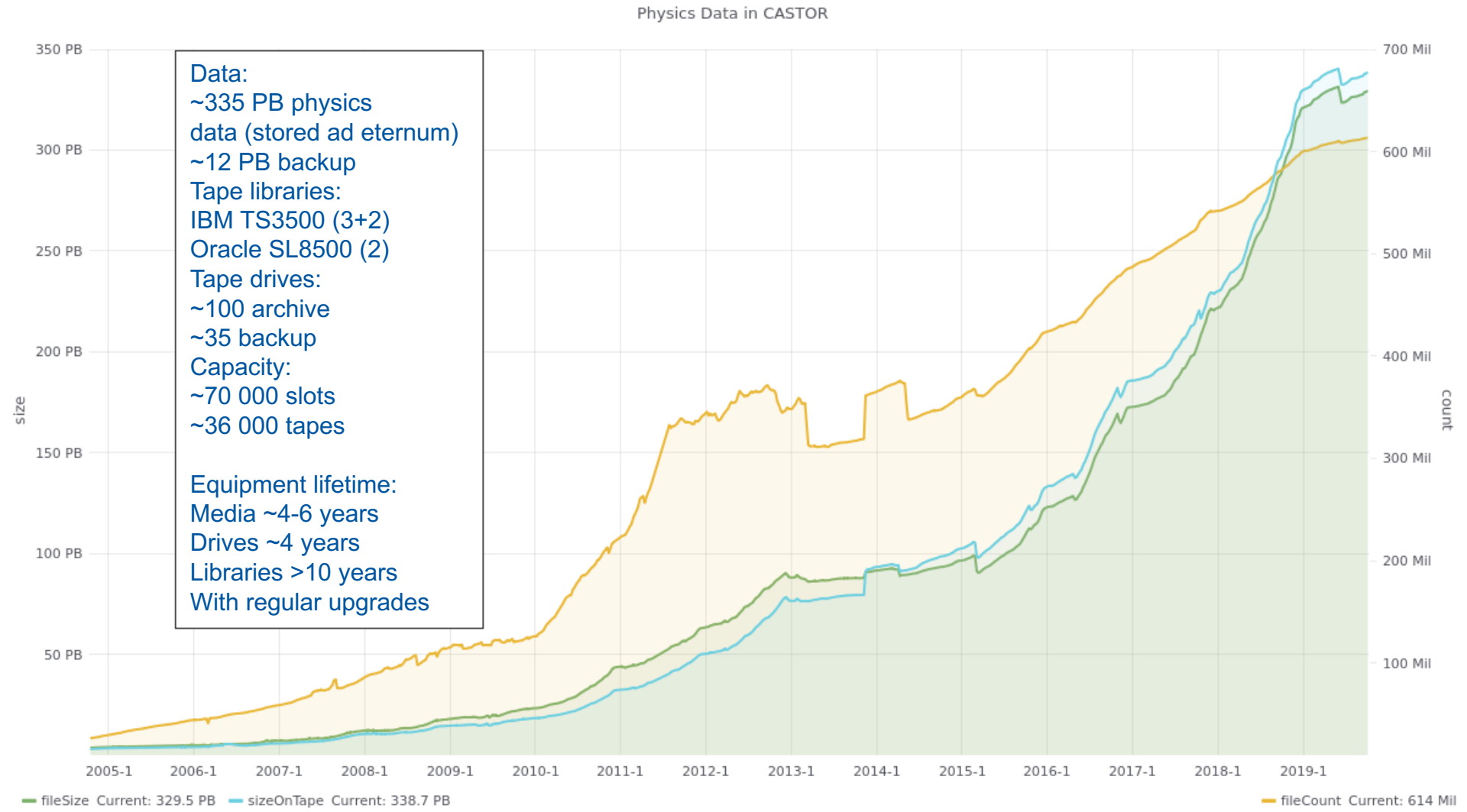
Current status of tape storage at CERN

Speaker Steven Murray

Authors

Vladimir Bahyl, Cedric Caffy, German Cancio, Eric Cano, Michael Davis, David Fernandez Alvarez, Julien Leduc, Steven Murray

Total amount of archived data up to 2019



CERN tape infrastructure 2019

- Repacked out of Oracle
  - 80 PB of data were repacked
  - Some “bottom of the barrel” files still to be processed
- Removed two Oracle SL8500 libraries
- Two Oracle SL8500 libraries remain but about to go
  - We are looking for a potential buyer
  - Please contact Vladimir.Bahyl@cern.ch if you are interested
- Converted one IBM library from TS3500 to TS4500
- Extended another IBM library to full size
  - From 5000 to 20000 slots
- Successful test of replacing FC links to tape drives with RoCE (RDMA over Converged Ethernet)
  - For enterprise drives only
  - Uses RDMA over Ethernet

