Co-design for the Science Data Processor

now, AA2 and beyond

Shan Mignot for the SCOOP team

Swiss SKA days 2023

Introduction

Science Data Processor (SDP)

- Purpose
  - receive visibilities & pulsar data from CSP: ~0.5 TB/s
  - close control & feedback loops for observation: streaming processing
  - elaborate image cubes, pulsar candidates & timing: batch processing
  - distribute data products to SRCNet: ~10 GB/s
- Requirements
  - streaming processing: latency requirements for control loops (~15 s)
  - batch processing: process 24h of data in 24h, ~50x data reduction
- Likely the weak point in the data chain with a possible impact on observing (not accounted for in the telescope availability budget)

Challenges

- Performance requirement estimates
  - 12.5 PFlops average / 125 PFlops peak
  - 40 PB data buffer (~24 h storage)
- Energy budgets
  - average: 1.3 MW Mid / 1.6 MW Low
  - peak: 2.0 MW Mid / 2.23 MW Low
  - Green500: Frontier, Lumy, Adastra achieve ~100 PFlops peak @ 2MW at maximum efficiency

SKA1-Mid SPC/ SOC Power Budget in Cape Town				
Products	AA4 Long TermAverage (>30min) [kW]	AA4 Peak Instantaneous (<5sec) [kW]	AA* Long TermAverage (>30min) [kW]	AA* Peak Instantaneous (<5sec) [kW]
PDT4 - MID Digitisation	230.8	323.4	230.8	323.4
CSP.CBF	230.8	323.4	230.8	323.4
PDT6 - Network & Computing	1641.7	2481.9	589.3	872.6
SDP Hardware MID	1300.0	2000.0	325.0	500.0
PSS Hardware MID	296.0	414.0	222.0	310.5
PST Hardware MID	16.4	26.8	12.3	20.1
OMC Hardware MID	12.9	18.1	12.9	18.1
NSDN MID	5.6	7.8	6.6	9.2
CPF-SPC link MID	8.2	11.5	7.9	11.1
NMGR	2.6	3.6	2.6	3.6
Building losses and cooling	374.5	561.1	164.0	239.2
Commissioning Margin	224.7	336.6	98.4	143.5
Site Total	2471.7	3702.9	1082.6	1578.8

SKA1-Low SPC/ SOC Power Budget in Perth				
Products	AA4 Long TermAverage (>30min) [kW]	Peak Instantaneous (<5sec) [kW]	AA* Long TermAverage (>30min) [kW]	AA* Peak Instantaneous (<5sec) [kW]
PDT6 - Network & Computing	1629.2	2270.9	429.2	598.4
OMC Hardware LOW	12.9	18.1	12.9	18.1
SDP Hardware LOW	1600.0	2230.0	400.0	557.5
NSDN LOW	6.5	9.1	6.5	9.1
CSP-SDP LOW	7.2	10.1	7.2	10.1
NMGR	2.6	3.6	2.6	3.6
Building losses and cooling	325.8	454.2	85.8	119.7
Commissioning Margin	195.5	272.5	51.5	71.8
Site Total	2150.6	2997.6	566.6	789.9

from SKA-TEL-SKO-0000035-06

Staged deployment & heterogeneity

- Array assemblies

	Purpose	Mid		Low	
		date	dishes	date	stations
AA0.5	risk mitigation	01/2025	4	11/2024	6
AA1	risk mitigation	12/2025	8	11/2025	18
AA2	fully operational	01/2027	64	10/2026	64
AA*	fully operational	10/2027	144 (64 from MeerKAT)	01/2028	307
	End of construction	07/2028		07/2028	
AA4	fully operational	?	197	?	512

