RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

PROGRAMME DE RECHERCHE
NUMÉRIQUE POUR L'EXASCALE

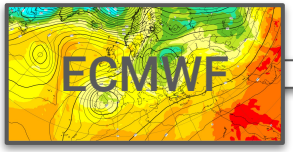
ExaDoST Illustrator - A Preliminary I/O Study of Radio Imaging Components

Francieli Boito (Université de Bordeaux)
François Tessier (Inria Rennes)

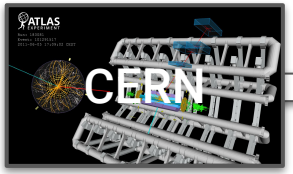
1. Context

Trends

> "A supercomputer is a device for turning compute-bound problems into I/O-bounds problems."
>
> *[Kenneth E. Batcher, Kent State Univ.]*

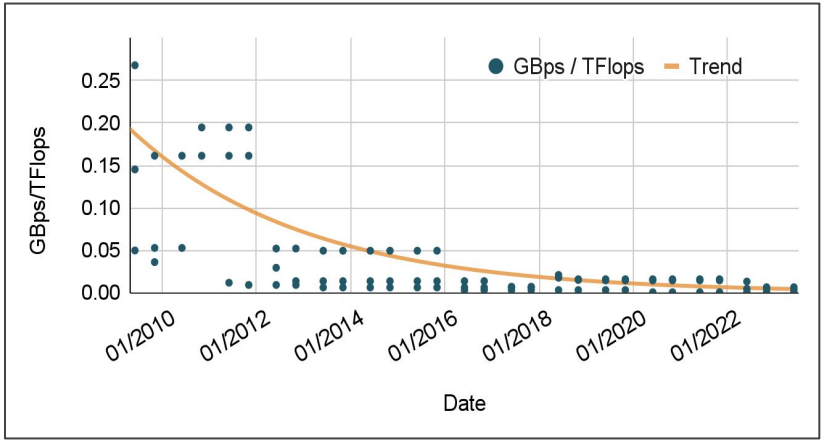


2023: **40 TiB / day**
"Shortly": **180 TiB / day**
"Near future": **700 TiB / day**

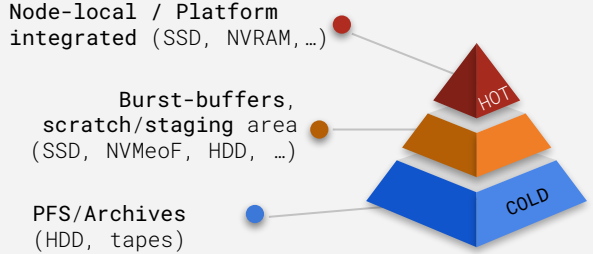


2021: **240 Gb/s** storage bw (T1)
2023: **> 1 EiB** of storage
2027: **2.4 Tb/s** storage bw (T1)
~350 PB / year (raw data)

2022: **2 PB** dataset
2023: **80 PB** generated by a **single job**
2023: **700 PB** storage system on **Frontier**
has only a **90 days retention policy**

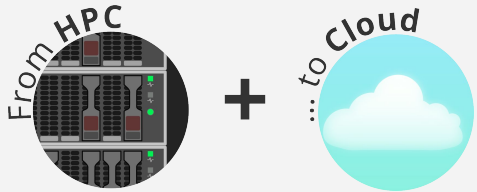


Trends



Deep storage hierarchy

New storage technologies



Hybrid infrastructures

→ Lustre FS

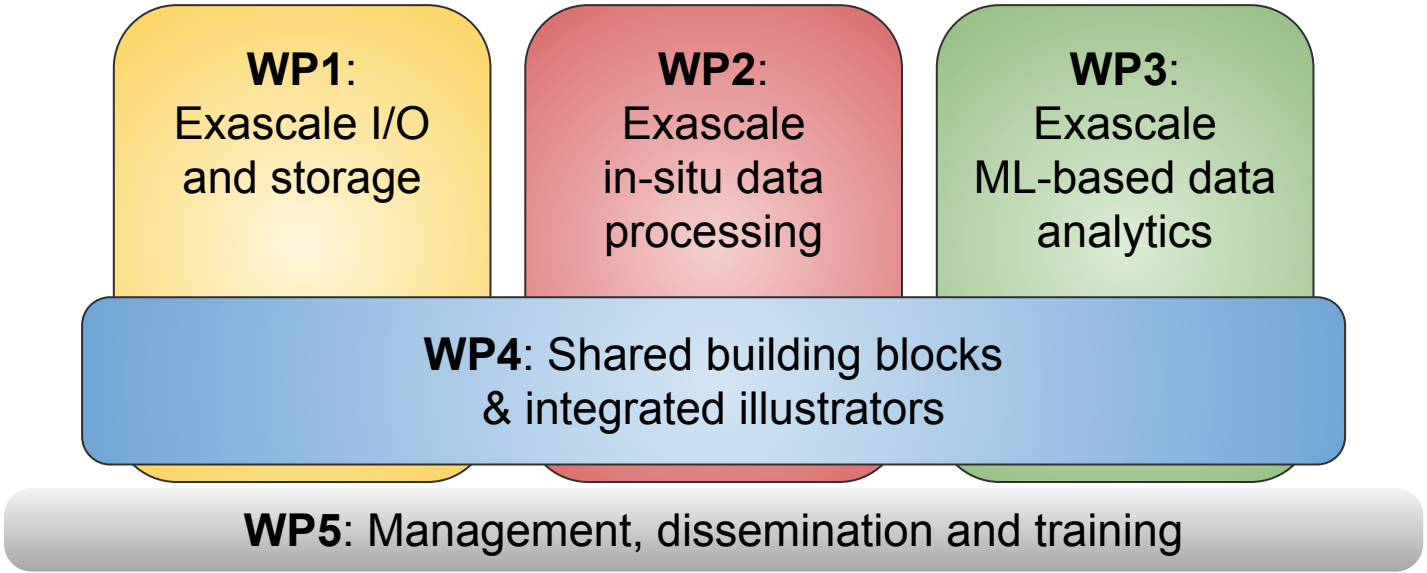
→ 679 PB capacity tier
- 47,700 HDD
- ~5 TBps

