Case Study: Deploying PetaGene's lossless compression for genomic data at a premier Children's Hospital in the United States

PetaGene
secure transparent compression

One of the top-three ranked Children's Hospitals in the United States (CH) has purchased and deployed PetaGene's genomic compression, two months after introductions.

Introduction

Understanding the genomic information of a patient is key in diagnosing a plethora of genetic and rare diseases. As a result, genome sequencing approaches such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) are growing in use in clinical and research laboratories such as at CH.

A critical challenge emerges as genome sequencing scales up - how to manage the increasing cost of storing these data. At CH, the volume of sequencing data is growing rapidly along with the associated storage costs. CH were looking for solutions to reduce ballooning costs of storing these data that needed to be accessed for critical research and analysis.

To solve this challenge CH adopted PetaGene's lossless compression software that consists of:

PetaSuite - the lossless compression binary that efficiently compresses FASTQ(.gz) and BAM files, making these files 60-90% smaller.

PetaLink - An LD_PRELOAD Linux library that runs in user-mode and allows tools and pipelines to transparently access NGS data in their original file formats.

Deploying PetaSuite Software

CH were keen to purchase PetaGene's software from the initial call, especially given the clear ROI, but wanted to evaluate the software before making a final purchasing decision. PetaGene supplied software that allowed CH to do this evaluation quickly and in two phases:

Eval Phase 1: PetaSuite Preview - test thousands of files at scale

The evaluation leveraged PetaSuite's Preview Tool, which allowed CH to test an unlimited number of files in a production environment. CH used it to do dummy compressions on 83,205 files with a total input size of 116 TB. These files were passed through the PetaSuite compression tool and validation step allowing CH to determine how much compression savings would be achieved, as well as speed, resources used, and software robustness.

File	# of files	Input size (TB)	Output size (TB)	Savings (TB)
FASTQ.gz	15,116	43.8	16.4	27.4
BAM	68,042	72.3	26.5	45.8
Total	83,158	116.1	42.9	73.2

The PetaSuite Preview version of the software provides real-world data that organizations can use to benchmark the compression performance of PetaSuite against thousands of files.

The results of the evaluation with PetaSuite Preview convinced CH to move forward to the next phase of evaluating the software.

Eval Phase 2: Testing PetaLink decompression

CH used the full PetaSuite tool on an eval licence with limited quota to produce 16 compressed FASTQ.gz files to test with the PetaLink Library in a selection of their standard pipelines - for example, mapping with Bowtie2 and alignment in the SunGrid Engine cluster. PetaGene compressed files appear to the tools in the original, uncompressed form with decompression managed by the PetaLink library in a transparent manner.

How it's going

These tests showed that in addition to storage savings, PetaGene compressed files can be accessed directly by all bioinformatics tools in any computing environment without requiring an intermediate state in which the files are first decompressed.

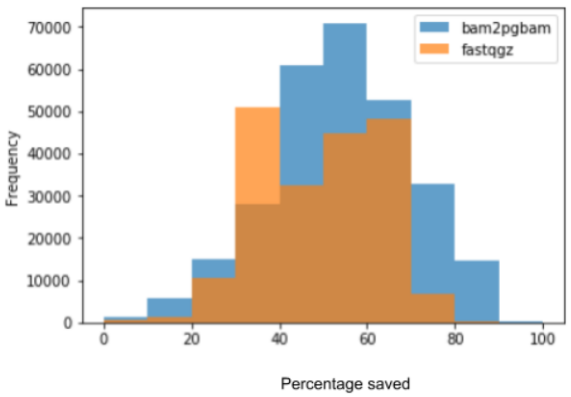

Figure 1. Distribution of files by compression savings (Percentage saved).

File	# of files	Input size (TB)	Output size (TB)	Savings (TB)
FASTQ.gz	663,669	854	360	494 (58%)
BAM	849,465	2,165	879	1,286 (59%)
Total	1,513,134	3,019	1,239	1,780 (59%)

Customer testimonials:

Infrastructure Manager, CH:

"Within two months of deploying PetaGene's compression solution, CH had made a return on investment, recovering the full cost of the PetaGene licence in storage savings. With PetaGene, we now have better control over the growth of our NGS data, allowing us to reduce storage costs while freeing up financial resources for more compute and analysis to further the research and clinical goals of CH. Hospital senior management consider the PetaGene purchase a big success."

Principal Systems Engineer, CH:

"Deploying PetaGene at CH was a straightforward process. Our users do not have to modify any of their tools or workflows since PetaGene's decompression library transparently serves original uncompressed data to the tools/pipelines. Since users are essentially able to work with the original data except that the data is more than 50% smaller, there are additional benefits such as faster data transfer speeds and analysis times."