URSI GASS 2023, Sapporo, Japan, 19 – 26 August 2023

Processing ASKAP's Big Data - Challenges and Lessons Learnt

Wasim Raja*[1], Maxim A. Voronkov[1], Matthew T. Whiting[1], Mark H. Wieringa[1], Daniel Mitchell[1], Stephen M. Ord[1], Minh Vuong[1], Eric Bastholm[2], Matt Austin [2], Paulus Lahur[3], Perica Manojlovic[1], Ozgur Cekmer [2] Tim Galvin [2]
(1) CSIRO Space & Astronomy, PO Box 76, Epping, NSW 1710, Australia
(2) CSIRO Space & Astronomy, PO Box 1130, Bentley, WA 6102, Australia
(3) CSIRO IM&T, Clayton South VIC 3169, Australia

The Australian Square Kilometre Array Pathfinder (ASKAP) is a wide field-of-view ($\sim 30^{\circ 2}$) radio interferometer comprising of 36 fully steerable 12 m dishes operating at low radio frequencies ($700 - 1800\ MHz$) with instantaneous *correlated* bandwidths of 288 MHz. ASKAP provides researchers not just the ability to survey the radio sky in new and efficient ways, but also to witness and strategise mitigation of inevitable challenges in data processing necessitated by ASKAP's innovative, yet very complex technologies. In this paper we highlight some of the challenges that influenced the evolution of ASKAP data processing strategies from its conception to full survey science operations that began late in 2022.

The huge amount of raw (post-correlated) data ($> 100TB/day$) coming out from ASKAP makes data processing using traditional software packages practically unfeasible. *ASKAPsoft* - a high performance computing (HPC) software package - was therefore developed in-house as part of the telescope design, and is used to process ASKAP data on HPC clusters at the *Pawsey Supercomputing Centre* in Western Australia. In addition to compute, storage is a major operational overhead. When processed to reveal full ASKAP capabilities, the intermediate data sizes can grow upwards of a few hundred TBs per day. The I/O of the disks too is a major factor impacting processing efficiency, especially in a shared environment. The original design aimed to mitigate these by reducing data on-the-fly and producing science-ready images in near real-time, avoiding multiple iteration over the data where possible. The feasibility of such single-pass strategy relies heavily on overall statibility of the system, and/or its calibratibility using a-priori model of the sky - assumptions that have practical limitations.

In any case, establishing feasibility of this radical strategy has its own challenges. While the processing algorithms we use are mature and robust, different science goals require different optimisation of the processing parameters to achieve desired image quality. The number of such parameters, as well as the extent of each parameter space are rather large, thereby increasing significantly the turn-around times for establishing optimised science-dependent automated workflows. The compute architecture too opens up another dimension to optimise against. Finally, the complex network of components in ASKAP can lead to complex processing failure modes that are hard to anticipate and factor in without extensive experience of at-scale processing of real data. Understanding of these issues during the commissioning phase thus required means to flexibly interact with the data, resulting in addition of significant technical debts to the goal originally envisaged. The processing continues to be done in offline mode using *ASKAPpipeline* - a highly configurable, high performance processing pipeline that orchestrates execution of core *ASKAPsoft* applications to meet a given science goal.

Over the years, the ASKAP *Software & Computing* group has worked closely with commissioning scientists & engineers in assessing and understanding the various practical challenges posed by ASKAP & its data, and in devising effective strategies to generate images as close in quality as permissible by the instrument's limits. While the deviation from on-the-fly to offline processing resulted in significant improvements in achieveable image quality, the need to have multiple passes through data drastically reduces processing throughput, making operations untenable on the supercomputer *Galaxy* dedicated for ASKAP data processing. For sustained operations, ASKAP requires an HPC ecosystem with configurations suitable for its most intensive workflows. Experiments on *Pawsey Supercomputing Centre's* newest cluster Setonix, with its higher core count and shared memory per node, is showing great promise. Nevertheless, migrating operations to a new system presents unavoidable challenges ranging from software building & deployment, to execution optimisation. The *ASKAP Software & Computing* group has designed robust solutions to address these issues that includes packaging our software into portable containers, as well as providing means of setting cluster-dependent processing configurations. While the current focus of the team has been to facilitate completely un-supervised processing, efforts are also being put in making our software use hardware acceleration when these are available in a future upgrade of the cluster.