→ 12 PB performance tier
- 5,400 NVMe
- ~10 TBps

Vertical and horizontal scaling

NumPEx Exa-DoST

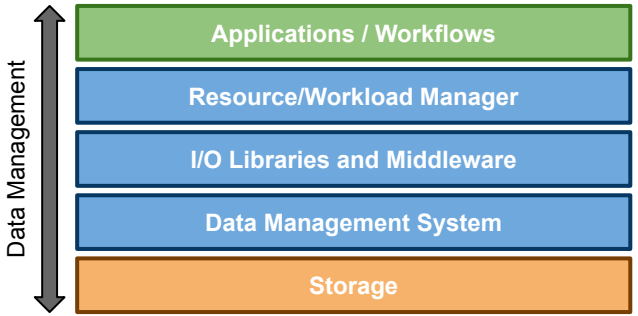
Data-oriented tools and software



NumPEx Exa-DoST - WP1 Objectives

Optimize the I/O performance of applications and workflows, and leverage emerging storage technologies

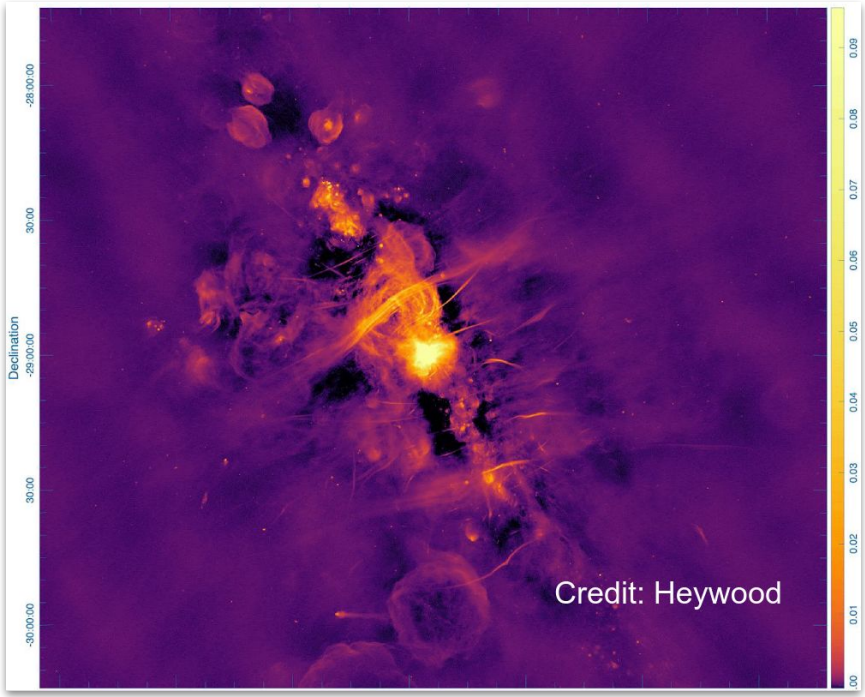
- **Support the I/O and storage requirements** of complex simulation/analytics/AI workflows running on hybrid HPC (+cloud, +edge) systems
- Promote **efficient I/O resource usage**
- Make the **I/O infrastructure adaptable to applications**' characteristics
- **Scale up modern I/O** and data storage methods and tools
- Develop and integrate **new output formats** for checkpoint/restart and for scientific analysis



2. The SKA use-case

MeerKAT, a Precursor of SKA

- Radio telescope consisting of 64 antennas in the Meerkat National Park, South Africa
  - Launched in 2018
  - Dish diameter: 13.5m
- Africanus: a Cloud-based PyData radio astronomy research ecosystem
  - Ease the development of new tools and algorithms



SC24 Atlanta, GA | hpc creates.

