Harnessing cloud computing for high capacity analysis of neuroimaging data

Daniel J. Clark[1], Christian Haselgrove[1], David N. Kennedy[1], Zhizhong Liu[1], Michael Milham[1,1], Petros Petrosyan[1], Carinna M. Torgerson[1], John D. Van Horn[1], R. Cameron Craddock[1,1,]

[a] *Center for the Developing Brain, Child Mind Institute, New York, New York, USA*
[b] *Division of Informatics, Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA*
[c] *The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI), Keck School of Medicine of USC, University of Southern California, Los Angeles, CA, USA*
[d] *Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, USA*
[e] *Computational Neuroimaging Laboratory, Center for Biomedical Imaging and Neuromodulation, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, USA*

Abstract

Functional connectomes capture brain interactions via synchronized fluctuations in the functional magnetic resonance imaging signal. If measured during rest, they map the intrinsic functional architecture of the brain. With task-driven experiments they represent integration mechanisms between specialized brain areas. Analyzing their variability across subjects and conditions can reveal markers of brain pathologies and mechanisms underlying cognition. Methods of estimating functional connectomes from the imaging signal have undergone rapid developments and the literature is full of diverse strategies for comparing them. This review aims to clarify links across functional-connectivity methods as well as to expose different steps to perform a group study of functional connectomes.

Keywords: Functional connectivity, connectome, group study, effective connectivity, fMRI, resting-state

1. Introduction

Text from CPAC grant: Conventional neuroimaging tools alone have insufficient capacity for performing big data analyses in real-world contexts. Using current best practices, processing a single fMRI dataset on conventional computer workstations can take almost 5 hours. Using tools that do not permit automated processing, 1000 datasets will take 5,000 hours of time, or 2.5 man-years, to prepare for processing. With sufficient programming skills, an analyst can construct an automated pipeline to process the data unattended in 5,000 hours or nearly 7 months. If the analyst is more sophisticated and can parallelize the processing using multi-core or cluster computer systems, and has access to adequate resources, this processing can be reduced to anywhere between a few hours and 2 months. C-PAC makes these levels of processing speedups accessible to all neuropsychiatric researchers, without requiring the advanced programming background necessary to implement them. Additionally C-PAC offers other improvements over conventional tools, such as the ability to restart a pipeline with different parameters and only re-computing the affected stages. All of these innovations make it possible for researchers to perform large-scale data analyses

*Corresponding author

in reasonable amounts of time. The lack of access to sufficient computational systems needlessly bars many researchers, who would otherwise be able to make substantial contributions to the field, from performing big data analyses. The cost of purchasing and maintaining high performance computing systems is a significant barrier for many researchers. Since most neuroimaging researchers only periodically need computational resources, but require fairly large resources when they do, they can benefit substantially from pay-as-you-go computational services such as the Amazon Web Services (AWS) Elastic Compute Cloud (EC2).

To meet this demand, C-PAC has been ported to EC2 under a contract from National Database of Autism Research (NDAR; see letter of support from NDAR manager Dan Hall). The proposed work will extend this capability to a Software-as-a-Service (SaaS) model that makes the entire C-PAC ecosystem available over the Internet using an execution dashboard (thin client) running in a web browser. This approach simplifies the cloud-computing model so that it is readily accessible to the large audience of researchers who lack computational resources and knowledge to provision a system in the cloud. Pushing the envelope for what is possible in connectome analyses requires taking advantage of new computing technologies. Exploiting the lack of data dependencies between datasets, pipelining systems have been able to achieve substantial reductions in processing times by running several pipelines in parallel on multicore and cluster computer systems. An advantage of this approach is that it can use pre-existing software tools that are designed to work with conventional computer processors (CPUs). A disadvantage is that it is costly to scale, as an incremental increase in performance will require a similar increase in hardware. Enabling even larger scale computation, such as higher resolution graph analyses, and non-parametric statistics that require thousands of iterations, will require taking advantage of new computing technologies such as Graphical Processing Units (GPU). When matched by computational power, GPUs are significantly cheaper than conventional CPUs in both cost and energy consumption 21,22. Although, there have been a number of reports lauding the ability of GPU implementations of neuroimaging algorithms to achieve 195x speedups for nonlinear registration23, 33x speedups for permutation testing24, 6x speedups for surface extraction25, 100x speedups for diffusion tractography26, and 250x speedups for computing functional connectivity27, the use of GPUs in this field has yet to become widespread. This is due to the lack of off-the-shelf tools that support GPU architectures, and the specialist knowledge required to develop software for these architectures.