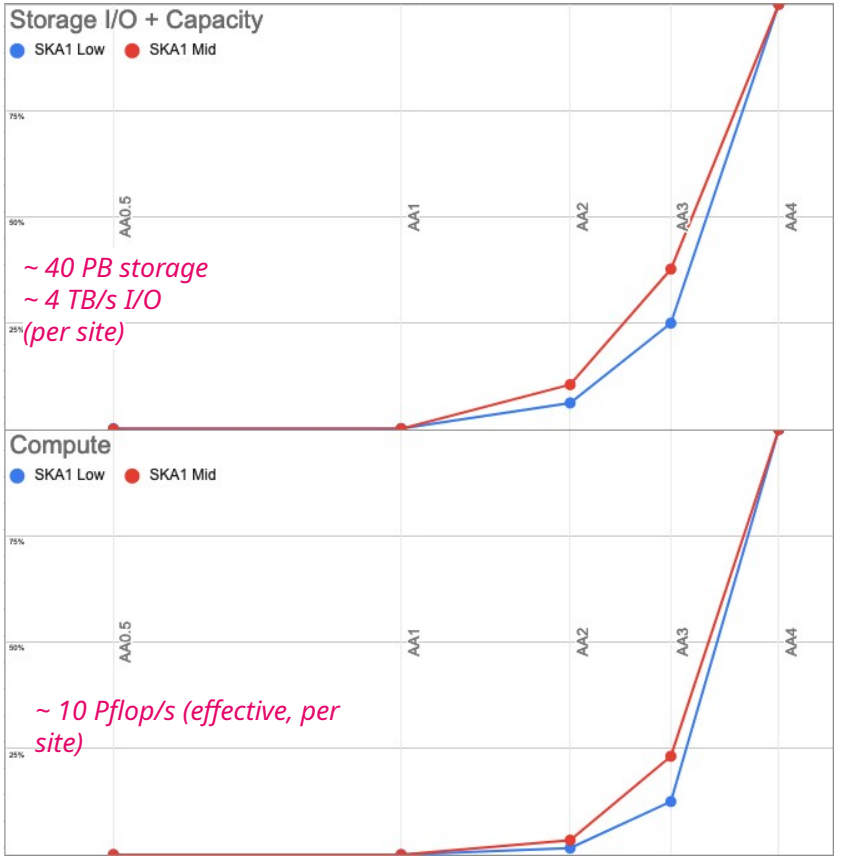
- Extension & maintenance over 50-year observatory lifetime

AA2 milestone

- Comparable to largest aperture arrays (LOFAR, MeerKAT)
- Carry out scientific observations
- SDP procurement starting this week, contract placed by Q4 2024
- Validate the SW/HW design and its ability to scale to AA4

Estimated SDP Scaling: AA1→AA4

(~50x in 17 months! Qualitative only, underestimates the AA2 situation, graph by P. Wortmann)



Co-design

Feasible? Achievable?

- Critical design review assumptions
  - Dennard's scaling & Moore's law: predictable increase of performance over time so procure SDP as late as practicable
  - estimated peak performance ranks 7th in Top500 today
  - 10% efficiency >> 1-3% of major Top500 systems (HPCG 2022)
- SKA computing hardware risk mitigation plan
  - financial risk vs. evolution of performance: capital cost of hardware
  - estimate uncertainties (software under development)
  - procurement strategy: collaborate with suppliers
  - power consumption: operational cost (single largest cost item)

Tooling: the benchmark suite

- Purpose
  - measure execution on different machines: environments, policies, architectures
  - automate and harmonise data collection and analysis to track performance
- Manage dependencies
  - SPACK recipes for casacore, DP3 and WSClean (recompiled)
  - Anaconda for utilities (not recompiled)
- ReFrame as a deployment and execution engine: installs, compiles and runs benchmarks from a central repository

Tooling: execution metrics

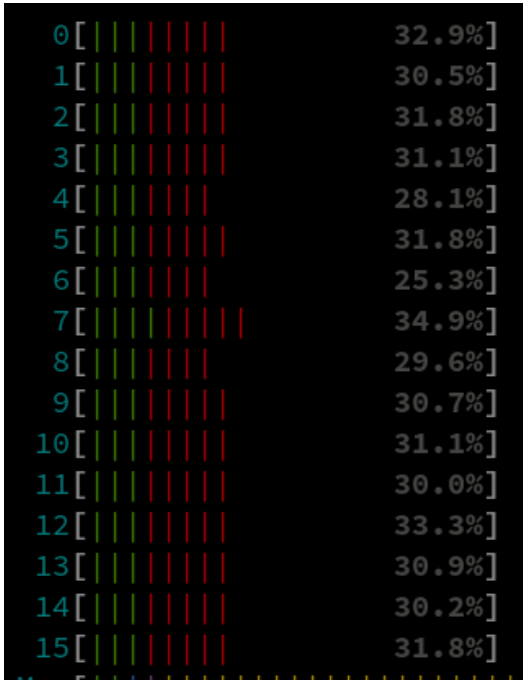
- Measure execution in a non-intrusive way
  - obtain a representative picture of performance
  - as a prerequisite to in-depth analysis (profiling)
- Data collection
  - RUM (Ressource Usage Monitoring): collect cpu/disk/memory /network usage statistics based on system files
  - PMT (Power Measurement Toolkit): measure energy usage
    - interfaces with RAPL for CPUs
    - interfaces with NVML for NVIDIA and roc-smi for AMD GPUs
- Jupyter notebooks: process and visualise measurements