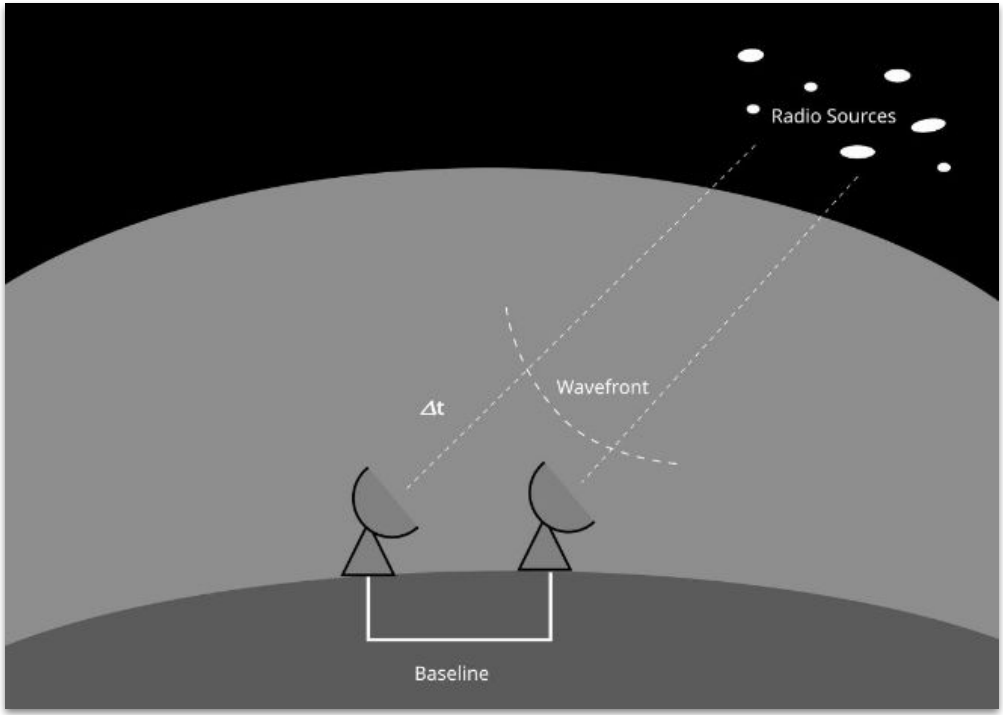
SC'24 SuperCompCloud Workshop

"The Africanus Radio Astronomy Ecosystem"
Dr. Simon Perkins of the South African Radio Astronomy Observatory

Source: Simon Perkins, Jonathan Kenyon (SARAO)

On the Scientific Instruments Side

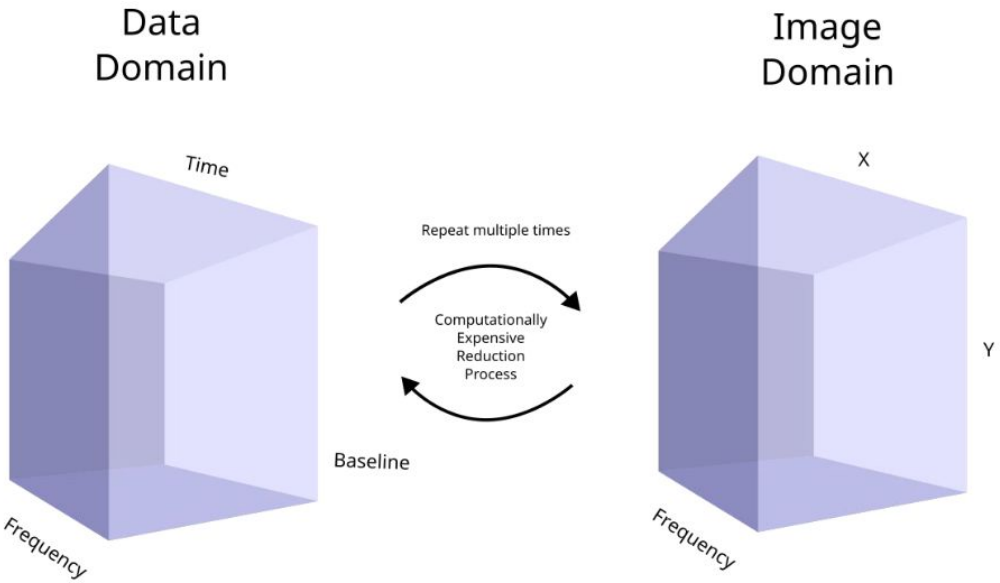
- Antenna measure signal at
  - Time
  - Frequency
- Same emission received at slightly different times (Δt) by an antenna pair
- Signal correlated along a baseline
- Number of Baselines quadratic in the number of antenna
- **Grid: (time, baseline, frequency)**