Moving away from Oracle and moving towards LTO

CERN tape infrastructure 2020 plans

- Prepare for run 3
  - Write > 150 PB in one year at > 40GB/s
  - Increase number of tape drives from 100 to ~160
  - Install LTO-9 drives if available in Q4
  - Install TS1160 drives if available in Q4
- Install new LTO tape library in Q1
  - Following open tender

EOS CERN Tape Archive (EOSCTA)

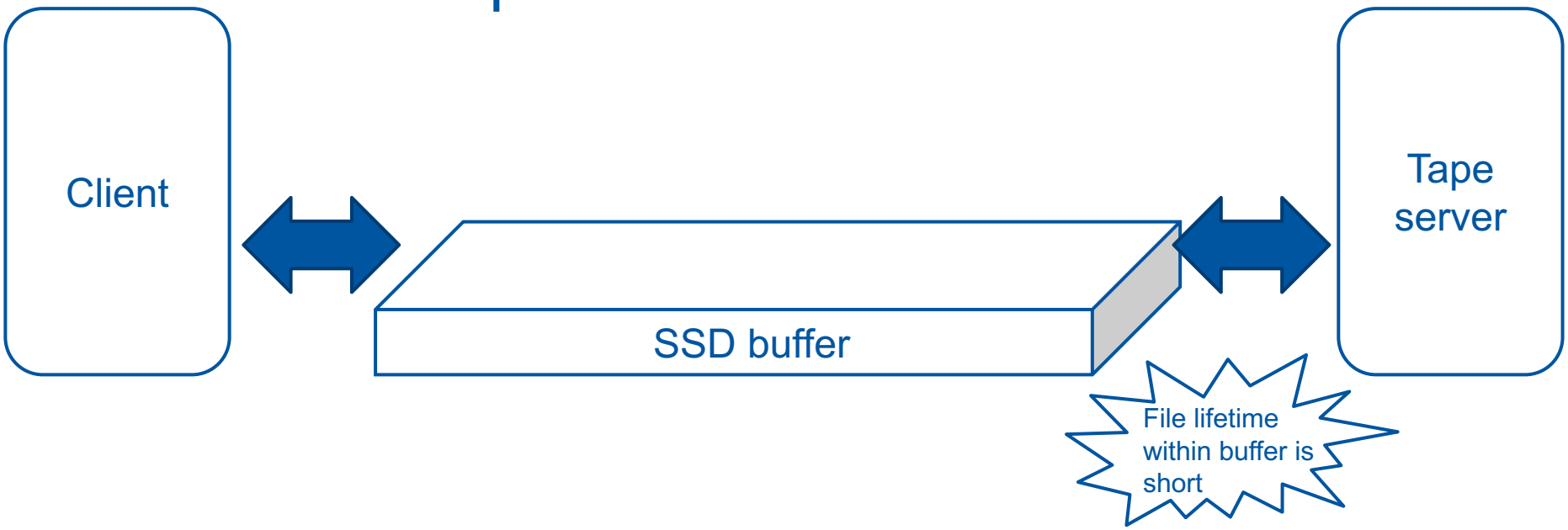
- EOSCTA is tape on the back of EOS
- EOSCTA will replace CASTOR
- EOS is the disk storage solution for LHC physics at CERN
- Natural evolution from CASTOR
  - Reused existing high performance CASTOR tape server
  - Same tape format – only need to migrate metadata
- Removed duplication between CASTOR disk and EOS
- New scheduler
  - Single step with complete global view
  - Preemptive - Use tape drives at full speed all of the time
- Single relational database with no PL/SQL business logic
  - Supports multiple database backends
  - CASTOR had 5 different databases
- Less networked components than CASTOR

EOSCTA solutions

- Ideal software bundle/solution for tape
  - FTS + EOSCTA
- Tailored solutions to match each LHC experiment
  - Alice with AliEn → EOSCTA + HSM like
  - ATLAS with Rucio → FTS + EOSCTA
  - CMS with Rucio → FTS + EOSCTA
  - LHCb with Dirac → FTS + EOSCTA (Drop SRM)

The SSD buffer of EOSCTA

Strict producer/consumer buffer



- EOSCTA is bandwidth driven
- SSD buffer solves the thrashing problem of spinners
- Scaling spinners to meet bandwidth requirements is too expensive
- The life time of a file in the buffer is short lived

