nextgenio newsletter

Welcome to the seventh issue of the NEXTGenIO newsletter!

NEXTGenIO at ISC 2019 Partner booths

arm	BSC	epcc	FUJITSU	intel	TECHNISCHE UNIVERSITÄT DRESDEN
F-910	A-1412	B-1251	J-600	F-930	J-622

The best place to come and talk to us at ISC is at any of our partners' booths, but you will also be able to find us at different events through the conference.

Research Poster:
Kronos Development and Results - HPC Benchmarking with Realistic Workloads (RP26), Antonino Bonanni, Simon D. Smart, Tiago Quintino (ECMWF), Tuesday 18th June, 8:30am-10:00am, Substanz 1-2.
More information at: http://bit.ly/ISC19Poster

Booth Presentation:
Intel Datacenter Persistent Memory Modules for Efficient HPC Workflows, Adrian Jackson (EPCC, University of Edinburgh), Tuesday 18th June, 1:40pm - 2:00pm, Intel booth (F-930).

Booth Demo:
PyCOMPSs 2.5, Tuesday 18th June, 2:00pm, BSC booth (A-1412).

ISC High Performance
FUELING INNOVATION

ISC 2019 Birds-of-a-Feather Session

Multi-Level Memory and Storage for HPC and Data Analytics & AI

Tuesday 18th June, 1:45pm – 2:45pm, Kontrast

Join NEXTGenIO members Hans-Christian Hoppe (Intel Datacenter Group) and Michèle Weiland (EPCC, University of Edinburgh), along with Kathryn Mohror of Lawrence Livermore National Laboratory, for a BoF discussing use cases and requirements for next-generation multi-level storage/memory systems, present proof of concept prototype results, and system software and tools development.

More information at:
http://bit.ly/ISC19BoF

ISC 2019 Workshop

HPC I/O in the Data Center

Thursday 20th June, 9:00am – 6:00pm, Basalt

HPC I/O in the Data Center Workshop

Adrian Jackson (EPCC, University of Edinburgh) will present results from NEXTGenIO in his talk, "An Architecture for High Performance Computing and Data Systems using Byte-Addressable Persistent Memory".

More information at:
http://bit.ly/ISC19Workshop

NEXTGenIO Prototype Delivered

Ingolf Staerk, Fujitsu

The NEXTGenIO project is now into the final straight. After a successful project review in December, production and testing of the NEXTGenIO prototype was completed at Fujitsu. The integrated and validated NEXTGenIO Prototype Hardware was delivered to EPCC in March, to be made available to all project partners for final software porting, tests and performance measurements.

The NEXTGenIO prototype system is presented to the EC project officer and reviewers at Fujitsu

Based on the NEXTGenIO hardware architecture that was created earlier in the project, Fujitsu has developed the NEXTGenIO system motherboard and system-ware, using Intel's Optane™ Data Centre Persistent Memory Modules (DCPMM). A 4-node demonstration cluster based on the new motherboard was already available at Fujitsu for project partners to port and test their middleware and application software.

Live demonstrations of NEXTGenIO cluster hardware and software were performed during the project review and the first impressive I/O performance achievements demonstrated successful milestones in the development of this new, scalable, high-performance computing platform, which has been designed to address the challenge of delivering scalable I/O performance to applications at the Exascale.

The review team and project partners had the opportunity to inspect the final production of the NEXTGenIO prototype system at Fujitsu during the review. Each of the prototype's 34 compute nodes is equipped with two Intel Xeon "Cascade Lake" CPUs and 3TB of DCPMM, and includes a software stack to seamlessly support I/O and memory intensive workloads. Compute and management nodes as well as network components are integrated within two system racks.

In early March, the NEXTGenIO prototype system was delivered to the EPCC Advanced Computing Facility (ACF) just outside Edinburgh. Over the following two weeks, a joint team of Fujitsu and EPCC ACF staff finished the cluster hardware and software installation and configuration. Over the course of several training sessions, the Fujitsu team passed on their expertise in cluster management software, operating tools and hardware maintenance to the EPCC ACF team. Finally, with a comprehensive set of tests for individual cluster nodes, interconnects and network switches, full cluster functionality and HPL performance, the handover of the NEXTGenIO prototype to EPCC was completed successfully.

The system is now in the final phase of middleware and application porting, testing and integration before entering into production during the summer.

