Comprendre les besoins IO de SKA : un retour d'expérience

Atelier Technique du Laboratoire Eclat

jtacquaviva@ddn.com

November 28, 2024

The Square Kilometre Array, world's largest radio telescope

10 light year away
The SKA will be so sensitive that it will be able to detect an airport radar on a planet at this distance

2'000'000 years
The data collected by the SKA in a single day would take nearly two million years to playback on an ipod

1'000'000+
of 500GB laptops can be filled with SKA data every year

On two sites

South Africa
SKA1-MID

≈200
dishes

5x
more sensitive than any other radio telescope

33'000 m²
of total collecting area
(=126 tennis courts)

Western Australia
SKA1-LOW

≈130'000
antennas spread between 500 stations

8x
more sensitive than any other radio telescope

420'000 m²
of total collecting area
(=58 football pitch)

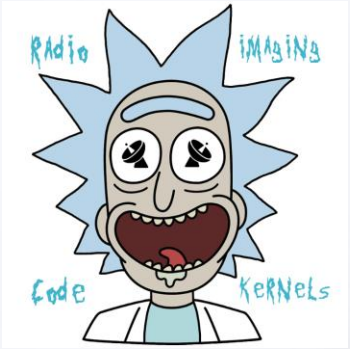
Already supported by : Australia, Canada, China, France, India, Italy, the Netherlands, New Zealand, South Africa, Spain, Sweden and the United Kingdom.

Dealing with uncertainties

- Extreme scale project
  - Path finders have been deployed, but no reference sites matching size and complexity
- Long term project in a fast-evolving technological environment
  - CPU / GPU / SSD / HDD / IB / ETH
- Science still dealing with unknown
  - Intrinsic to the nature of the scientific problems addressed
- Code stack not yet fully sedimented
  - Multiple code variants

Dealing with uncertainties: example RICK

- RICK (Radio Imaging Code Kernels) is a code that addresses the **gridding**, **FFT** and **w-correction**, combining parallel and accelerated solutions.
- It is being designed not to substitute radioastronomy codes but to provide specific solutions, portable and fast
- **C**, **C++**, **CUDA, HIP** (for AMD GPUs)
- **MPI** & **OpenMP** parallel, fully working in parallel
- All the aforementioned steps can run on GPUs (both CUDA and OpenMP for GPU offloading), in particular the FFT using the distributed CUDA library **cuFFTMp**
- An optimized version of the reduce has been developed on both CPU (combining MPI+OpenMP) and GPU (using NCCL or RCCL, for Nvidia and AMD respectively)
- **Weighting** and **uv-tapering** are being implemented



Courtesy of

CLAUDIO GHELLER
INAF – INSTITUTE OF RADIOASTRONOMY BOLOGNA

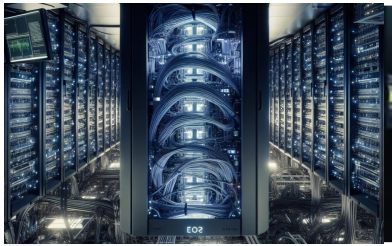
Emanuele De Rubeis (UniBO-IRA), Giuliano Taffoni (OATS), Giovanni Lacopo (OATS), Luca Tornatore (OATS), David Goz (OATS)

Dealing with uncertainties: Co-Design

- SKA-O set-up the co-design Raccoon Team
  - Design future proof architecture
  - Focus on the storage component
  - From an IT stands point SKA is a huge data acquisition and processing device
  - EPFL and DDN
- Optimized Project Management
  - Risk mitigation
  - Considering size of the project, even a modest improvement leads to strong ROI
- Bring together science and technology expertise
  - IT specialists are seldomly radio-astronomers :-(
  - Radio-astronomers are skilled in IT, but we have our secret sauce :-)

