Reprocessing all the XMM-Newton scientific data: a challenge for the Pipeline Processing System

José-Vicente Perea-Calderón[1], Pedro Rodriguez-Pascual[2], and Carlos Gabriel[2]

[1]*RHEA for ESA/ESAC, European Space Astronomy Center (ESAC-ESA), Madrid, Spain;* jose.perea@sciops.esa.int

[2]*XMM-Newton SOC, European Space Astronomy Center (ESAC-ESA)*

Abstract.

2019 will mark the 20-year anniversary of the XMM-Newton Mission[1]. So far, the mission has successfully completed a total of around 14.000 pointing observations, and it is expected to continue for many more years, producing a huge number of high-quality science data products.

Data processing of those observations is carried out by the XMM-Newton Pipeline Processing System (PPS)[2] and the products are delivered to the XMM-Newton Science Archive (XSA)[3]. During the two decades many changes have been implemented in the data processing software, partly following improvements to the calibration of the science instruments. Several re-processing campaigns have been undertaken along the mission in order to have an up-to-date and uniformly processed set of high-level science data products in the archive.

This paper is a review of the analysis that has been carried out to achieve a new whole mission re-processing and the way to do it in the future.

1. Introduction

A whole mission re-processing campaign requires a power computer infrastructure to be done. But even so the overall processing time could be considered too long depending on the number of observations, the complexity of the software algorithms and the calibration, etc. Substantial reduction of the processing time of the mission would allow to change the frequency of the renewal of the archive contents.

Unlike the daily mission operations where a a limited number of observations have to be processed by PPS, a whole mission re-processing is a real challenge. An individual XMM-Newton Pipeline job (of one observation) can take up to six hours of computer processing time, some of them even longer. To achieve the processing of thousands of observations in a reasonable period of time requires a special preparation including a deep analysis of the computing resources. An extreme optimization of the resources sharing becomes essential in this case.

---

[1]XMM-Newton SOC home page http://xmm.esac.esa.int.

[2]XMM-Newton Pipeline https://www.cosmos.esa.int/web/xmm-newton/pipeline.

[3]XMM-Newton Science Archive http://nxsa.esac.esa.int.

Besides the optimization of the computing infrastructure usage, a set of software tools had to be developed in order to cope with the management and monitoring of this enormous number of individual Pipeline jobs.

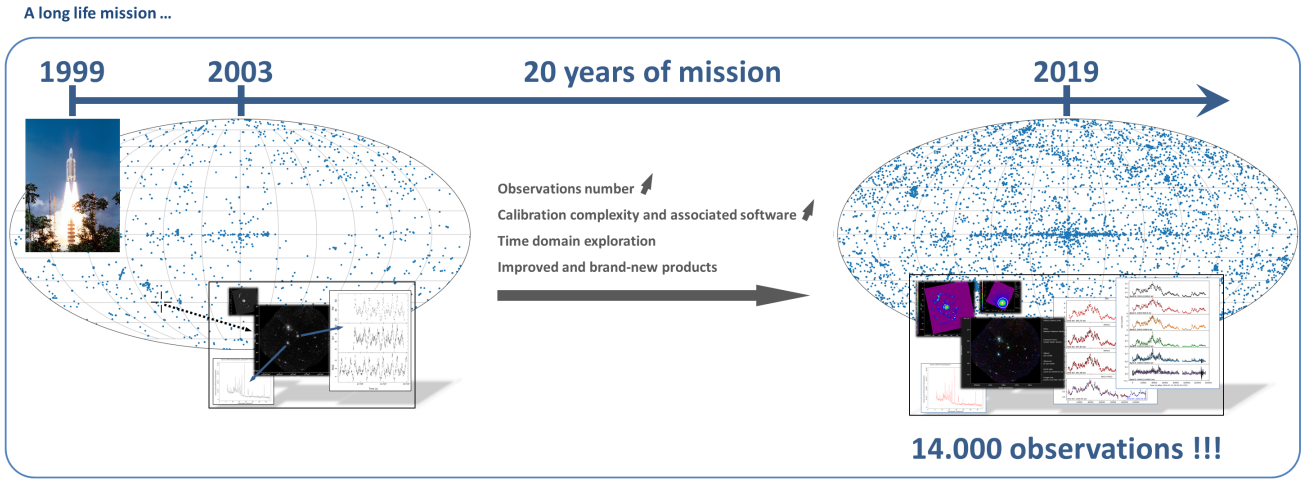

Figure 1. Increase of complexity over the years

2. All XMM-Newton mission re-processing

In addition to the increase of the number of observations over the years, the complexity of the calibration and the associated software has been also increased significantly. As a result the Pipeline produces improbed and brand-new data products. It is also expected to provide new results from the time domain exploration of the data: variability, transients detection and others.

The considerable improvements in the quantity and quality of the results have resulted in subsequent increased workload in the computer infrastructure and, as a consequence, an important increase of the processing time.

The real challenge is not processing all the observations of the mission itself but finding out how fast we can do the job. Processing all the mission in a short period of time would allow to populate the archive on a continuing progressive and dynamic basis. So any significant change in the calibration or important software upgrade would produce a new "all-mission pipeline products pack" ready to be ingested in the archive.

In order to put this idea in place we have set ourselves the goal of processing the whole mission within a week.

3. XMM-Newton Pipeline single-job

The first approach to speed up the process is splitting every Pipeline job into 4 threads which provides a a process time reduction by a factor 2. But still, most of the Pipeline jobs may take up to 6 hours to complete in any of the computer nodes of the ESAC/ESA Grid infrastructure. In addition, many of those jobs might need up to 8 GB of memory to be processed (figure ...).

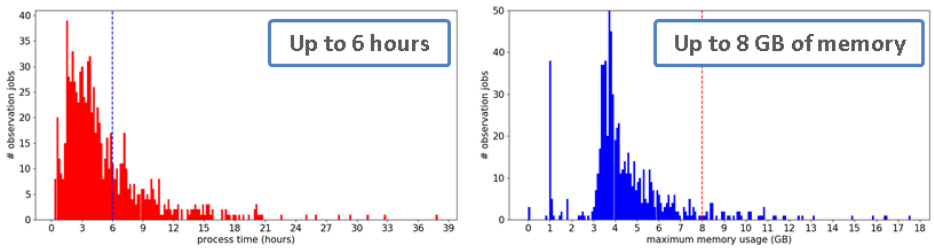

Figure 2. Pipeline single-job analysis

4. Processing computer infrastructure

The setup of the XMM-Newton Pipeline processing system is settled into the ESAC Grid computer infrastructure. Conservatively so far, the Pipeline has been initially set up to request a limited amount of memory and CPU resources. Following a testing phase of high workload experiments on the computer infrastructure we found out two significant features of the Pipeline jobs:

- The maximum demanded memory by the jobs only happens in very short time peaks, so there is a lot of free memory most of the time
- The CPU's load hardly reach the 50 % of the total infrastructure CPU power

These findings imply that we can further intensify the shared resources of memory and CPU in the computer infrastructure. The system is therefore forced to reach almost the hundred percent of CPU load capacity and the jobs are launched with no memory restriction. In this way we are able to have a 140 % more simultaneous jobs than in the initial conservative setup. And that is the key point to reduce the overall processing time.

Figure 3. Speep up the overall processing time

...
...
...
...
...
...

5. Conclusions

- We can reprocess all the mission at any time
- Requirements for every single-job matter
- Deep analysis of the computer capabilities is essential to accomplish the process of this huge number of jobs within a reasonable time
- New monitoring and management tools become a necessity