The NEXTGenIO prototype system, newly installed at EPCC's Advanced Computing Facility

NEXTGenIO Prototype

EC project officer, reviewers and NEXTGenIO project partners at the project review at Fujitsu

EC project officer, reviewers and project partners in front of the NEXTGenIO prototype system at Fujitsu

Full view of the NEXTGenIO prototype system

Extending Slurm to Support Data-Driven Workflows

Alberto Miranda, Ramon Nou (BSC), Adrian Jackson, Iakovos Panourgias (EPCC)

As HPC systems in the Top500 reach hundreds of petaFLOPs, researchers are turning their attention to problems requiring large-scale analysis of experimental and observational data, such as the computational analysis of ITER reactor designs or the simulation, filtering, and evaluation of large-scale experiments such as the Compact Muon Solenoid at the Large Hadron Collider.

What separates these data-intensive problems from traditional, compute-bound large-scale simulations is that, even though they exhibit comparable computational needs, they also have significant data requirements. Further, in many cases these problems are run as a composition of separate tasks that are executed as a scientific workflow in the context of a large parallel job, with each task representing a separate phase in a complex model that may depend on data generated by previous phases. This requires communicating large amounts of data between tasks by relying on the parallel file system, which results in severe I/O performance degradation.

Unfortunately, the currently available interfaces between users and HPC resource managers do not make it possible to convey these data dependencies between jobs. For example, if a Job A generates data that should be fed into a Job B, there is no way for users to express this dependency, nor to influence the job scheduling process so that Job A's output is kept in burst buffers or node-local storage until Job B starts. Worse still, since the I/O stack remains essentially a black box for today's job schedulers, Job A's output could end up being synchronized to the cluster's parallel file system and, at some point in the near future, staged back into the new node allocation for Job B, which might end up including some of the original nodes reserved for Job A.

To tackle this issue, researchers at BSC and EPCC have been working on extending Slurm and developing new services to capture these I/O requirements and dependencies from data-driven workflows, so that they can be used to improve the overall performance of the I/O subsystem. Thus, the NEXTGenIO Slurm job scheduler allows users to define an HPC workflow in terms of which jobs belong to it and how they depend on each other. Each workflow receives a unique Workflow ID and all jobs that are part of a workflow are handled by the scheduling algorithm. Slurm guarantees that each job gets updated priorities and resource allocations as the workflow progresses and also monitors their execution so that, if an intermediate job fails, all remaining jobs in the workflow can be cancelled. Furthermore, the Slurm interface tools and programmatic APIs have been extended to allow users to query Slurm about the overall status of a workflow and also to get a list of all jobs and their current status.

Additionally, to efficiently transfer data between the different tasks in a workflow and appropriately exploit the high performance and density of Intel®'s Optane™ DCPMM node-local storage, a new service called NORNS has been developed. NORNS is an infrastructure service that coordinates with Slurm to orchestrate asynchronous data transfers between the different storage layers in an HPC cluster. NORNS enables Slurm to support scientific workflows by providing a framework for executing and monitoring data transfers between node-local, node-shared and remote storage. By coordinating with NORNS, Slurm can schedule transfers between jobs without having to wait for it to actually be available at the intended destination. This frees it to continue scheduling oncoming job submissions, while NORNS focuses on monitoring transfers and prioritizing them as needed.

NEXTGenIO will thus contribute a framework that will allow users to easily define and execute their scientific workflows while transparently benefiting from upcoming storage technologies such as Intel®'s Optane™ DCPMM.

