 million rows representing individual sequence variants and 125,200 columns representing a 50 fold expansion in the original 2,504 subjects. Although the resulting 3.9 TB sparse matrix cannot fit into memory, CompStor Analytics performs PCA within 77 minutes.

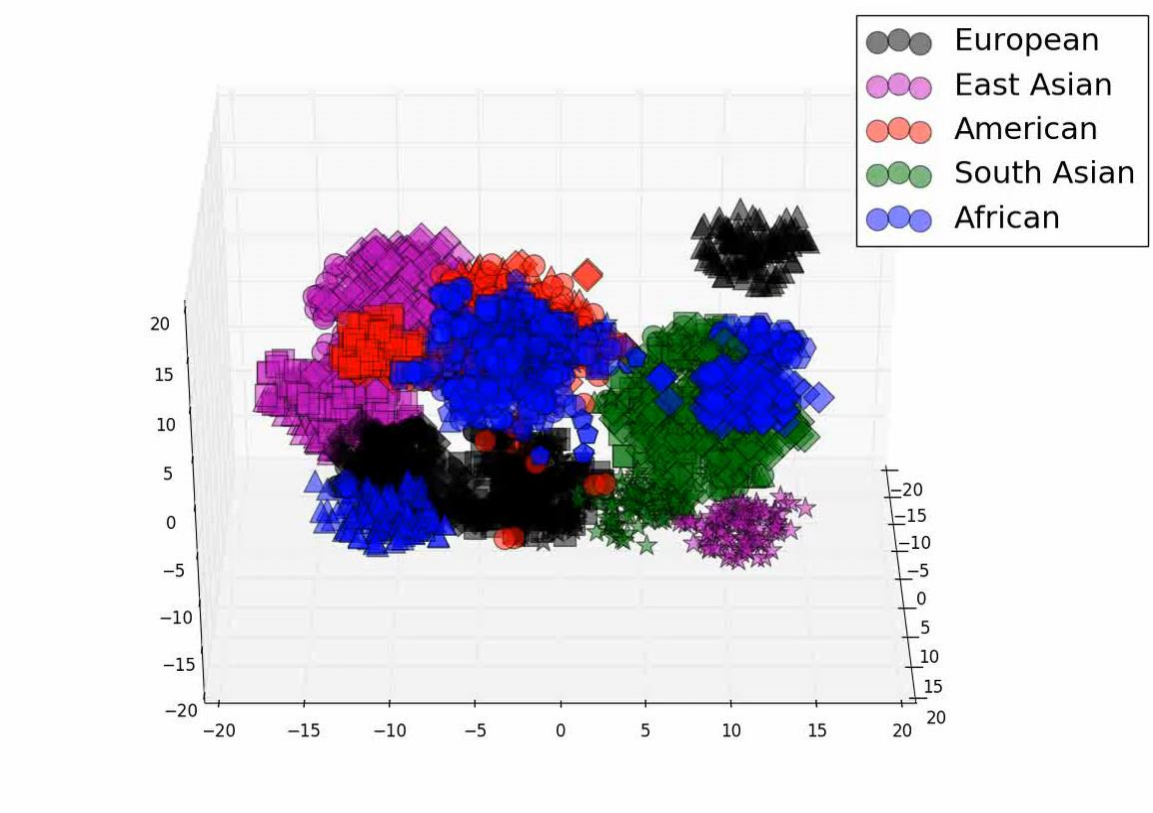


Figure 2. *10 PCAs visualized with a 3D t-SNE mapping 2,504 subjects with 84M autosomal variants. Matrix represents over 10.9 billion elements completed on 1-node CompStor® cluster in 7.5 minutes. Continental populations and subpopulations are colored and shaped.*