Source: Simon Perkins, Jonathan Kenyon (SARAO)

On the Scientific Instruments Side

- Antenna measure signal at
  - Time
  - Frequency
- Same emission received at slightly different times (Δt) by an antenna pair
- Signal correlated along a baseline
- Number of Baselines quadratic in the number of antenna
- **Grid: (time, baseline, frequency)**



Source: Simon Perkins, Jonathan Kenyon (SARAO)

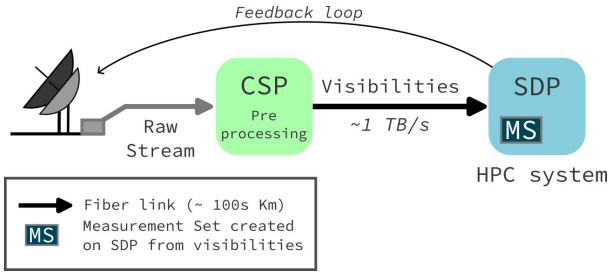
The Data Challenge

- Time, baseline (quadratic) and frequency sampling increasing

Instrument	Dump Rate	Antenna	Baselines	Channels	TB/hour
MeerKAT	8 seconds	64	2,016	4K/32K	0.237/1.896
SKA-MID	2 seconds	197	19,306	64K	145.75

Source: Simon Perkins, Jonathan Kenyon (SARAO)

- Huge stream of data + limited buffer capacity = need for continuous fast processing
- Current processing pipelines needs to scale
- **In Exa-DoST: focus on I/O and data access in general**



4. QuartiCal

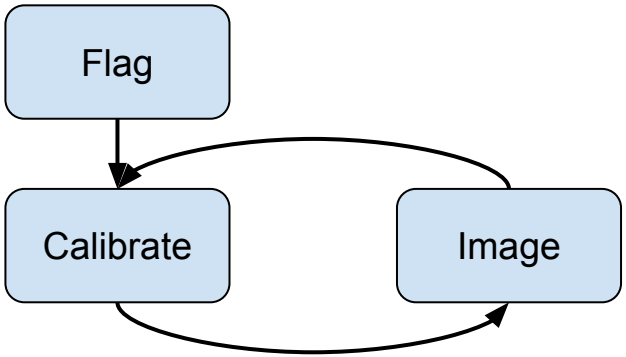
SKA - QuartiCal

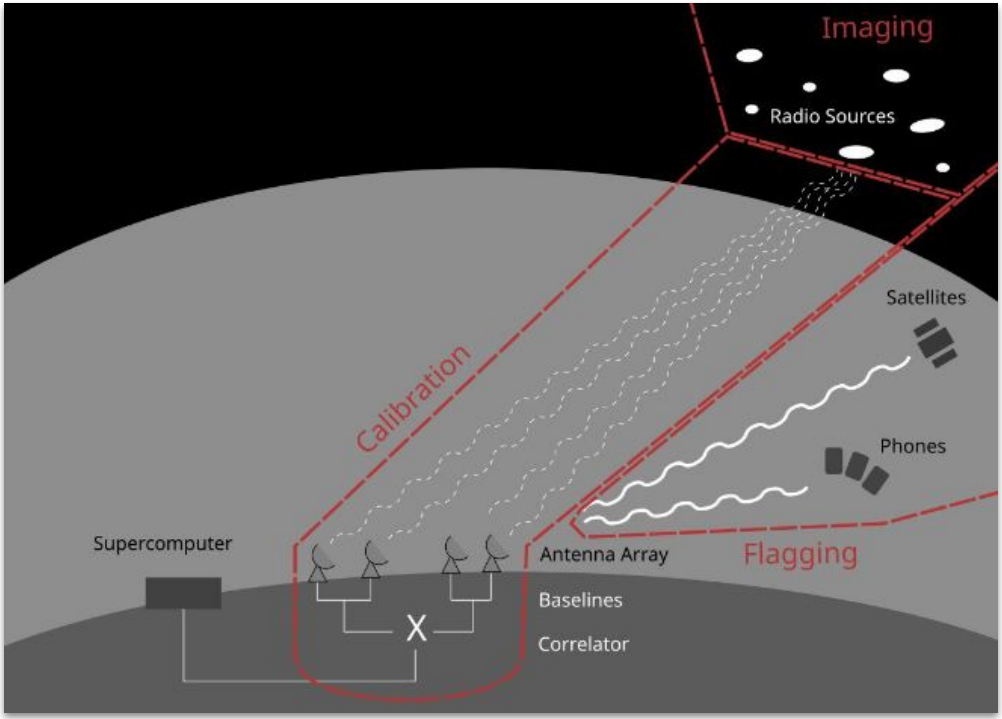
- Fast radio interferometric calibration routines exploiting complex optimisation
- Python code, developed by the MeerKAT team in South Africa (Jonathan Kenyon, Simon Perkins)
- Ugo Thay's M1 internship in the Inria KerData team (Rennes):