Why DDN? We deliver at Scale

THE AI DATA COMPANY



NVIDIA Eos
World's Fastest SuperPOD

NEC AI Research
Japan

Berzelius
Linköping University

Cambridge-1
UK Life Sciences

PARAM Siddhi AI
India Research and R&D

NAVER AI Cloud
South Korea AI Services

HiPerGator
Uni of Florida

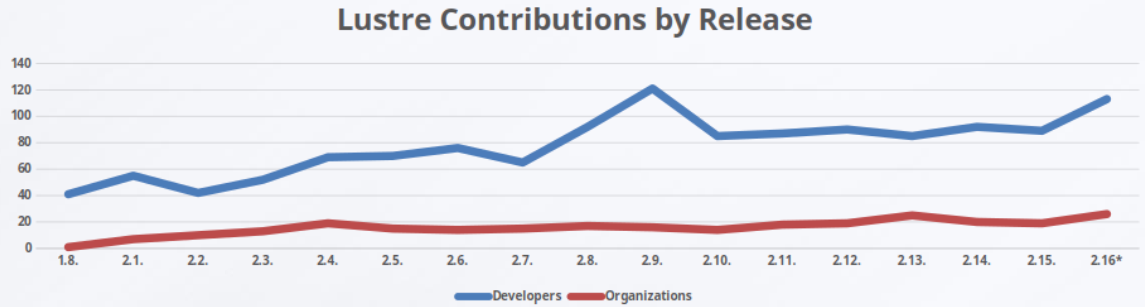
Lambda
US Cloud SuperPOD

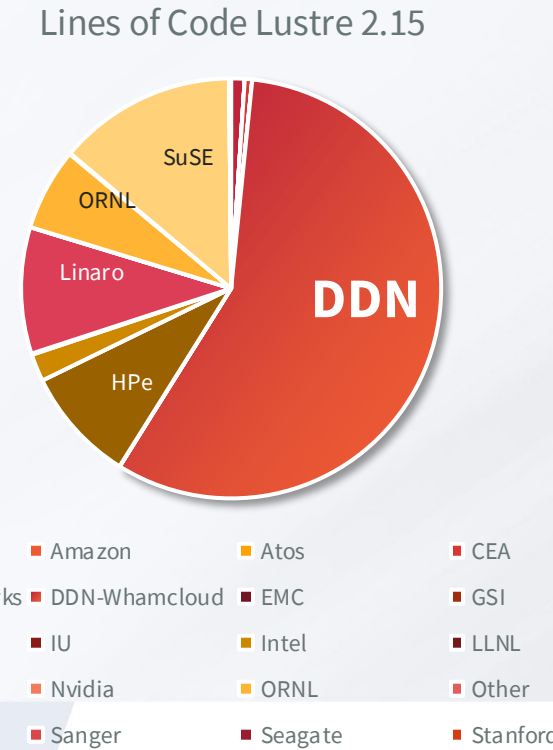
NVIDIA Selene
World's First SuperPOD

Scaleway
Europe Cloud SuperPOD

Why DDN? An Open-Source Driven Company

Lustre Open-Source Parallel FRile system (OpenSFS)





- Designed for HPC: data extension of the compute platform
- DDN is the lead contributor to Lustre
  - User meetings in Europe organized by EOFS
  - User meetings in Asia organized by DDN
- Open-source shields form vendor locking
  - Strong asset for long-terms projects

Why DDN? Open to the Ecosystem

A – EuroHPC Infrastructures powered by DDN

Leonardo


Meluxina

Discoverer

Vega


Deucalion

B – Collaborative Research Programs and DDN

- EuroHPC design of next-Gen IO system
- AI-automated features extractions from Satellite Images
- Nuclear Fusion code optimization
- Excalibur (UK)
- Sandia National lab.
- RIKEN: AI for Science

C – DDN R&D Spending in Europe

- Significant portion of our WW turnover is spent on R&D
- 25 persons R&D in EU
- Lab set-up in CINECA

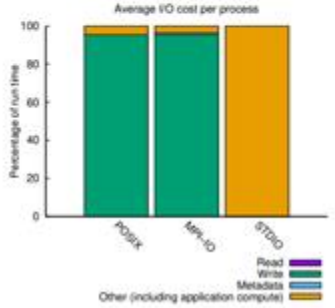
Cross-Fertilization: Performance Analysis

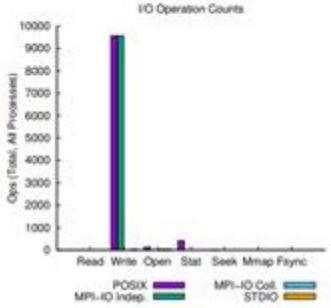
Workshop on IO tracing e.g. Darshan

Courtesy L. Bellentani from CINECA

jobid: 6134195	uid: 33678	nprocs: 32	runtime: 7.7400 seconds

I/O performance *estimate* (at the MPI-IO layer): transferred 76293.9 MiB at 9974.31 MiB/s
I/O performance *estimate* (at the STDIO layer): transferred 0.0 MiB at 14.47 MiB/s

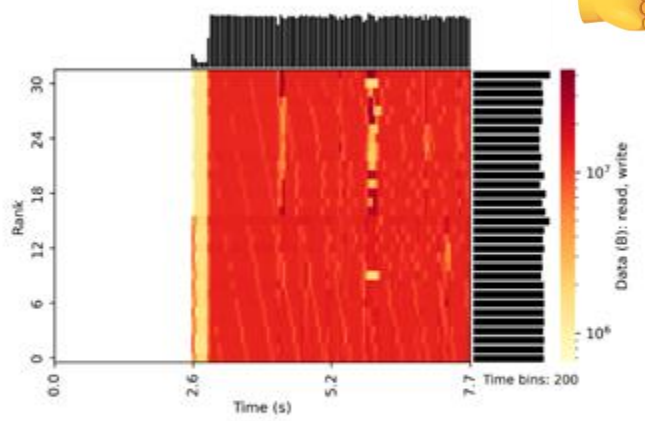




Variance in Shared Files (POSIX and STDIO)

File Suffix	Processes	Fastest			Slowest			σ	
		Rank	Time	Bytes	Rank	Time	Bytes	Time	Bytes
...misterio.out	32	11	6.331536	2.4GiB	26	7.412161	2.4GiB	0.264	0
...<STDOUT>	32	16	0.000005	49B	8	0.000021	49B	0	9.76

Heat Map: DXT_MPIIO



Heat map of I/O (in bytes) over time broken down by MPI rank. Bins are populated based on the number of bytes read/written in the given time interval. The top edge bar graph sums each time slice across ranks to show aggregate I/O volume over time, while the right edge bar graph sums each rank across time slices to show I/O distribution across ranks.