As a part of the proposed improvements to C-PAC, high scale graph theoretic methods and non-parametric statistical group-level analyses will be implemented using GPU architectures. This advance in C-PAC will make the sophisticated analyses enabled by the higher throughput of GPUs accessible to the wider community of neuroimaging researchers. By implementing C-PAC as a Software-as-a-Service in Amazons EC2, it creates the first-of-its-kind, pay-as-you-go solution for performing big data science on neuroimaging data. Other neuroimaging centric Amazon Machine Images exist, allowing users with advanced computational background to process large datasets on clusters provisioned in the cloud. C-PAC will be the first to do it using thin clients that simplify all of the technical implementation details inherent in this process. To a user, it will look no different than using a web application. The Cloud Computing Model: To make large-scale processing available to researchers without access to computational resources, C-PAC has been ported to the Amazon Web Services (AWS) Elastic Compute Cloud (EC2) under a contract from NDAR (see letter from NDAR manager Dan Hall). Using a freely available preconfigured C-PAC Amazon Machine Image (AMI, a virtual machine), users can quickly provision a computing system and preprocess data without having to deal with software installation or hardware configuration. Distributing a standardized software ecosystem avoids errors related to incompatible or unsupported software versions and permits pipelines to be exactly reproduced down to the fine details. Data can either be transferred into and out of EC2 using conventional network transfer mechanisms (i.e., scp)