I/O monitoring of QuartiCal

SKA - Radio Astronomy Algorithm Cycle

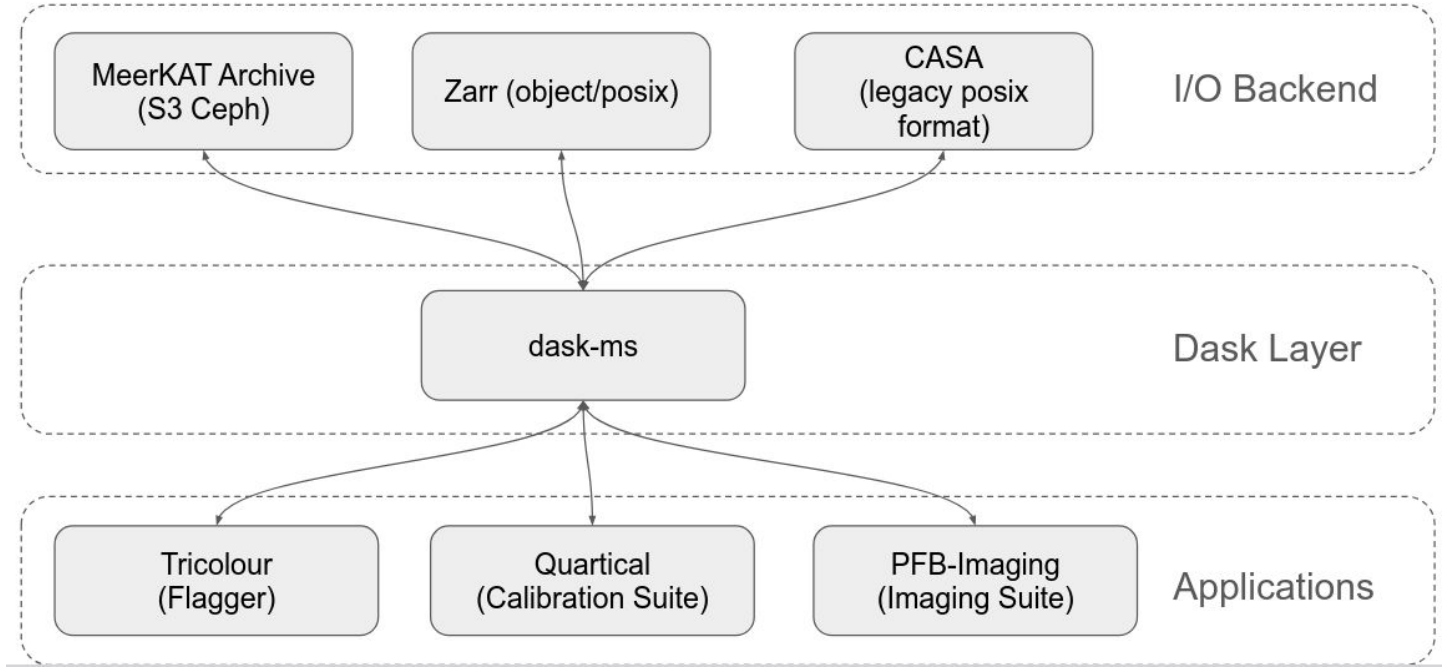
- Flagging
  - Remove Radio Frequency Interference
- Calibration
  - Account for Systematic Error
- Imaging
  - Transform Data into the Image Domain