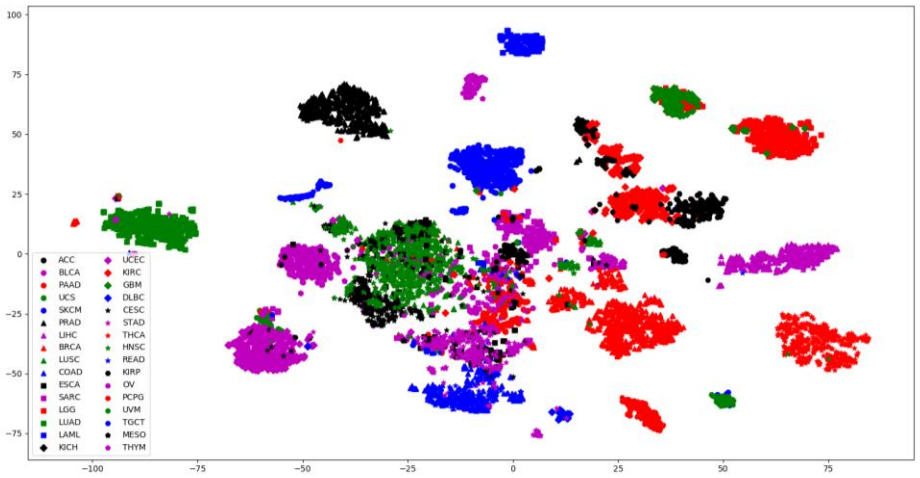
Methylation of cancer genomes

The epigenetic pattern of DNA methylation greatly influences cellular development and oncogenesis. Both somatic mutations and methylation patterns contribute to cancer aggression and treatment resistance. In fact, observing methylation patterns is used in the detection and characterization of cancer. The Cancer Genome Atlas (TCGA) consortium has made the methylation patterns across 33 prevalent cancers publicly available within the Pan-Cancer Atlas. This dataset consists of the methylation β-values across 396,066 CpG sites for 9,821 cancer patients. Methylation β-values measure the degree of methylation across the sites on affected chromosomal alleles.

CompStor Analytics generates twenty principal components within 1 minute. Twenty principal components are enough to describe 51% of the variation for this dataset. Distinct patterns of methylation are accentuated for several cancer types after applying t-SNE algorithm on the principal component scores for the patient population (Figure 3a). Focusing on populations with glioblastoma multiforme (GBM) and low-grade glioma (LGG) we can observe methylation signatures lead to distinct clusters (Figure 3b). Observed cross-cluster placement and outliers illuminate the evolving nature of cancer diagnosis. With further sampling of tissues from a control population, distinct differential methylation patterns may be even more striking.

As with the genetic variants above, we augment this data set by a factor of 100. This yields a matrix for 982,100 subjects in a 2.8 TB dense matrix. CompStor Analytics performs a PCA on this augmented matrix within 1 hour.

a)



b)

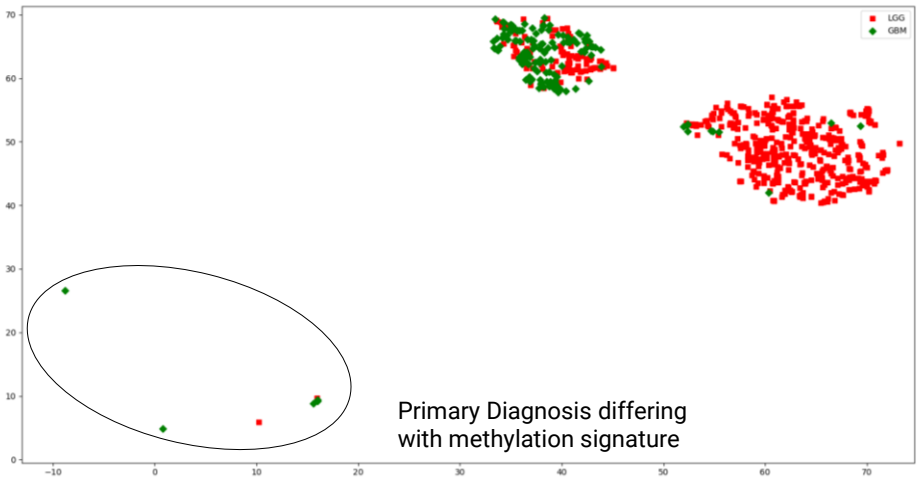


Figure 3. *2-dimensional t-SNE of the first 25 principal components of DNA methylation data in cancer. (a) The 33 cancer origins are colored, showing differential methylation by cancer type or tissue of origin. (b) focuses in on low-grade glioma (LGG) and glioblastoma multiforme (GBM) demonstrating how primary diagnosis can disagree with methylation signatures.*