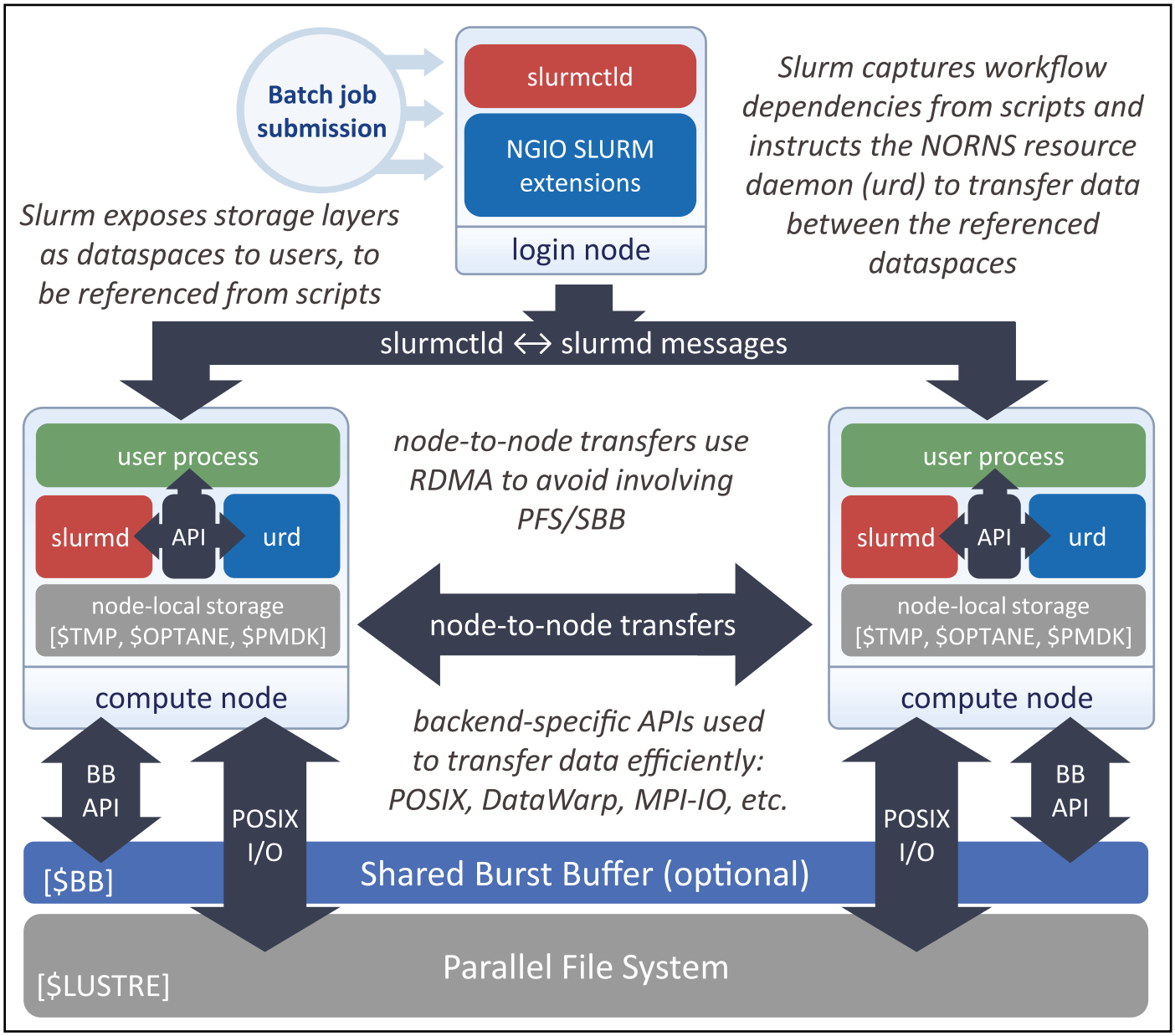

Figure 1 Supporting data-driven workflows in the NEXTGenIO project

Top-Down Performance Analysis of HPC Workflows

Christian Herold, TU Dresden

Modern HPC workflows consist of coordinated sequences of interdependent applications. Typically, workflows are implemented as batch jobs composed of multiple dependent steps with each step executing a single application. Due to dependencies, inefficiencies in one step may delay work depending on its associated job and increase the runtime of the whole workflow. Identifying a bottleneck in a complex workflow can be a challenging task. NEXTGenIO develops tools that assist in analysing such complex workflows.

In order to identify and optimise the job step responsible for a potential bottleneck, details of the runtime behaviour are collected. A top-down approach is deployed to study the performance data at both a global (whole workflow) and a detailed (application level) perspective.

The top-down approach provides performance summaries for each level of a typical workflow, from the entire workflow to individual job steps. Figure 1 depicts an example workflow that consists of two jobs where each job contains two steps. On the first level, our approach provides performance summaries for the overall workflow. We define the Runtime Share as the distribution of CPU time among computation, communication, and I/O categories. Using these summaries, the user can identify the location and nature of potential performance problems in the workflow.