EOSCTA shared vs dedicated storage

Each VO has its own **dedicated** SSD buffer

Alice ⟷ SSD buffer for Alice

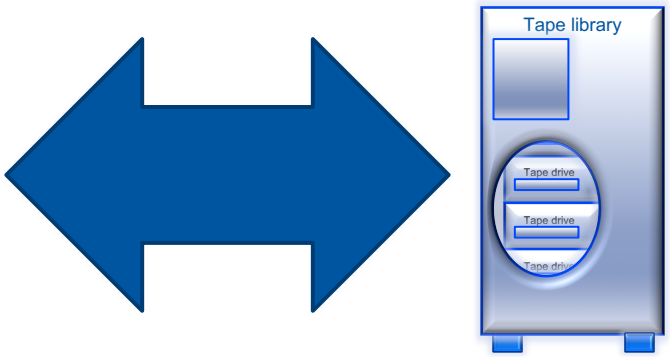
ATLAS ⟷ SSD buffer for ATLAS

CMS ⟷ SSD buffer for CMS

LHCb ⟷ SSD buffer for LHCb

Public ⟷ SSD buffer for Public

Tape drives are **shared**

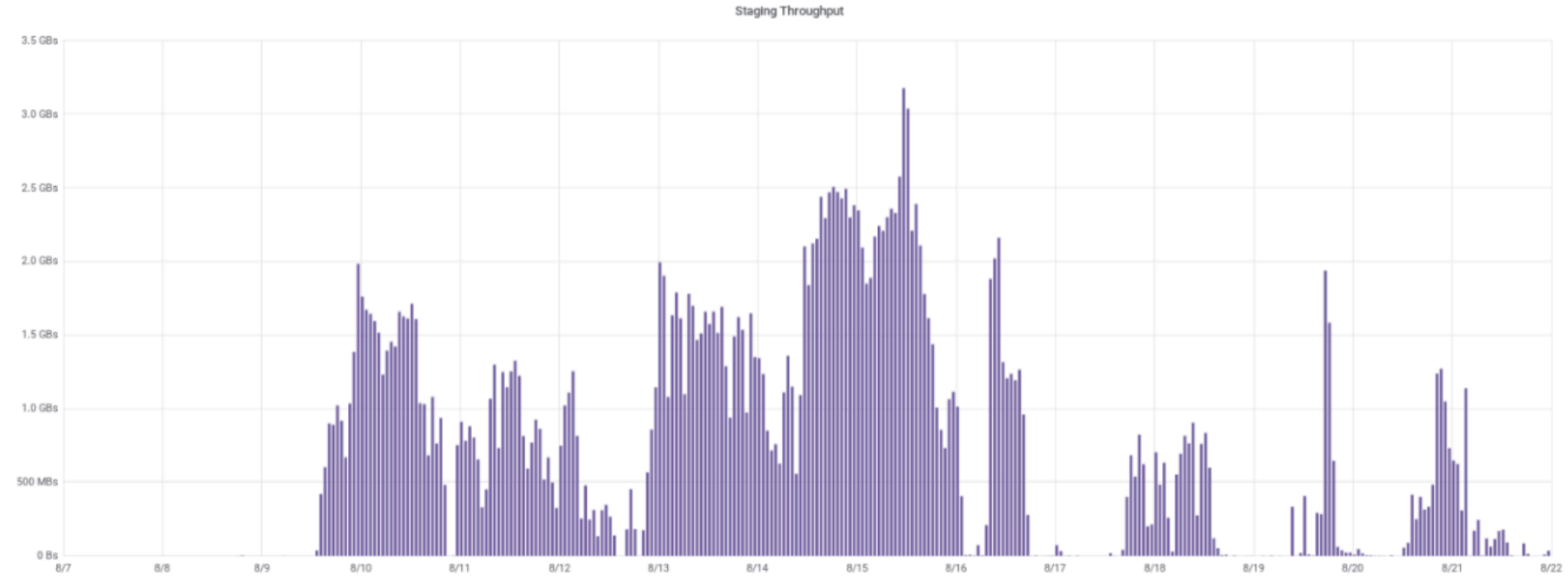


Each VO has its own **dedicated** tapes

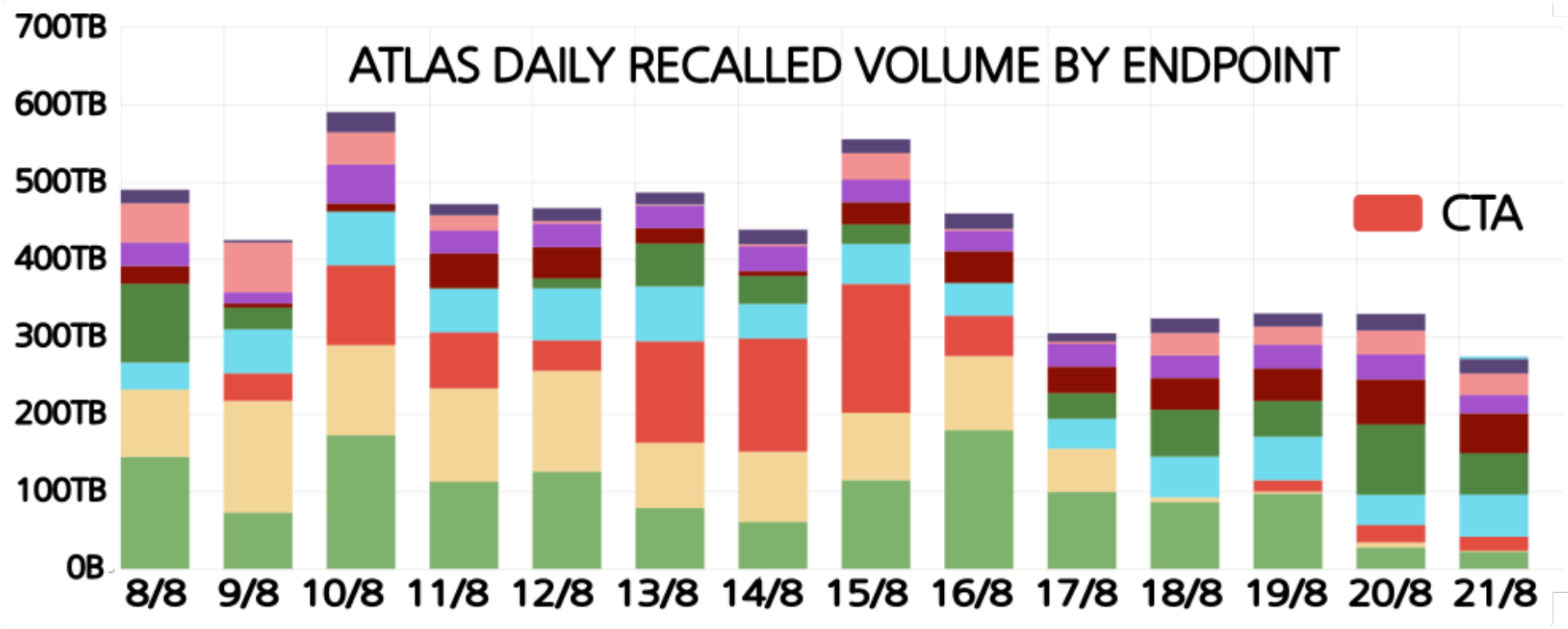
EOSCTA operations 2019

- Moved from spinners to SSDs
- Deployed “retrieve only” instance for ATLAS
  - 5 hyper-converged servers
  - 20 TB of SSDs
  - 10 tape drives
- Migrated ATLAS metadata from CASTOR to EOSCTA
  - See Giuseppe Lo Presti’s “CERN Storage Evolution” presentation
- Successful ATLAS 2018 recall campaign
  - All ATLAS data in CASTOR was available via CTA
  - "From ATLAS side, the recall campaign went smoothly and we managed to recall all the files (326k files, 741 TB) in a timely manner (CERN was the first site to achieve the full recall of all the files).This gave us additional confidence in CTA performance and in the migration strategy”
- Developed EOSCTA operations tools, dashboards and workflows
- ATLAS deployment of EOSCTA is possible from November

EOSCTA performance during ATLAS recall campaign



Relative EOSCTA performance during ATLAS recall campaign



EOSCTA operations 2020 plans