Cross-Fertilization: Compression

Conduct testing to validate the client-side compression features from EXAScaler / Lustre

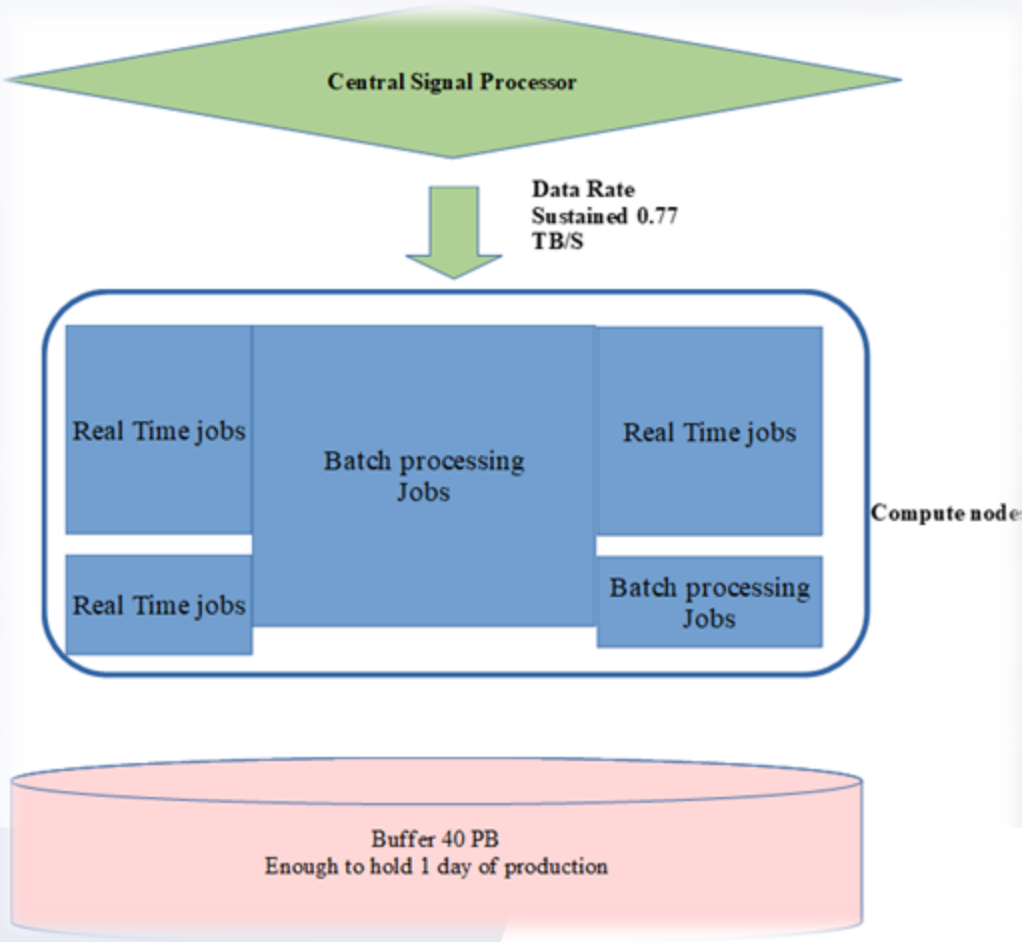
- **Real data-set provided by Shan Mignot / SKA-O**
- **Initially encouraging results were not validated**
- **Entropy prevents significant compression ratio**

