# Amazon EC2 Pricing for MLMI Projects

*Brian B. Avants*

*8/10/2018*

> *"Make it work, make it right, make it fast" ... and then cost efficient.*
>
> - kent beck

## Cost of data processing and storage in Amazon Elastic Computing environment

The cost of EC2 will be a function, primarily, of the size of the data and the amount and type of computing, storage and data transfer a project requires. In EC2 terms, this breaks down into four factors:

- N: number of data elements, here assumed to be medical images which leads us to focus on compute instances that come with sufficient RAM.
- C: type of computing that is needed to process the images (CPU, GPU, RAM).
  - type: https://aws.amazon.com/ec2/instance-types/
  - spot-pricing cost: https://aws.amazon.com/ec2/spot/pricing/
- E: storage required on EBS
  - Elastic block storage (EBS) cost: $0.10/GB - $0.20/GB https://aws.amazon.com/ebs/pricing/
- T: transfer costs that will relate to how much of the compute output that one wants to save locally and/or transfer to S3. These can get complicated and range from free to several factors higher depending on degree of backup and number of transfers between regions ( which seems to be the main contributor to rising costs ).
  - Amazon EBS Snapshots to Amazon S3 currently $0.05 per GB-month of data stored https://aws.amazon.com/ebs/pricing/
  - S3 cost: https://aws.amazon.com/s3/pricing/

There are also relatively controllable "support costs" which act like taxes: https://aws.amazon.com/premiumsupport/pricing/.

A simple model of total cost will then be:

$$\text{cost} \propto N * (u * C + \gamma * (E + T)) * m \tag{1}$$

where $m$ is a multiplier for support costs, $\gamma$ scales the output for the expected amount of storage, $u$ indicates utilization (e.g. 10 for 10 hours) and $E, T$ are costs of storage on EBS and transfer to S3.

Why are we computing this when amazon provides a calculator? This lets us begin to build a predictive estimate for future work and modify/adjust over time. Furthermore, it's faster and more scriptable, versus endless poking around a website GUI, to have such a tool. Finally, we implement to aid understanding.

## Amazon EC2 pricing for example instances

Some instances come with storage. Note also that reserved pricing or spot pricing will reduce these costs by a factor of 50% or more.

Table 1: Cost for select EC2 instances as of August 2018

| | Name | vCPU | Memory..GiB. | On.Demand.Hourly.Cost |
|---|---|---|---|---|
| **6** | t2.large | 2 | 8 | $0.0928 |
| **7** | t2.xlarge | 4 | 16 | $0.1856 |
| **15** | m5.large | 2 | 8 | $0.0960 |
| **21** | m5d.large | 2 | 8 | $0.1130 |
| **28** | m3.large | 2 | 7.5 | $0.1330 |
| **29** | m3.xlarge | 4 | 15 | $0.2660 |
| **49** | c3.xlarge | 4 | 7.5 | $0.2100 |
| **64** | r3.large | 2 | 15.2 | $0.1660 |
| **75** | r5.large | 2 | 16 | $0.1260 |

useful resource for this material

EC (elastic computing) provides a linux environment (e.g. ubuntu, debian, etc) with a specified number of cores and memory available.

## Amazon EC2 pricing for an example medical imaging project

Let us assume we want to compute BrainAge with ANTs processing based on public data. This would be similar to a cloud-based project, NAPR, in which models were trained using healthy control data from the ABIDE, CoRR, DLBS and NKI Rockland neuroimaging datasets (total N = 2367, age range 6-89 years).

Additional datasets might include IXI (531), OASIS (313), SALD (494), SLIM (580), PNC ($\approx$ 1200), ABCD (up to 11,500, currently 4,500) the Human Connectome Project (1,200-1,400) as well as the UK Biobank (10,000-20,000). In total, these would yield over 25,000 normative T1-weighted neuroimages covering the human lifespan.

To process this smaller dataset (2367), we choose r5 instances, though others would do. Description from amazon: "The memory-optimized R5 instances use custom Intel Xeon Platinum 8000 Series (Skylake-SP) processors running at up to 3.1 GHz, powered by sustained all-core Turbo Boost. They are perfect for distributed in-memory caches, in-memory analytics, and big data analytics, and are available in six sizes."