- Deploy LHC instances
  - Alice
  - ATLAS
  - CMS
  - LHCb
- Deploy non-LHC “public” instance after LHC instances
  - Users and VOs will be consulted accordingly
  - Need to understand public specific workflows
  - CASTOR will not be turned off without agreement with users and VOs

EOSCTA software developments 2019

- Repack
- Full integration with FTS and XRootD
- Support for multiple database backends
- “Activities” and priorities for retrieves
  - ATLAS requested support for “activities”
  - Integrated with FTS

EOSCTA database backends

- Oracle
  - The current production solution at CERN
- PostgreSQL
  - The future production solution at CERN
  - Recommended for Tier 1 sites
- MySQL
  - Contributed by IHEP, China

EOSCTA service improvements 2020

- CERN LTO RAO
  - Travelling salesman optimization for read positioning
  - Mass recalls are ~2.5x-3x faster than linear
  - Please see
    - LTO performance, Make Tape Reading Great Again, HEPiX Fall 2018, Barcelona
    - https://indico.cern.ch/event/730908/contributions/3153156/attachments/1732268/2800425/LTO-CERN-HEPiX-Oct-2018-germancancio.pdf
- Pre-emption of tape drives
  - Use tape drives at full speed all of the time
- Tools and modifications to facilitate operations
  - Tape and drive dedications
- Migrate to XRootD 5 (extended attributes)
- Collocation hints in collaboration with ATLAS