Engage discussion with DotPhoton

- **No low hanging fruits**

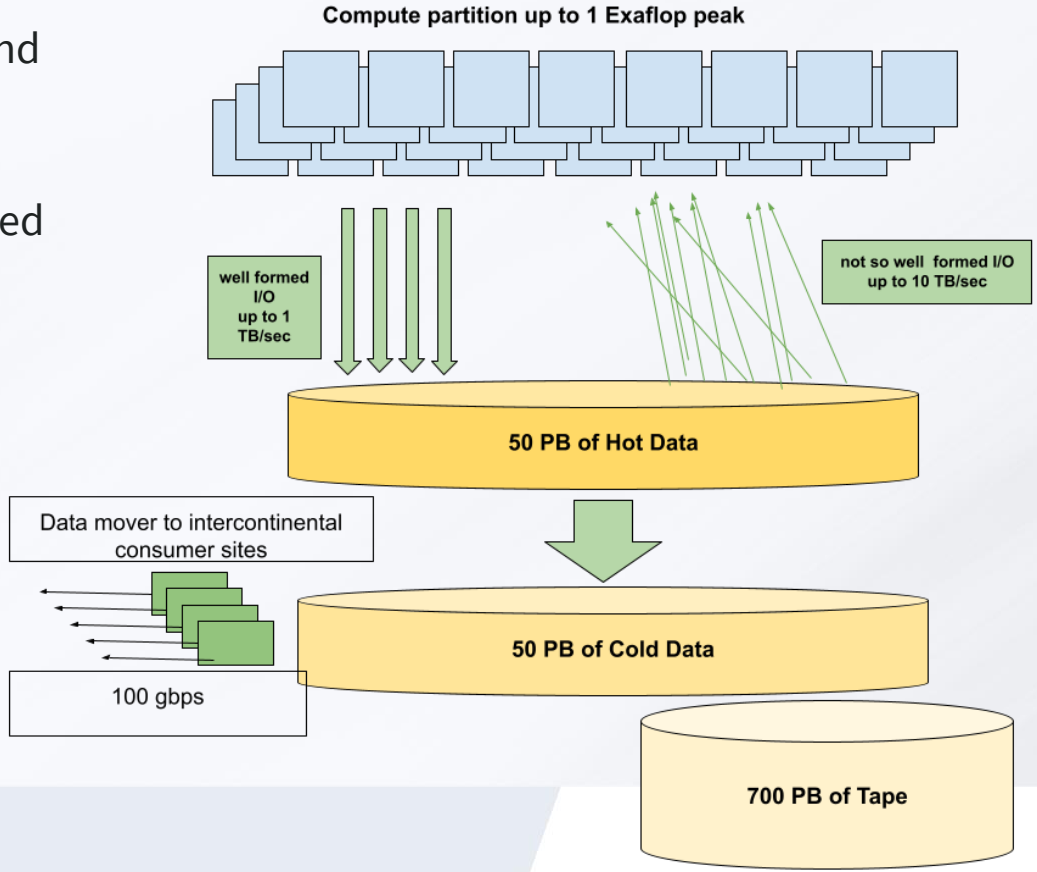
Extreme Scientific Challenges: key numbers

- 40 PB / day of raw observational data
- 700 PB / year of scientific data product
- Read Write ratio: 10:1
- Focus on power requirements
- 24/24 7/7 with mixed workload
- Radio-Telescope life expectancy 50 years

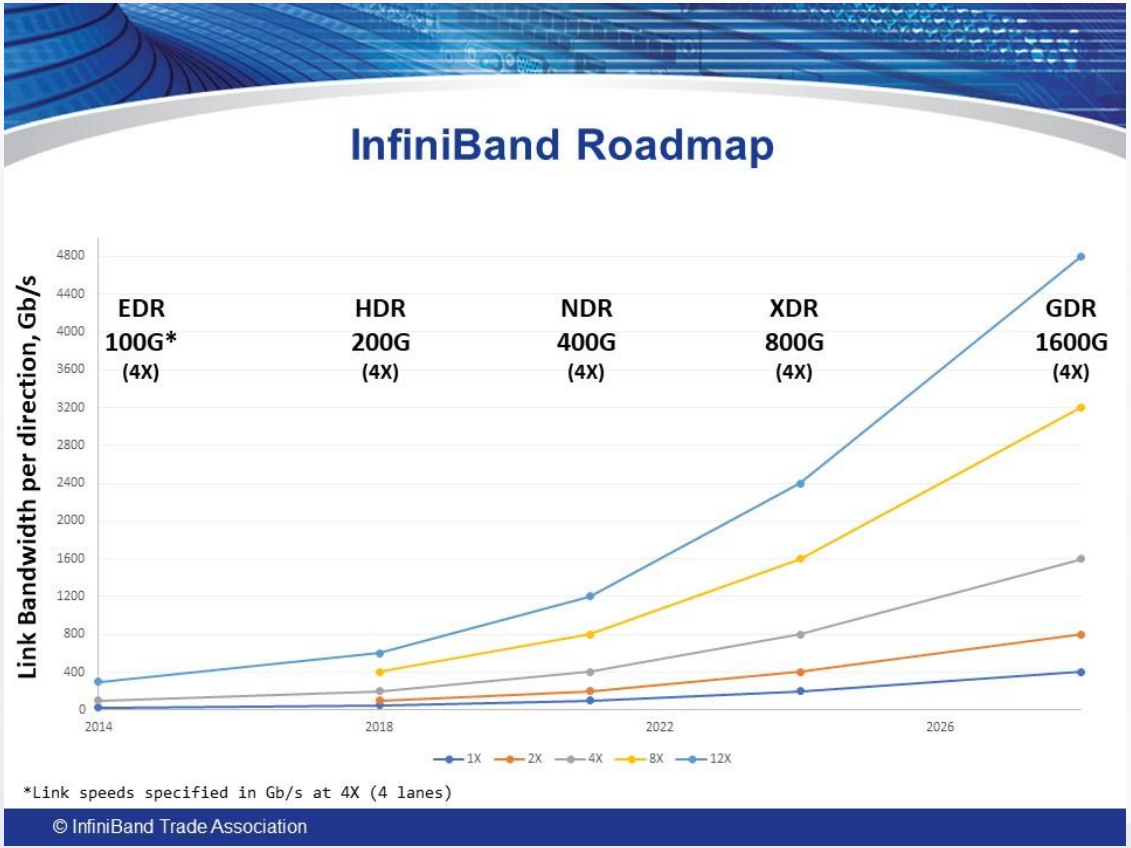


CPFs: a possible Architecture

- Central Processing Facilities in Australia and South African follows the same design patterns
- Scientific results produces by CPF are stored as cold data and shared to outer world
- Outer world can also browse tape archive
- A scalable metadata catalog is needed to interface CPFs and outer world