State-of-the-art software evaluation

- Rapthor
  - self-calibration pipeline in production for LOFAR: mix of calibration and imaging
  - based on DP3 (calibration, prediction) and WSClean (imaging): C++ code
  - Python pipeline implementation of first 3 iterations by team SCHAAP
- Assess if AA2 objectives can be met with it
  - in SKAO context: Low telescope properties, dedicated SDP
  - identify bottlenecks and feasibility of improvement: for SKAO to use it or as guidelines to software design
- Evaluate ability to scale beyond AA2

DP3 spin locking

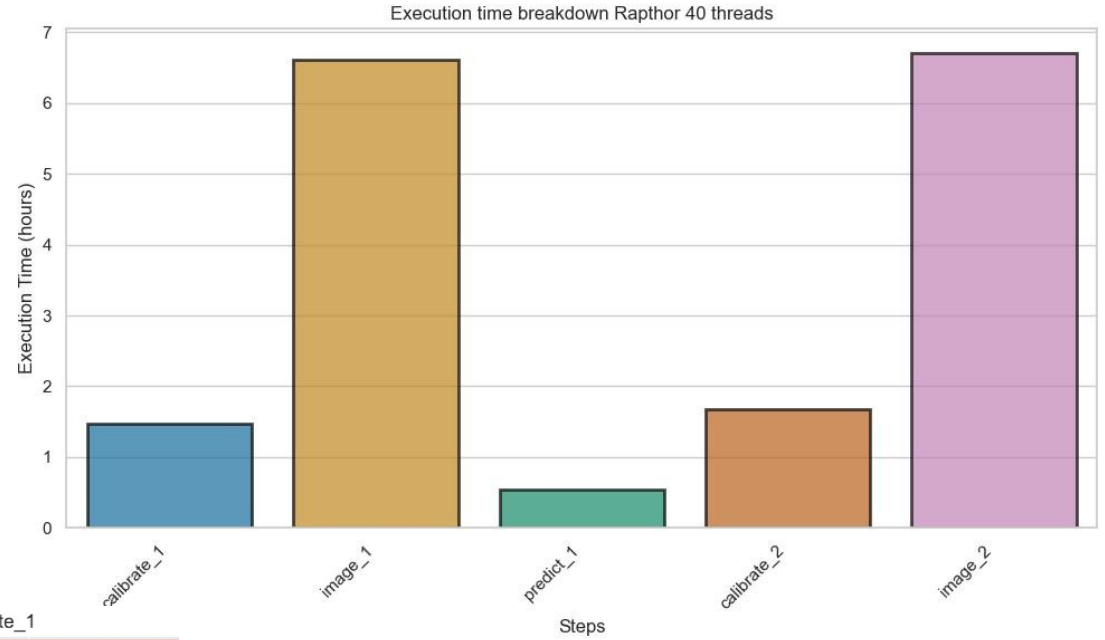
- In gaincal, ApplyBeam leads to spin locking
  - mechanism for synchronising threads
  - 50% of core usage: resources used by kernel to check mutex
- Results from casacore::Direction not being thread-safe
- Significant performance improvement likely
  - use different synchronisation model
  - replace casacore::Direction



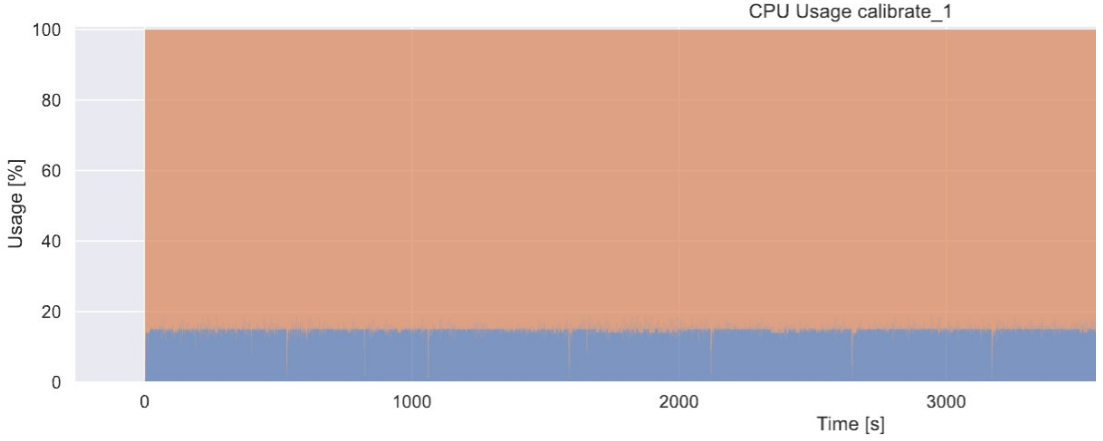
partial htop snapshot of core usage (green: user, red: kernel)