or C-PAC can be configured to access data from the AWS Simple Storage Service (S3). The C-PAC AMI supports the StarCluster cluster computing toolkit to allow users to easily create computing clusters in EC2 that can be dynamically rescaled based on the users needs. The C-PAC team has used this AMI to perform cortical thickness extractions on nearly 3,060 structural scans from the NDAR database (doi:10.15154/1165646), and recently preprocessed all 1,102 datasets from ABIDE in approximately 63 hours. Until now the enthusiasm for performing neuroimaging analysis in EC2 has been tempered by the lack of clear models that illustrate the cost and performance tradeoffs. Based on lessons learned from the aforementioned preprocessing efforts, a comparison of processing data on local equipment versus in the cloud is illustrated in figure 5. For this model, each dataset includes a structural MRI and one resting state fMRI, is processed using 4 pipeline configurations, takes 4.5 hours to process on 8 processors and 15 GB of RAM, and produces 4.5 GB of outputs. The local workstation is assumed to be a Dell Precision 7910 with dual Intel Xeon E5-2630 2.4 GHz 8-core processors (32 total virtual CPUs when hyper threading), 64GB of RAM, two 400GB solid state drives (SSD) for local storage, and a 1100 Watt power supply, that costs $8,642 (from dell.com on 1/31/2015). The cost of electricity is estimated based on 90% usage of the workstation power supply at the US average commercial rate for November 2014 of $0.1055 per kilowatt hour (http://www.eia.gov, accessed 1/31/2015). The costs associated with software and hardware maintenance was conservatively estimated to require 5% of a research technicians effort ($50,000 a year salary + 25% fringe) for a year. For the cloud computation model, c3.8xlarge EC2 instances are the most comparable to the specified workstation and include 32 virtual CPUs, 60 GB of RAM, and two 320 GB SSDs and cost $1.60 per hour for on-demand instances and average $0.26 per hour for spot instances in the US east zone (aws.amazon.com, accessed on 1/31/2015). Spot instances offer a method to pay reduced rates for spare cycles that are not currently being used at on-demand prices and their price fluctuates based on demand. These instances are reserved at a bid price, and if the cost of the instance goes above that price, the spot instance is terminated, making them less reliable for computation. This simulation allows up to 20 instances to be used for the computation. Additional costs include persistent storage ($0.10 per GB per month) to hold the inputs and outputs of the processing and the cost of transferring processed data out of the cloud, at a rate of $0.09 per GB. Importantly, the cloud computation model does not require software maintenance costs since the C-PAC development team maintains the AMIs. Figure 5 illustrates that the results of this simulation strongly favor cloud computing. The stepwise nature of the AWS line reflects that since up to 80 datasets (20 nodes, 4 datasets each) can be processed in parallel, the processing time and hence the price increases for every 80th dataset that is added. When using the more expensive and robust on-demand instances up to 5,000 datasets can be processed for less than the cost of owning and maintaining a workstation (Fig. 5A). The cost drops substantially when using spot instances, even for a more conservative model that uses twice the average spot instance price in its formulation. Using spot instances is cheaper than maintenance and electricity costs alone for processing up to 4,000 datasets. Another advantage of cloud computing is that additional nodes can be added with no additional overhead costs, resulting in much faster computation for the same bottom line (Fig. 5B), whereas adding twenty nodes to the local cluster would increase the capital costs 20 fold. The cost of processing 1000 subjects is almost 16 times cheaper for EC2 spot instances than for local computation (Fig. 5C). Thus illustrating that cloud computing is a very cost-effective alternative for neuroimaging data processing when these infrastructures are not available and can provide a simpler and more scalable solution even when they are. AIM 3. EXTEND C-PAC TO LEVERAGE CLOUD COMPUTING AND GRAPHICAL PROCESSING UNIT (GPU) TECHNOLOGY TO FURTHER OPTIMIZE COMPUTATIONAL EFFICIENCY. Implementing C-PAC as a Software-as-a-Service in the cloud: Although C-PAC has been ported into the Amazon EC2, it is far from the turnkey solution that the C-PAC team will build in releases 3 through 7 of the development timeline.

Software-as-a-service is a software distribution model in which applications are hosted by a service provider and accessed over the Internet. It can provide a user access to a considerable amount of computational resources, without any need to deal with software and hardware maintenance. Figure 8 illustrates the basic concept for the implementation of C-PAC in the cloud. Rather than installing the entire C-PAC ecosystem locally, the user will log into the Amazon cloud and start the C-PAC AMI on a medium size on-demand instance. The C-PAC AMI will be equipped with a web server that is running the C-PAC dashboard. By connecting to the C-PAC dashboard, the user will be able to configure pipelines and data bundles, initiate pipeline execution and monitor the pipelines progress. When the user starts a pipeline, the server will provision a computing cluster in the cloud based on the configuration provided by the user, and will start executing the pipeline. When processing has completed, the computing cluster will be terminated, but the master node will remain running until terminated by the user. Our goal is that the cost of process a single data bundle should not exceed $0.75 for spot instances and $2.50 for on-demand instances. Development of the C-PAC SaaS infrastructure will occur in three phases. The first phase, to be completed in release 2, infrastructure will be developed for building, maintaining, and testing C-PAC AMIs. The C-PAC dashboard will be constructed in releases 3 through 6. The dashboard will be developed in the Django Python web framework (https://www.djangoproject.com/). Although specifically developed for the cloud, the dashboard will also be usable locally to submit and monitor jobs on cluster systems. The third component, developed in releases 5 and 6, consists of quality assessment tools that will enable to user to view and rate pipeline outputs using a web browser. From this information, the user will be able to create subjects list for group-level analysis. Although the development of the dashboard and the quality assessment are listed as different features, they will be tightly integrated.