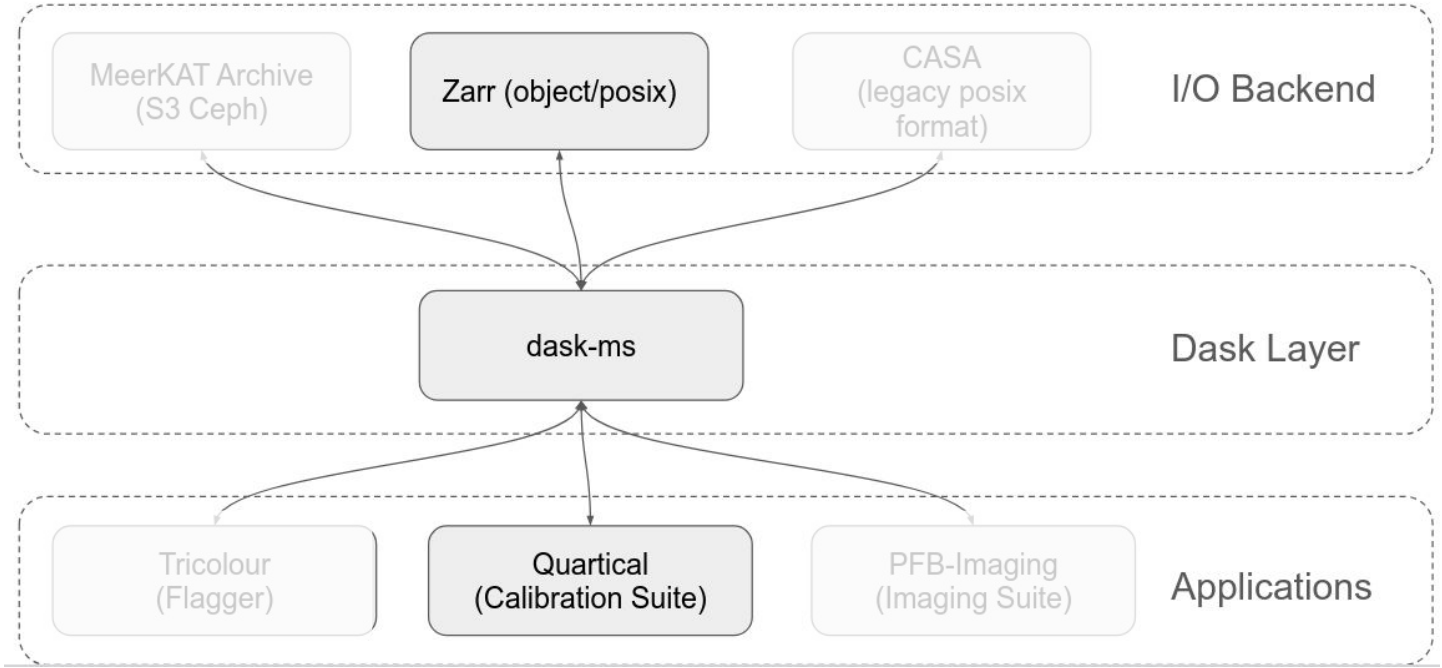
Source: Simon Perkins, Jonathan Kenyon (SARAO)

SKA - QuartiCal



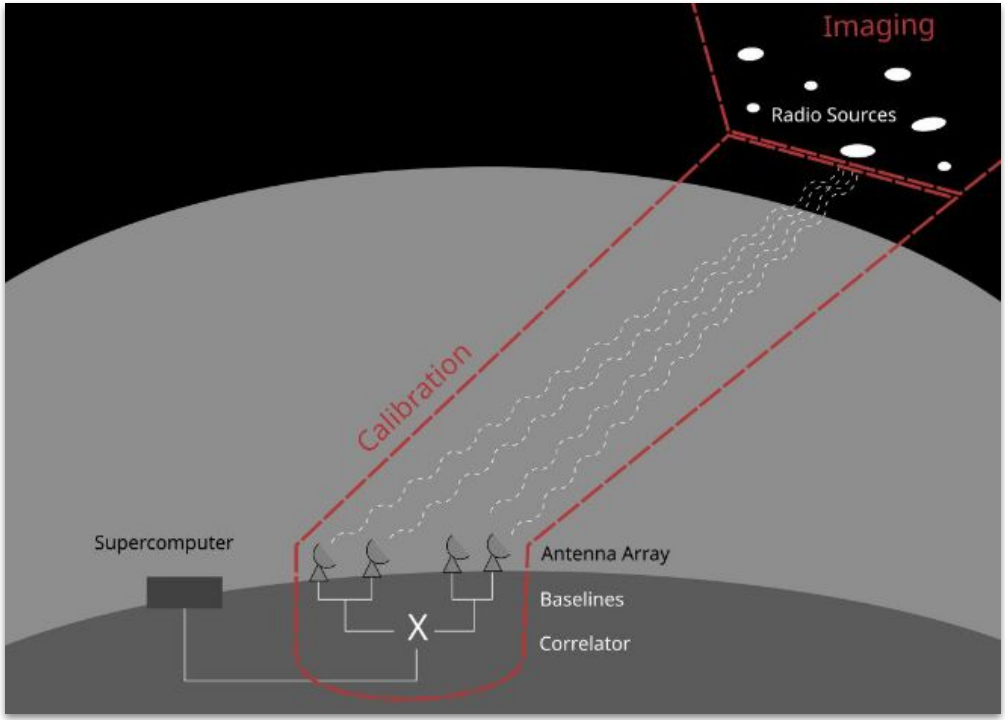
Source: Simon Perkins, Jonathan Kenyon (SARAO)

SKA - QuartiCal



SKA - QuartiCal

- Accounts for Systematic Error
- Non-linear least squares (NNLS) problem
  - Minimize the impact of errors
- Dense Linear Algebra
  - 2x2 Jones Matrix
    - Describe the polarization state of an electromagnetic wave, and its evolution
  - Highly Configurable
- Subdivide into Time-Frequency chunks
- Embarrassingly Parallel



Source: Simon Perkins, Jonathan Kenyon (SARAO)