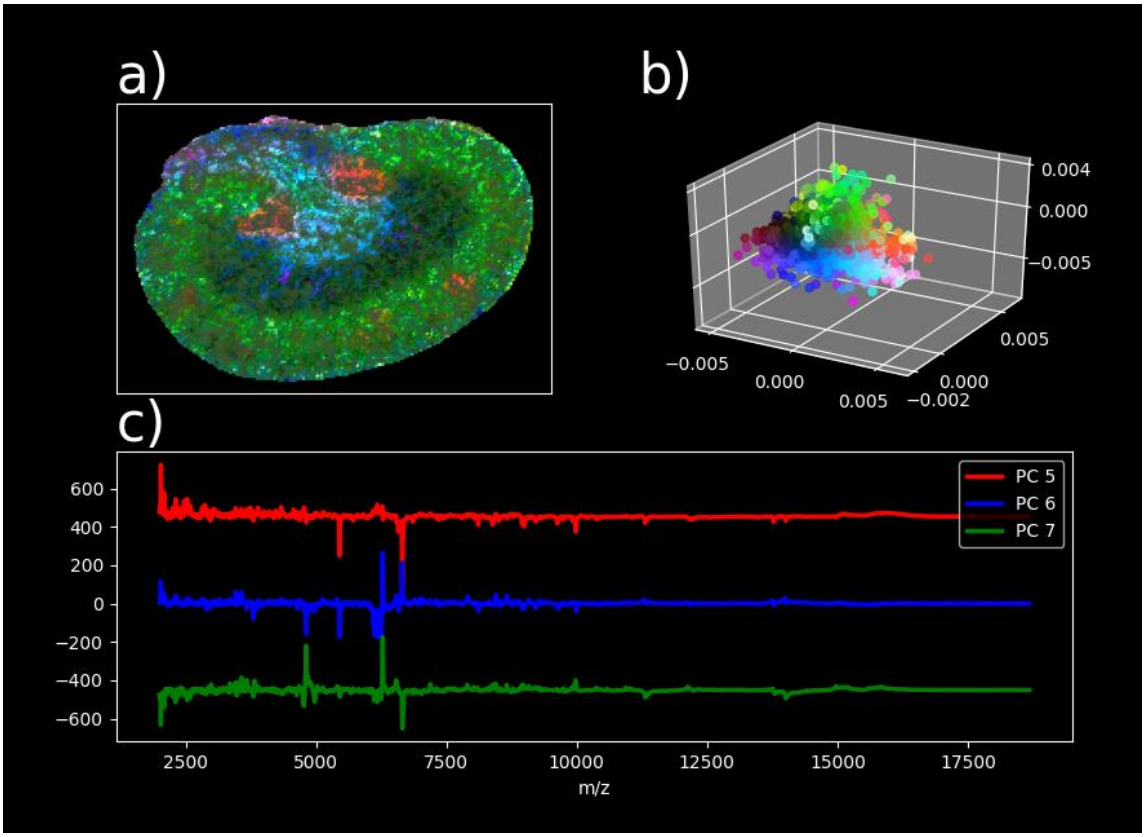
3D MALDI IMS of a mouse kidney

3D MALDI IMS is a mass spectrometry technique characterizing the three-dimensional mass spectrum of, for example, an anatomical specimen using sequential tissue section sampling. It combines specialized MALDI ionization with time-of-flight mass spectrometry to identify proteins or small molecules in their spatial orientation. Composing the results of MALDI MS throughout all tissue sections of an organ reveals the three-dimensional distribution of chemical components representing wild type, disease or other experimental conditions.

To facilitate reproducible research, benchmark datasets for 3D MALDI MS imaging have been utilized. One dataset comprises 75 sections from the central portion of a mouse kidney. The 7,680 mass-to-charge ratios, in the range of 2,000-20,000 m/z, were acquired for 1,362,830 spectra throughout the mouse kidney.

CompStor Analytics produced 40 principal components of the above dataset, within 8 minutes. The scores and basis vectors of these components describe 83.7% of the variation of the data. The scores for principal components 5,6, and 7 are colored by red, green, and blue intensities and plotted back into the original slice of the mouse kidney. The distinct anatomy of the renal cortex (green), the renal medulla (dark green), the renal pelvis (blue), and the surrounding of the renal pelvis (red) are clearly distinguishable. The component basis vectors are shown in their capacity to increase (positive) or reduce (negative) the spectra at distinct mass-to-charge ratios.

Once more, as above, we extend this dataset by a factor of 100 to 136,283,000 spectral samples. CompStor Analytics performs a 40-component PCA on this 3.9 TB dense matrix in 6 hours.