2. The Amazon Web Services Elastic Compute Cloud (EC2)

4000 characters including spaces total Introduction (1576) The National Database for Autism Research (NDAR, http://ndar.nih.gov) and other NIH/NIMH data repositories are amassing and sharing thousands of neuroimaging datasets. With the availability of this deluge of data and the development of the NDAR infrastructure for its organization and storage, the bottleneck for applying discovery science to psychiatric neuroimaging has shifted to the computational challenges associated with data processing and analysis. Maximizing the potential of these data requires automated pipelines that can leverage high-performance computing (HPC) architectures to achieve high throughput computation without compromising on the quality of the results. A disadvantage of this approach is that it requires access to HPC systems that are not always available, particularly at smaller research institutions, or in developing countries. Cloud computing resources such as Amazon Web Services (AWS) Elastic Compute Cloud (EC2) offers a pay as you go model that might be an economical alternative to the large capital costs and maintenance burden of dedicated HPC infrastructures. Realizing this need, the developers of the Laboratory of Neuro Imaging (LONI) Pipeline, the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) Computational Environment (CE) and the Configurable Pipeline for the Analysis of Connectomes (C-PAC) have implemented pipelines in EC2 that interface with NDAR. Each pipeline was used to perform a benchmark analysis of 2,000 structural images from the NDAR database to establish the feasibility of this approach.

2.1. EC2 Instances

2.2. Storage

2.3. Data Transfer

3. Processing Nueroimaging Data in the Cloud

Methods (1574) Each of three pipelines were installed into Amazon Machine Images (AMIs) and customized to perform structural preprocessing on NDAR data. The LONI Pipeline (http://pipeline.loni.usc.edu) was enhanced to permit direct access to NDAR collections for workflow-based data processing (Torgerson, in press). Workflows can be created from a combination of commonly available neuroimaging processing tools represented as Pipeline Modules. With respect the benchmark analysis, specifically developed Pipeline Modules captured the results from FreeSurfer and FSL FirstAll, updated the NDAR with the results and returned them back to the NDAR Amazon Cloud storage server. C-PAC (http://fcp-indi.github.io) is a configurable pipeline for performing comprehensive functional connectivity analyses that was extended to include the Advanced Normalization Tools (ANTs) cortical thickness methodology (Tustison, 2014) and to interface it with NDAR (https://github.com/FCP-INDI/ndar-dev). Results of this workflow include 3D volumes of cortical thickness and regional measures derived from the Desikan-Killiany-Tourville atlas (http://mindboggle.info/faq/labels.html). NITRC-CE (http://www.nitrc.org/projects/nitrc_es/) is an AMI that is pre-installed with popular neuroimaging tools. A series of scripts were developed for NITRC-CE to interact with NDAR, calculate a series of quality assessment measures on the data, perform structural imaging analysis using FreeSurfer and FSL FirstAll results, and to write the results back to NDAR (https://github.com/chaselgrove/ndar). Results(513) Speeds obtained for processing structural data in EC2 were consistent with those obtained for local multi-core processors. For example, using an EC2 instance with 4 processors and 15 GB of RAM (m3.xlarge), the C-PAC pipeline was able to complete the ANTS cortical thickness pipeline in 8.5 hours per subject, in comparison to 9 hours on a local workstation with 12 processors and 64 GB of RAM. EC2 processing cost $1.94 per image for on demand instances and an estimated $0.26 per image when using spot instances. Conclusions (349) Analyzing data using cloud computing is an affordable solution, with low hardware and software maintenance burdens; this can be beneficial for smaller laboratories and when data is already in the cloud. Further reductions in cost can be obtained using lower costs spot instances, which fluctuate in price and may get shut down if demand gets too high.

3.1. Cloud computing cost and performance

3.2. Other Considerations

Models for minimizing data transfer.

Optimizing allocation of resources.

Security and privacy. Kleinschmidt on on-going activity and Bertrand Thirion on statistical data processing. RCC would like to acknowledge support by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. The authors would like to thank the anonymous reviewers for their suggestions, which improved the manuscript.

4. Conclusion

References