SKA - QuartiCal

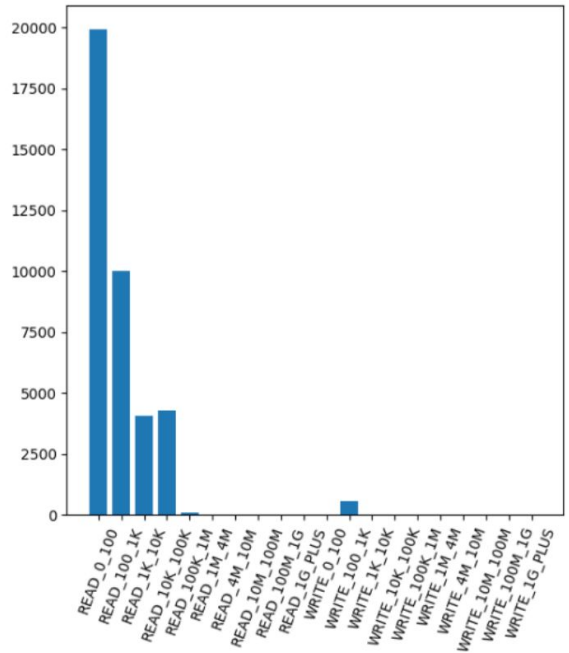
- Experiments on PlaFRIM (cluster in Bordeaux)
  - Single node, multiple nodes (up to 4 nodes)
- Combination of Darshan and strace for I/O tracing
  - Darshan: I/O monitoring tool from ANL. ~No overhead but designed for MPI applications
  - strace: syscall tracer. High overhead, high level of details.

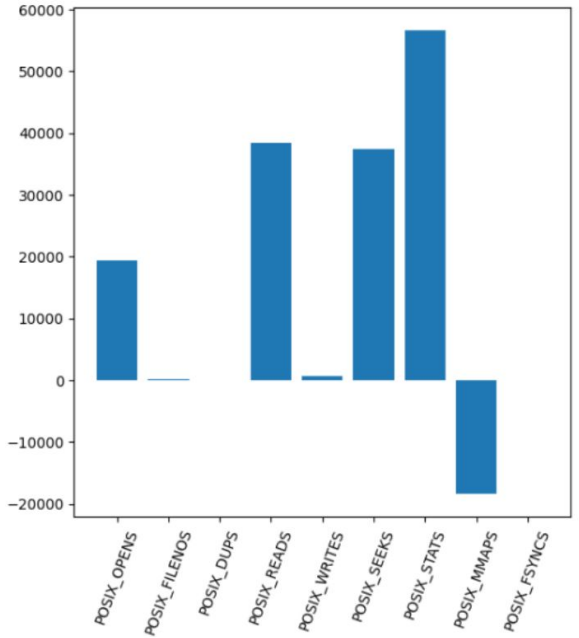
Appel POSIX	Darshan	strace
READ	28315	123294
WRITE	387	4619
OPEN	14196	87779
SEEK	27788	109089
STAT	~46000	421426

- Performance analysis using PyDarshan, darshan-utils, numpy/matplotlib, JupyterLab