CPFs Hardware discussion: Network link



10 TB / sec can be achieved with

- 500 HDR200 links
- 250 NDR links
- 125 GDR link

Do we need extreme network density to achieve the data throughput ?
Data ingestion rate is also defined by the storage devices capabilities

CPFs Hardware discussion: PCI lane

PCI Generation	Throughput per PCI lane	Initial release of the specification	Market Availability
Gen. 4	1 GB/s	2017	2019
Gen. 5	4 GB/s	2019	2023
Gen. 6	8 GB/s	2022	2024
Gen. 7	16 GB/s	2025	2027

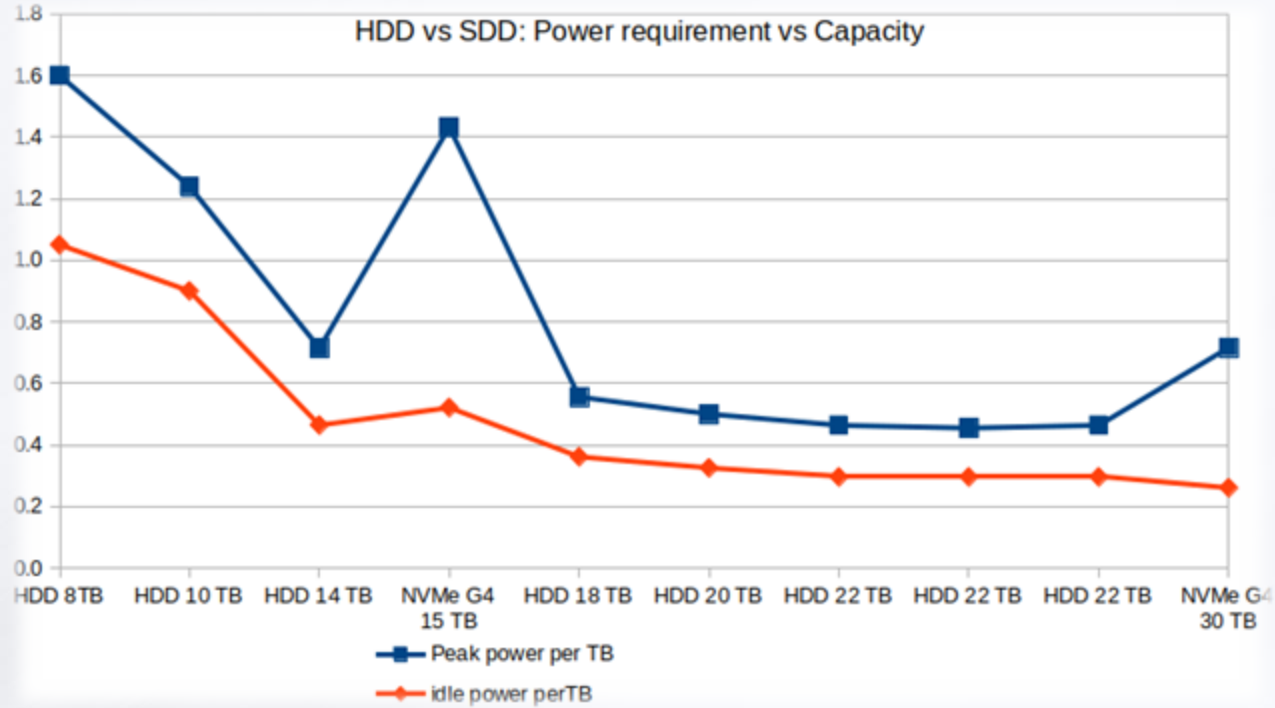
An NVMe used 2 PCIe lane
Production Gen5 device ~ 10 GB/s
10 TB / sec can be achieved with

- 1000 PCI Gen5 devices
- ~ 40/50 appliances with current form factor

- PCI technology tends to follow a 2.5-year development cycle
- Since PCI Gen 4 read and write are significantly imbalanced
  - Latest Gen 5 devices ~14 GB/s Read and ~ 8 GB/s Write
  - Imbalance due to flash (not pci) further increased by the data protection mechanism

Critical importance of crafting a carefully balanced architecture with alignment of Network / CPU / PCI / Device Bandwidths