For Research Use Only. Not for use in Diagnostic Procedures.

Figure 4. *The colors of MALDI MS sampling of a mouse kidney slice (a) are drawn from the distribution of PCA coordinates (b) in a 3D component space expressing linear combinations of the basis vectors (c). Anatomical features of the mouse kidney(a)—renal cortex (green), renal medulla (dark green), renal pelvis (blue), and surrounding renal pelvis (red)—are clearly distinguishable. Three PCA basis vectors (c), represent their capacity to increase (positive) or decrease (negative) the overall signal at distinct mass-to-charge ratios.*

LSQR and tomography

Generalized tomography techniques generate images from generalized one-dimensional projections (or, "measurements") of the image. Unlike standard tomographies such as computed tomography (CT), generalized tomography cannot rely on regular sampling techniques, for which formulaic Radon transforms may exist. In such cases, the image can be recovered by writing each projection's equation and solving the resultant linear system for the intensity of each voxel.

We simulate a 3D generalized tomography problem by generating random sparse projections of a 3D Shepp-Logan phantom of dimension 1000x1000x1000 voxels. Each random projection is sparse, containing contributions from 1000 voxels. We note this method is mathematically reasonable but not physically implementable; we merely seek a method to quickly generate a solvable 3D tomography problem. A real 3D tomography would specify some other version of the projections making up the measurement. The CompStor analysis portion of this simulation and a real system would exhibit similar complexities.

Figure 5 shows the results of a 1000x1000x1000 imaging problem, exhibiting 1 billion voxels and 1.5 billion equations representing 1.5 billion projections. The size of the resulting sparse matrix is 18 TB. This problem can be solved by an 8-node CompStor cluster in less than 2 hours.