Execution metrics

- Low CPU usage
  - < 20% for DP3
  - < 60% for WSClean

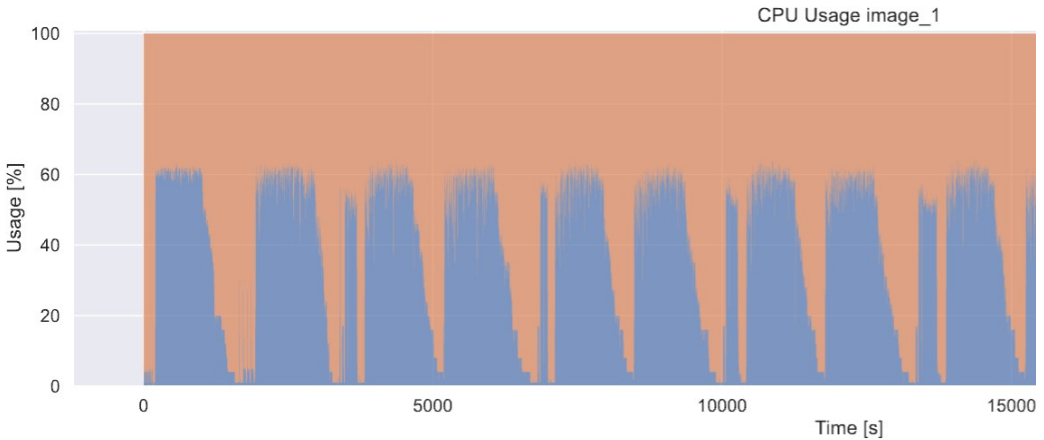


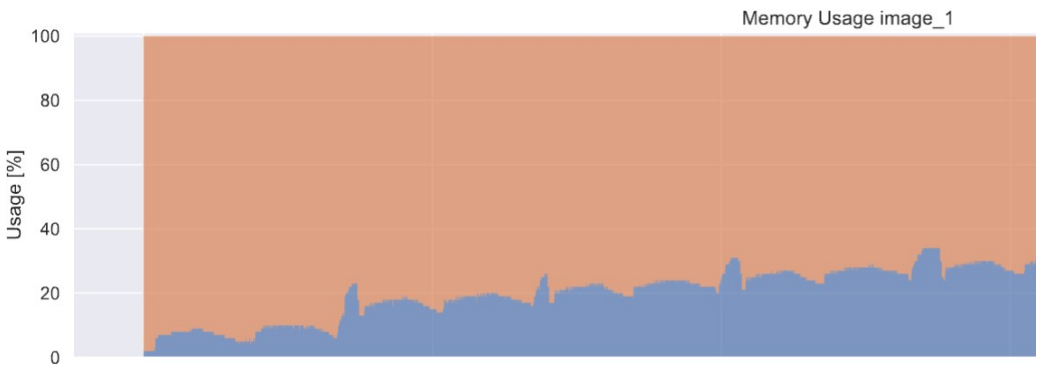
execution time for the different steps of the pipeline (iterations 1 & 2)