CPFs Hardware discussion: power envelop



Technology	Watt per TB moved
HDD	5120
SDD (Gen 4)	205
SDD (Gen 5)	73

DDN AI400X3

2U, 2.4KW | **140 GB/s** *READS* | **100 GB/s** *WRITES* | **1.5M** *IOPS* | **400Gb IB/Eth** *4x OSFP connections*

DDN Virtualization Technology **Eliminates cables, switches & servers**

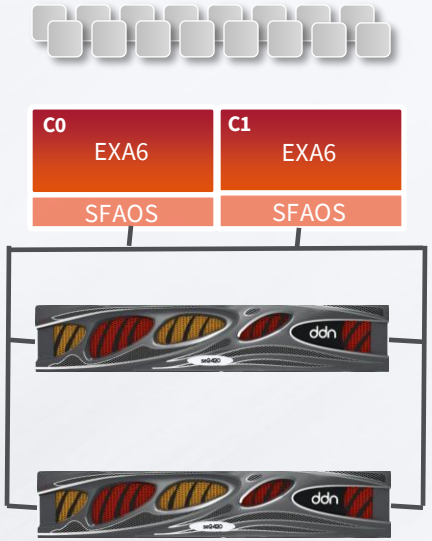
Large Capacity Flash Drives **Doubles Capacity per Watt**

End-to-End Parallel Datapath **Doubles Performance per Watt**

Workload Focused Performance Optimization **Speeds up Applications**

100% linear scale out Software **Removes costly silos**

CPFS: Hardware discussion



Intelligent Clients

Fastest Network

Fastest Servers

Fastest I/O

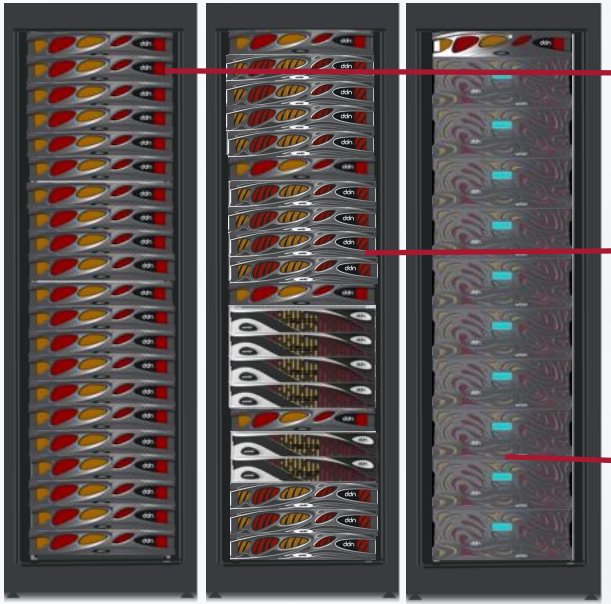
Resilient, High-Density Enclosures

Fully Redundant back-end

Massive Expansion NVMeoF or SAS

Fastest and Safest QLC, TLC SSD and HDD

TLC Flash **QLC Flash** **Hybrid**



Best IOPS & Speed
16PB of Flash
70M IOPs
2.5 TB/s

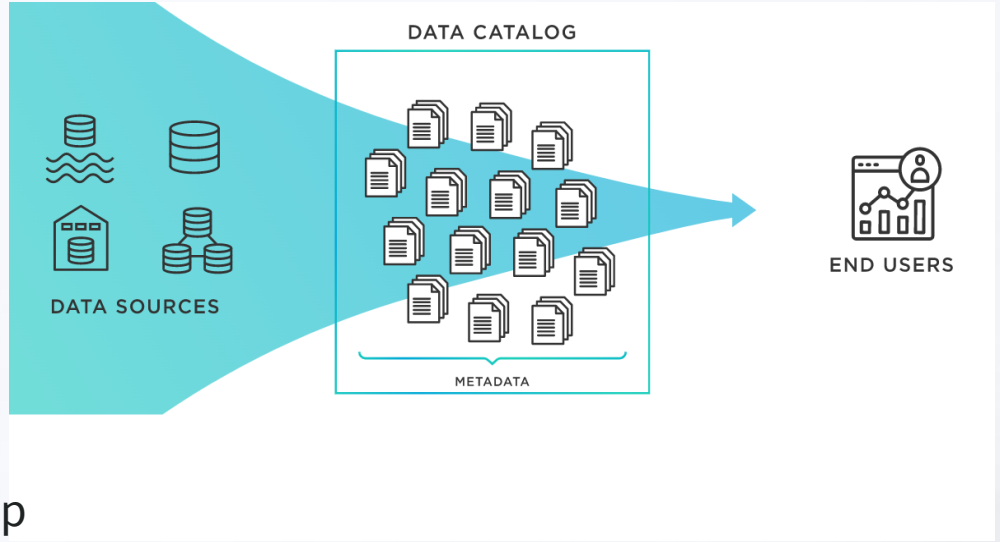
Best Price Flash
26PB of Flash
14M IOPs
400 GB/s