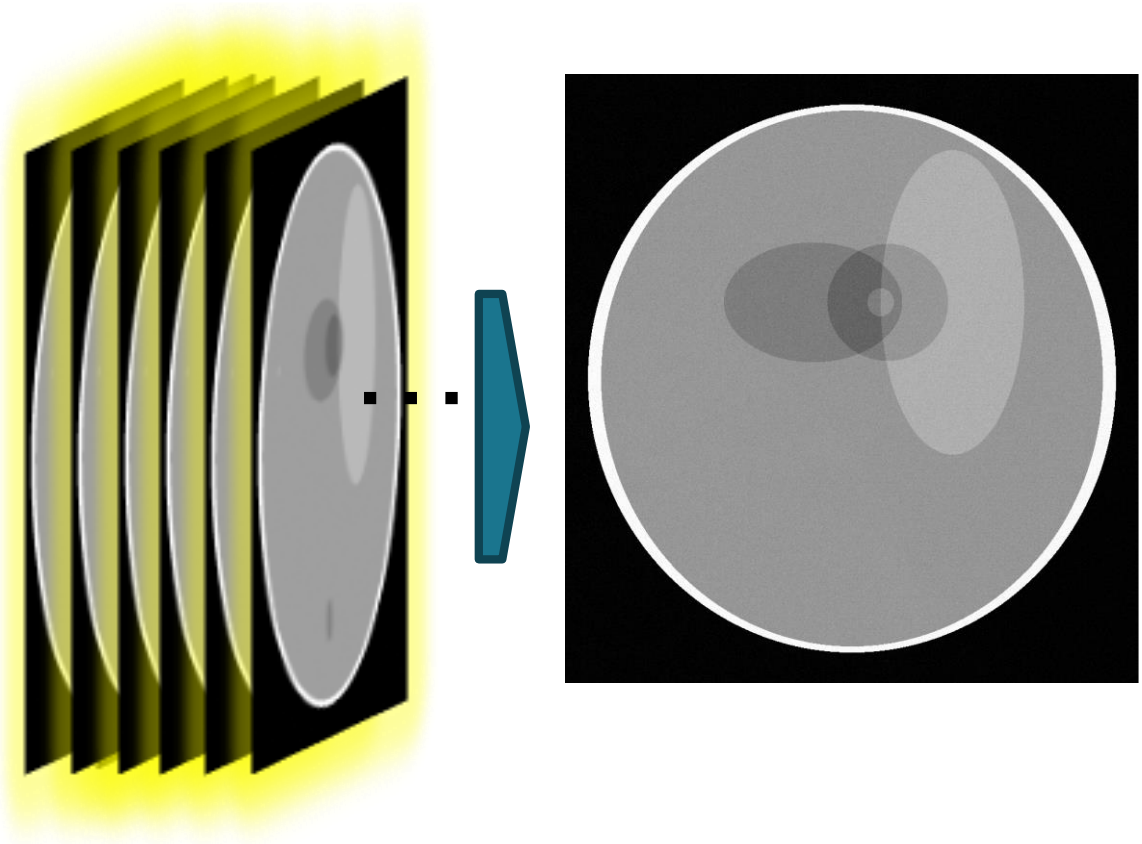


Figure 5. *Results of a 3D generalized tomography problem with 1 billion voxels.*

CompStor® appliance

A typical CompStor Analytics configuration is shown in Figure 6. It features extensive SSD memory capacity and an optimized, high speed ingress for data.

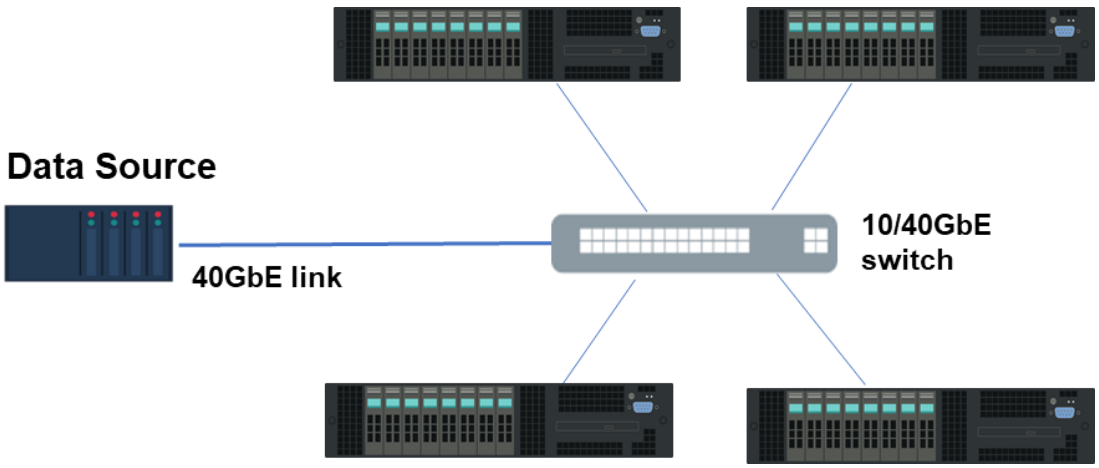


Figure 6. *A 4-node, CompStor® compute cluster with optimized high speed data ingress and multithreading.*

Conclusion

CompStor Analytics enables fast run times for matrix computation on very large datasets, solving problems in genomics, proteomics, metabolomics, and tomography. Data sets up to tens of terabytes in size can be analyzed on CompStor clusters. The multi-node features of CompStor allow users to run even larger problems in faster times, without invoking the cost penalties of a supercomputing system.

About OmniTier

OmniTier Inc., founded in 2015, develops and supports integrated software solutions for memory-centric (Principal component analysis, 2010) infrastructure applications, including high performance object caching, scientific analysis for machine learning, AI, and genomics. Its leadership team has a track record of delivering many industry firsts in data storage and access across different media types. The company has offices in Milpitas, California, and Rochester, Minnesota.

References

[1] The 1000 Genomes Project Consortium. (1 October 2015). A global reference for human genetic variation. Nature, Nature 526, 68-74. Retrieved from The 1000 Genomes Project Consortium: http://www.internationalgenome.org

[2] Human methylation results shown are in whole or part based upon data generated by the TCGA Pan-Cancer Atlas: https://cancergenome.nih.gov/newsevents/newsannouncements/pancancer_atlas.

[3] Oetjen, J. (2015). Supporting materials for "Benchmark datasets for 3D MALDI- and DESI-Imaging Mass Spectrometry.". GigaScience Database, BMC, 4:20. Retrieved from http://dx.doi.org/10.5524/100131.

[4] Principal component analysis. (2010). In H. &. Abdi, Wiley Interdisciplinary Reviews: Computational Statistics (pp. 433-459).

[5] Hinton, L. v. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 2579-2605.

[6] Ulyanov, D. (2016). https://github.com/DmitryUlyanov/Multicore-TSNE. Retrieved from Github.