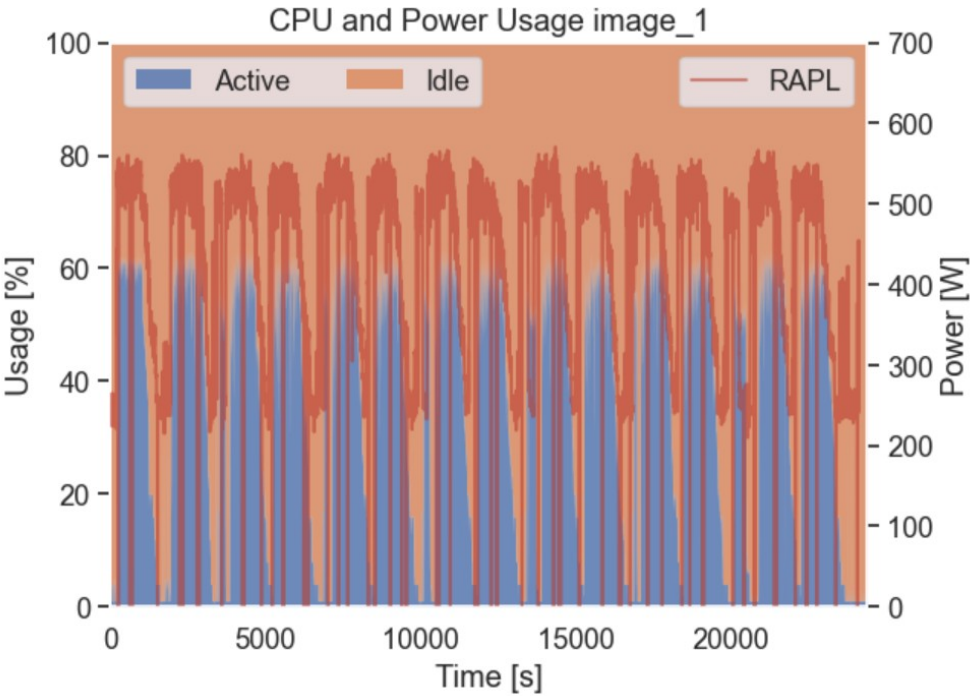


core usage for calibrate_1 (DP3)

Execution metrics (II)