Best Price TB
20PB of HDD
100 GB/s of speed

Exploitation discussion: Data set management

CFP will generate a large volume of SFP to be processed by the community

Metadata catalog required

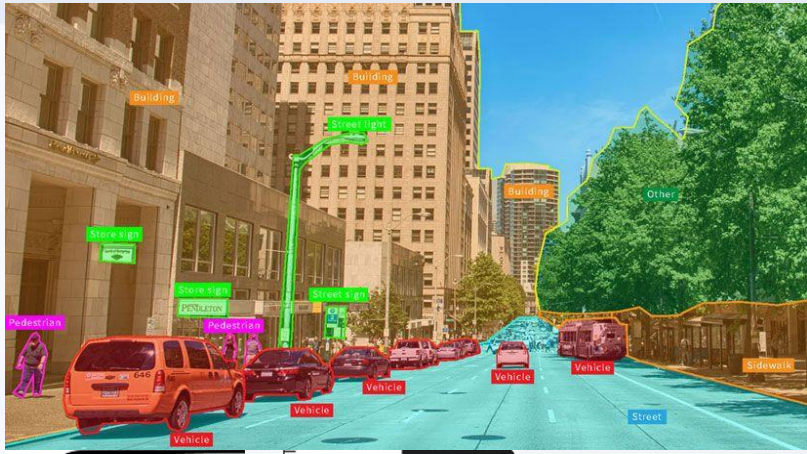


- Metadata allows to structure Datalake
- Prevents Data-lake to turn in Data Swamp
- Query-able Metadata Catalog
  - Difference between semantic and logistical metadata
  - Difference between query-able and structured

Exploitation: Metadata catalog

- Query-able Metadata: RUCIO
- DDN works with CERN on RUCIO + Lustre
  - Work conducted in a DaFab EU project
  - 3M€ / 8 partners
- Extension of the Query capabilities of RUCIO
- Support of GeoJSON metadata

- Metadata / data ratio evolves with metadata complexity
  - Metadata exceed data by a factor of 3 on some AI workoads