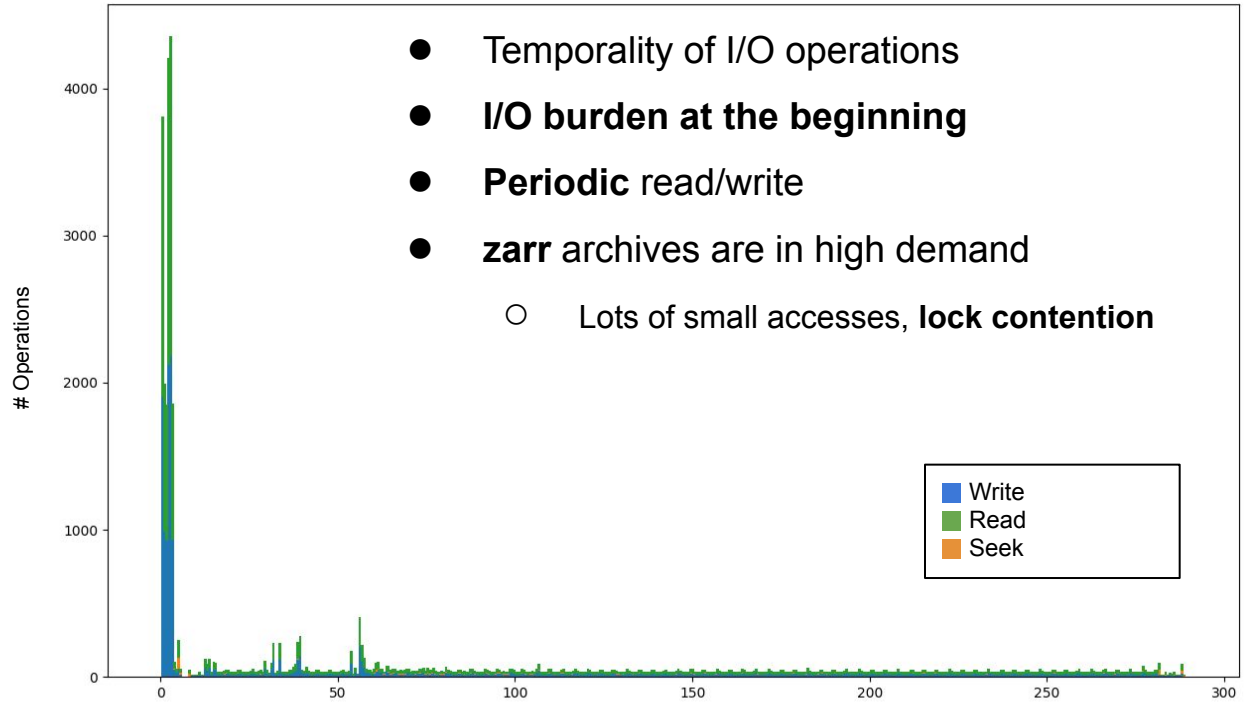
SKA - QuartiCal

- Distribution of I/O operations
- QC is **read and metadata-intensive**
  - Lots of small files accesses
  - Lots of seek/stat operations





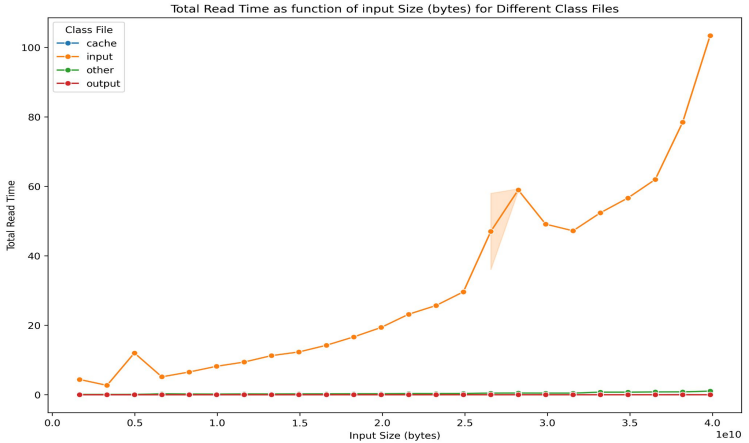
SKA - QuartiCal

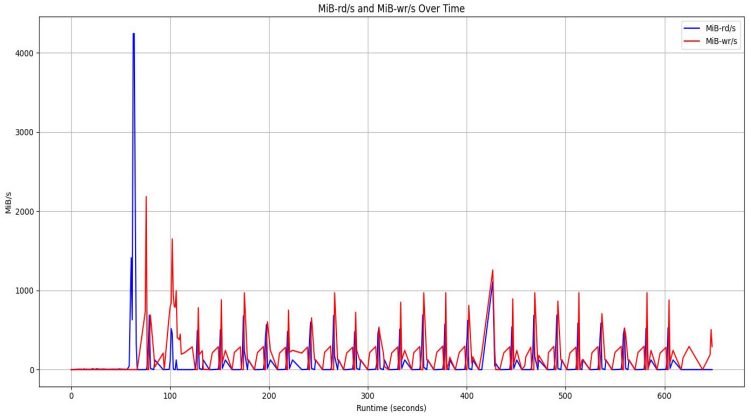


3. The DDF pipeline

DDF: KillMS + DDFacet

- Experiments with a single-node version
  - Iheb Becher's internship (LAB + Inria TADaaM)
  - Antsa Rasamoela, Valentin Hazard, Luan Teylo, Francieli Boito
- Read time is ~15% of execution time
  - scalability may be an issue
- The application **writes and reads in a "cache" folder** (~50GB when input is ~40GB)
  - output is negligible
  - the "cache" is reused by next steps of the pipeline
- Reads and writes throughout the execution
- **Darshan is NOT able** to properly profile the write operations





5. Perspectives

Next Steps

- Confirm the **zarr bottleneck** in QuartiCal
  - Data structure, benchmarking
- Continue the work on **profiling the DDF pipeline**
  - Improve it so it will use the I/O infrastructure better
  - Propose improvements to the I/O infrastructure that will benefit it
- Study and improve the libraries that handle the **MS format**
  - **Casacore** seems to not be fully parallel-IO ready (work exists already)
  - **Open M2 internship position** (KerData, Eviden)
- Extend our work to the Exa-AtoW **DDF pipeline**

Ideas? Suggestions? Want to collaborate?

Write to us! francieli.zanon-boito@u-bordeaux.fr and francois.tessier@inria.fr

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030

PROGRAMME
DE RECHERCHE

NUMÉRIQUE
POUR L'EXASCALE

Retrouvez toutes nos actualités

NumPEx