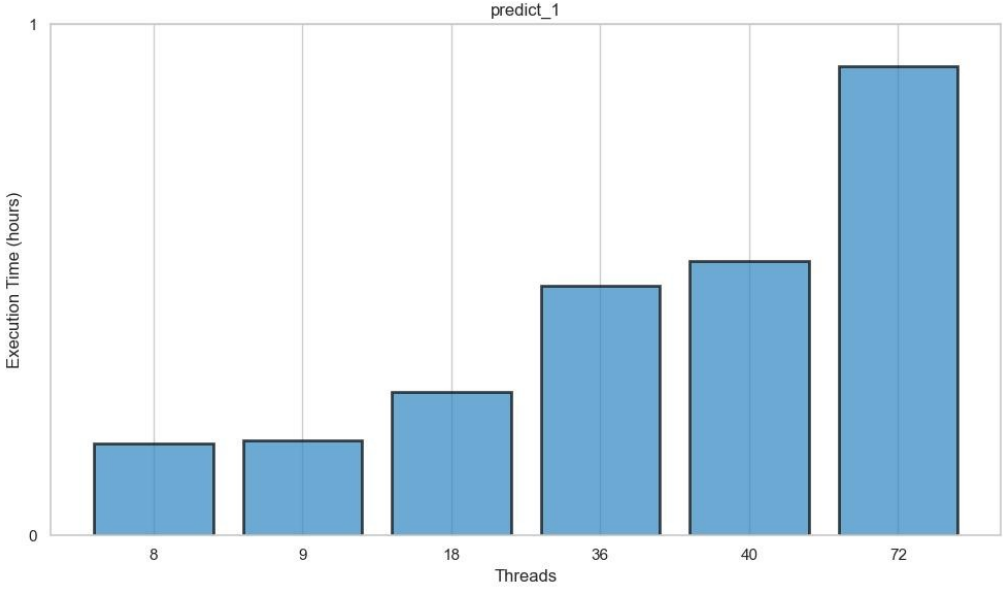




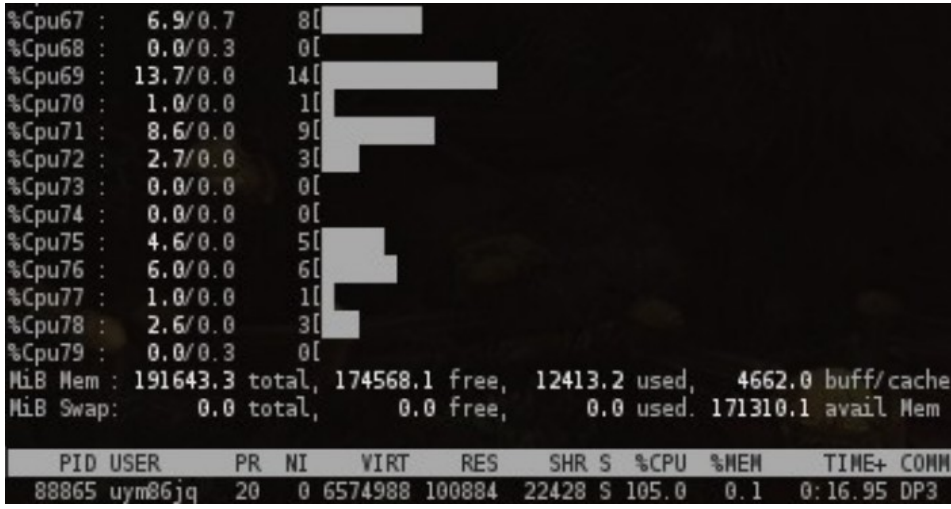
resource usage recordings for the image_1 step (WSClean)

DP3 multithreading

- Thread experiments
  - scaling: increases execution time!
  - disabling multi-threading not complete



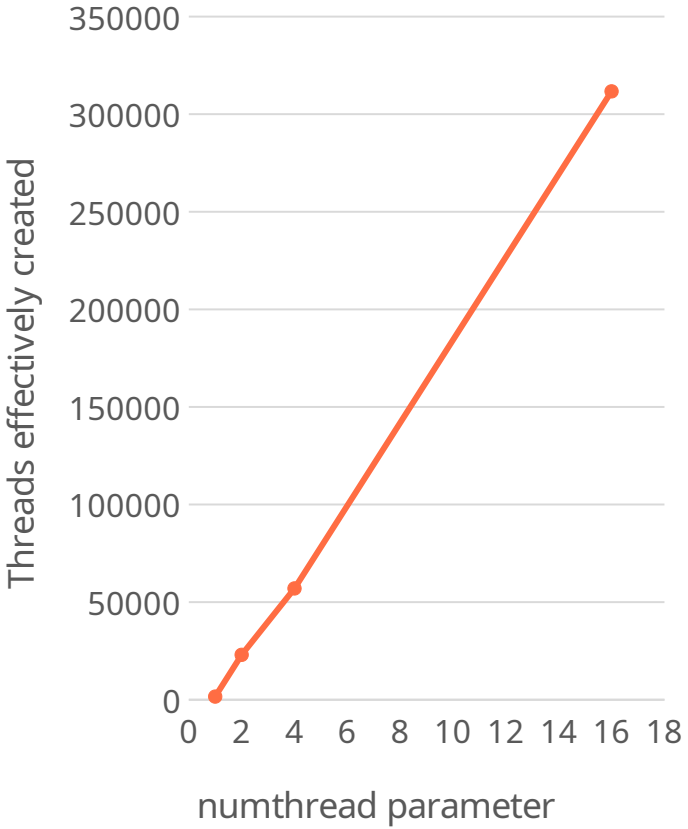
execution time as a function of user-requested number of threads



htop snapshots of core activity with multi-threading disabled: still carries out distribution on multiple cores and spin locks for synchronisation