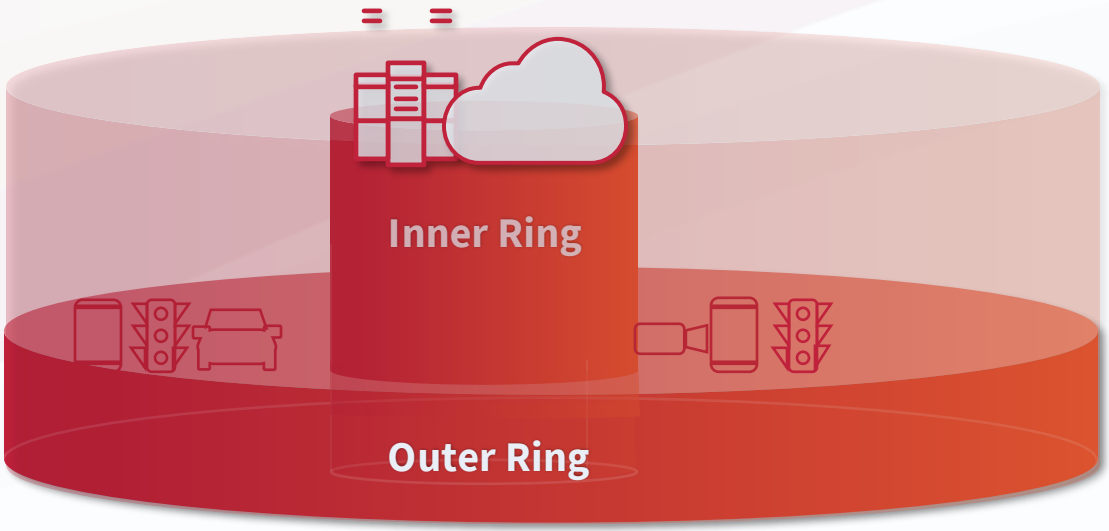


DaFAB Project: 101128693 — DaFab — HORIZON-EUSPA-2022-SPACE

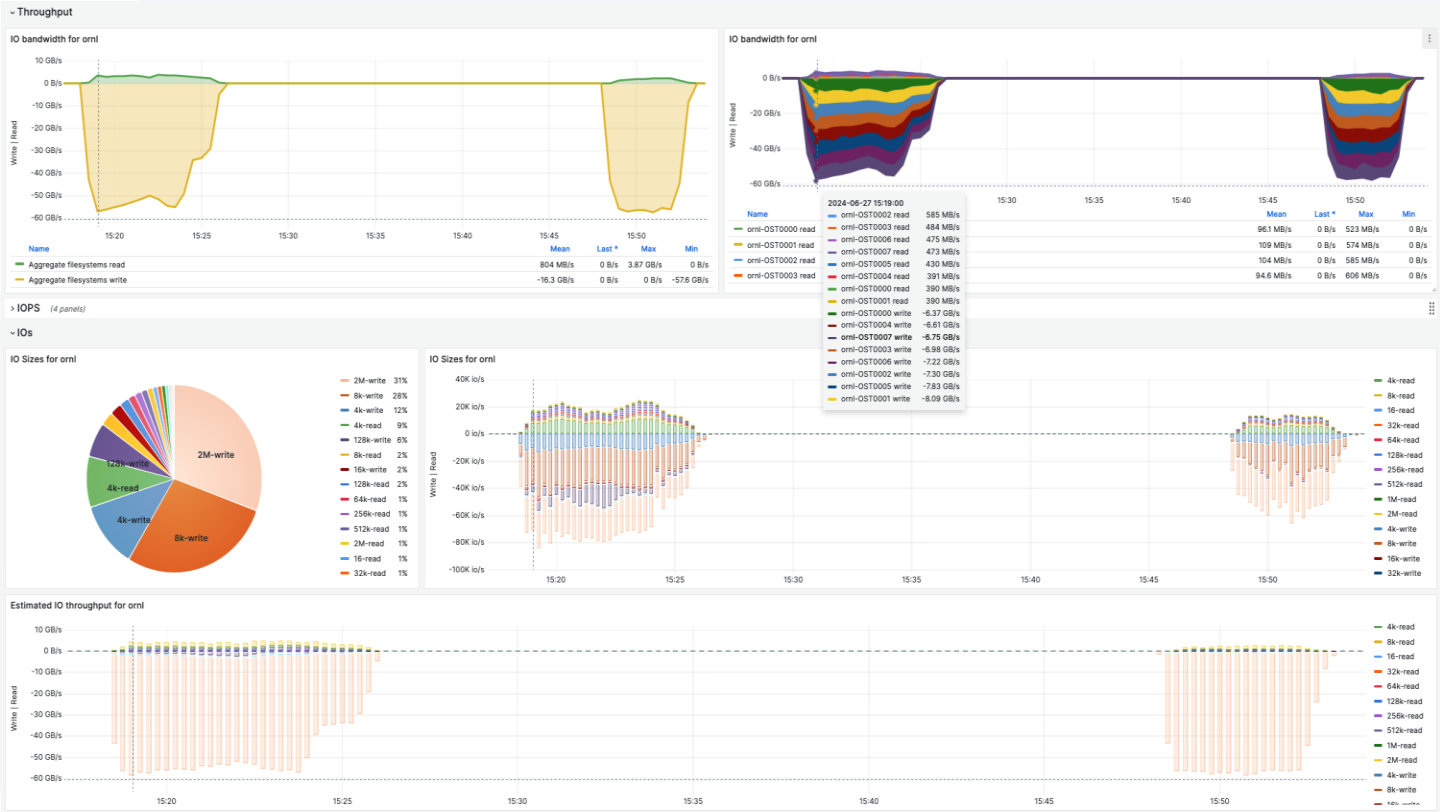
Exploitation: Metadata structuration and registration

Scientific Data Operation
Scale, Efficiency & Performance

Scientific Dataset Management
Multi-Tenancy, metadata-based governance, registration mechanism



On-going: Imaging IO Test



Benchmark Suite: Level1 imaging-iotest: multi-node analysis

Peter Wortmann from SKA-O

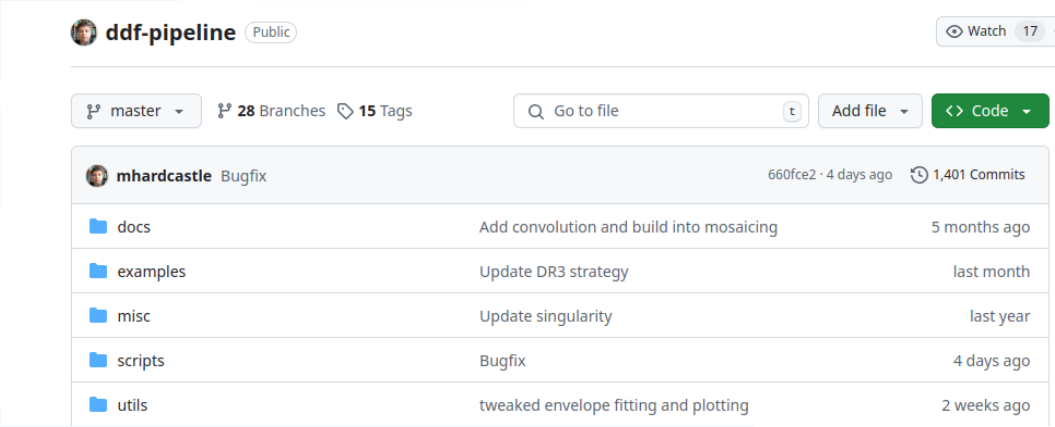
On-going: Performance Assessment

Deploy Existing scientific pipelines on existing architectures

File systems are heavily instrumented: ability to extract detailed application behavior

Cooperate with the Scientific Community

Pipelines remains highly-technical workflows



Next Step: Production and Orchestration

1. **CFP workload will evolve over the year**
2. **Usage with evolve over the year**

Future Proof Architecture

- Implementation of QoS and SLA capabilities
- Prevent over-specialization of the components
  - Specification vs risk mitigation
- Sandboxing / experimental playground
  - Argo / K8
  - Multi-tenancy


ddn