We assign the following values for equation (1):

- N = 2400 images;
- u = 10 hours per instance, based on time we expect the processing will take per brain;
- C = r5.large: $0.126 is the on-demand instance cost vs $0.05 hourly for reserved cost;
- $\gamma$ = 0.209 which leads to on-disk storage of 501.6GB;
- E = $0.1 nominal cost per GB (per month) on EBS
- S = $0.05 nominal cost per GB (per month) transfer to S3 ( backup )
- m = 1.1 a multiplier for service cost

**This yields an estimate of $3409.164 for the BrainAge project.** We validate this calculation against the official amazon estimate for this project:

Table 2: Amazon estimate for BrainAge project (continued below)

| Your.Estimate | X |
|---|---|
| Service Type<br>Amazon EC2 Service (US East (N. Virginia)) | Components<br><br><br>Compute: |

| Your.Estimate | X |
|---|---|
| | EBS Volumes:<br>EBS IOPS:<br>EBS Snapshots: |
| AWS Data Transfer In | US East (N. Virginia) Region: |
| AWS Data Transfer Out | US East (N. Virginia) Region: |
| AWS Support (Business) | Support for all AWS services: |

| X.1 | X.2 | X.3 |
|---|---|---|
| Region | Component Price | Service Price |
| | | $3099.00 |
| US East (N. Virginia) | $3024.00 | |
| US East (N. Virginia) | $50.00 | |
| US East (N. Virginia) | $0.00 | |
| US East (N. Virginia) | $25.00 | |
| | | $0.00 |
| Global | $0.00 | |
| | | $44.91 |
| Global | $44.91 | |
| | | $313.96 |
| | $313.96 | |
| Free Tier Discount: | | -$4.31 |
| Total Monthly Payment: | | $3453.56 |

It remains to be seen if these are accurate estimates of actual cost. Costs can be reduced by $> 50\%$ by using:

- reserved instances
- spot instances

Processing this smaller dataset will improve cost estimates for the full 30,000 subject processing (above data plus ADNI, PPMI and other aging and/or neurodegenerative disease cohorts).

Only implementation will reveal true costs. Extra costs are likely to occur due to the overhead of setting up the environment (possibly compilation of open source tools), to setting up/testing/managing cluster computing and/or any potential failures due to insufficient memory and/or bugs/inconsistencies in the EC2 environment.

## Other notes and comments below (ignore unless interested)

### EC2 types and on demand pricing

- types: https://aws.amazon.com/ec2/instance-types/
- m5 ("next generation EC"): https://aws.amazon.com/ec2/instance-types/m5/

### Spot pricing

Amazon spot pricing saves costs on using EC2 by scheduling your work to run when computing becomes available. This is similar to havign a "low priority" job on a standard queue scheduling system. Cost savings are roughly 50%, possibly more.

Current amazon spot pricing link

### Cluster computing

http://star.mit.edu/cluster/

https://alces-flight.com/

### Storing Data on AWS

So, where does data go? Project data, as well as files necessary to run bioinformatics software (reference genomes, bwa indicies, blast dbs, SNP dbs, vcfs, etc), should go on S3. (If you want archival storage for data you're rarely going to touch, there's another Amazon service called Glacier.) Here's the story of S3 vs EBS in bullet points:

- EBS Volumes are expensive
- S3 is cheap
- But: you can't compute on S3 (in general)
- And: you can compute on EBS
- So: pull S3 stuff onto EBS temporarily, compute, then delete the EBS volume when finished

In other words, S3 is a long-term storage solution; an EBS volume is a short-term storage solution while you're running your jobs.

Tip: To save money, don't keep big volumes kicking around for a long time after your jobs have finished.

Tip: some tools can work directly with bam files on S3 sans download (see Biostars: Tool for random access to indexed BAM files in S3?).

### Post-hoc analysis of S3/EC costs

https://aws.amazon.com/blogs/big-data/analyzing-aws-cost-and-usage-reports-with-looker-and-amazon-athena/

## amazon calculator

https://calculator.s3.amazonaws.com/index.html