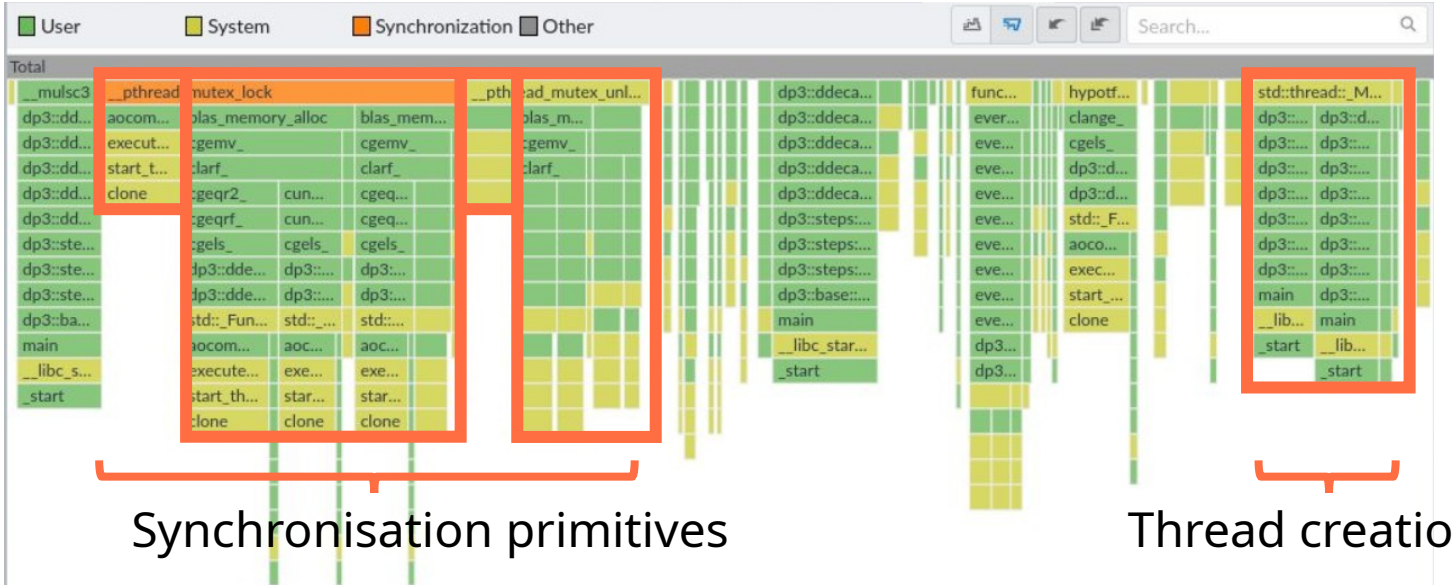
DP3 multithreading (II)

- Work in collaboration with Eviden (Atos)
  - very high number of threads: configured for 16 but >300k created
    - leads to spin time and thread management overhead
    - impedes collection of execution metrics (perf, vtune)
  - threads live less than 500 ms: code is mostly sequential
  - leads to competition for resources: strains system



DP3 multithreading (III)

- Significant improvements shown to be possible by revising thread management (work by Eviden)
  - modified ParallelFor.h to use OpenMP
  - ThreadPool still creating many threads (1600 threads created)



Prospects

Tooling perspectives

- Different flavours of the suite being envisaged
  - pipeline developers for performance regression testing
  - suppliers for tender and product acceptance
- Data collection
  - trace collection to track thread and MPI behaviours
  - define normalised benchmark outputs
    - compare software versions
    - compare machine performance
    - across teams

Pipelines and co-design perspectives

- Self-calibration pipeline
  - Low: work with SCHAAP to correct identified deficiencies
  - Low: parallelised time steps across multiple nodes using DASK
  - Mid: being developed by team HiPPO based on DP3 & WSClean
  - Mid: different size/resolution for images likely to lead to different performance behaviour
- Co-design
  - address scaling issue
  - focus on software/hardware match to prepare SDP procurement

Evolution of co-design organisation

- Co-design contract
  - ~ 2 FTEs for 12-month test period funded by SKAO on SDP hardware money
    - prove the value of investment: can co-design pay for itself via savings during procurement?
    - start date to be defined (possibly end of 2023)
  - France: Direct Data Network (storage provider)
  - Switzerland: EPFL (led by Stefano Corda)
  - separate team: SKAO not willing to mix cash and in-kind contributions
  - Product Owner: Miles Deegan (SKAO)

Evolution of co-design organisation (II)

- In-kind contribution
  - France willing to continue providing an in-kind contribution
  - seeking recognition by SKAO (in addition to cash contribution)
  - on-going discussion on scope: will contribute to characterisation of execution with a view to energy and sustainability aspects (eg. hardware trade-offs, operation, staged deployment and maintenance)
  - extension possible to Science Processing Centre for a system view of processing resources (infrastructure, CBF, PSS, PST, SDP)
  - team could join DP ART to strengthen the transverse nature of co-design
- In-kind contribution by Switzerland? By FHNW?

Credits and acknowledgements

- SCOOP team members
  - EPFL: Stefano Corda
  - FHNW: Manuel Stutz
  - OCA: Shan Mignot, Mathieu Carrère/Chiheb Sakka
  - Inria: Anass Serhani
  - Observatoire de Paris: Aristide Doussot
  - CGI: Neil Quinn
  - Astron: Chris Broekema
- Collaboration with Eviden
  - Clément Devatine (intern)
  - Erwan Raffin
  - David Guibert
- Computing resources
  - CSCS: Piz Daint
  - IDRIS: Jean-Zay
  - Inria: Grid5000
  - EPFL: Jed cluster