The second level of our approach provides a similar breakdown for a specific job and gives an overview of its steps. Summaries inform users about the job's Runtime Share, allowing the most time-consuming steps or paradigms to be identified.

The third and most detailed level provides an outline for a specific job step. From here, users can switch to an in-depth single application analysis as provided by the tools MAP and Vampir.

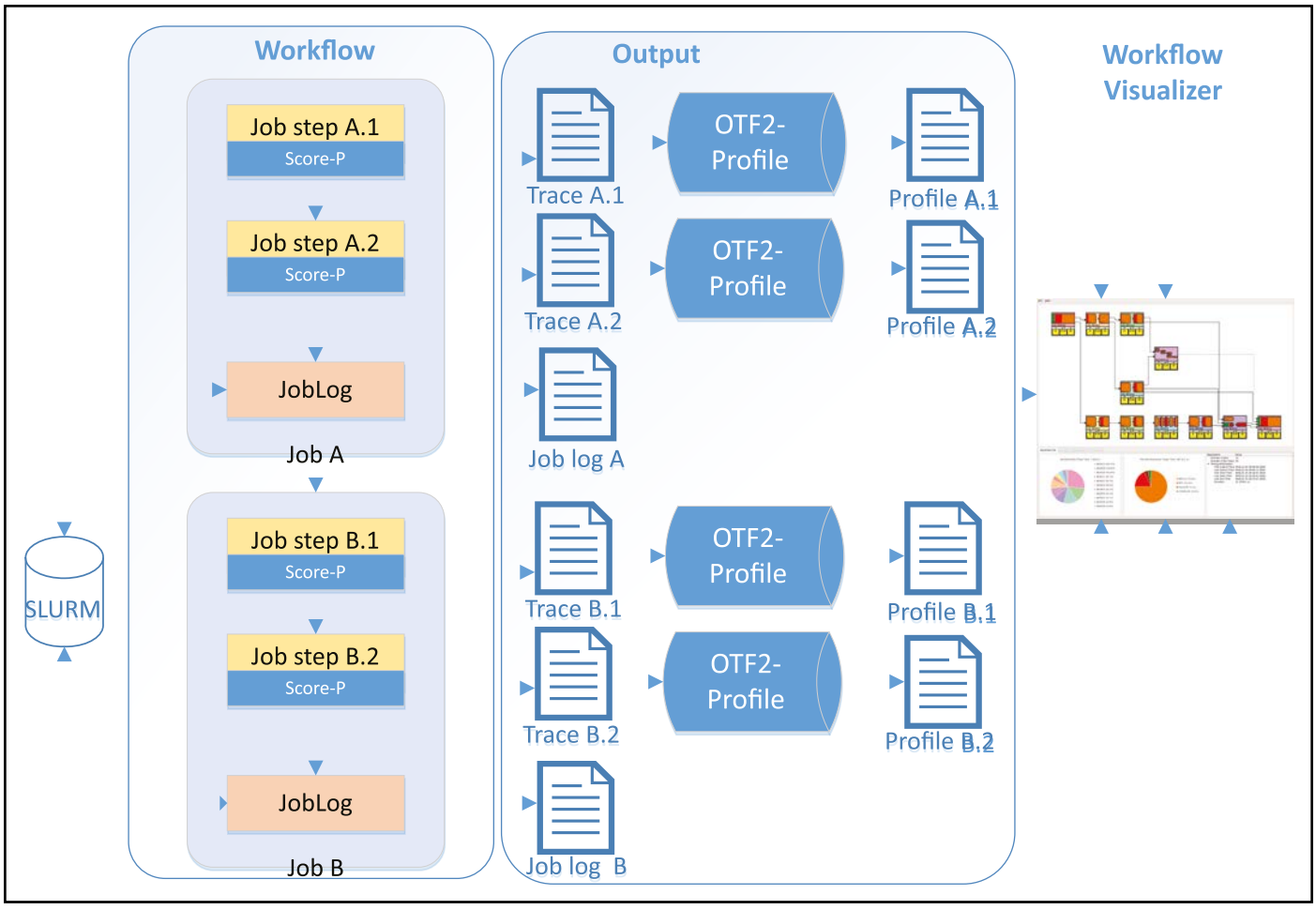

Figure 1 Workflow data collection and visualization in NEXTGenIO

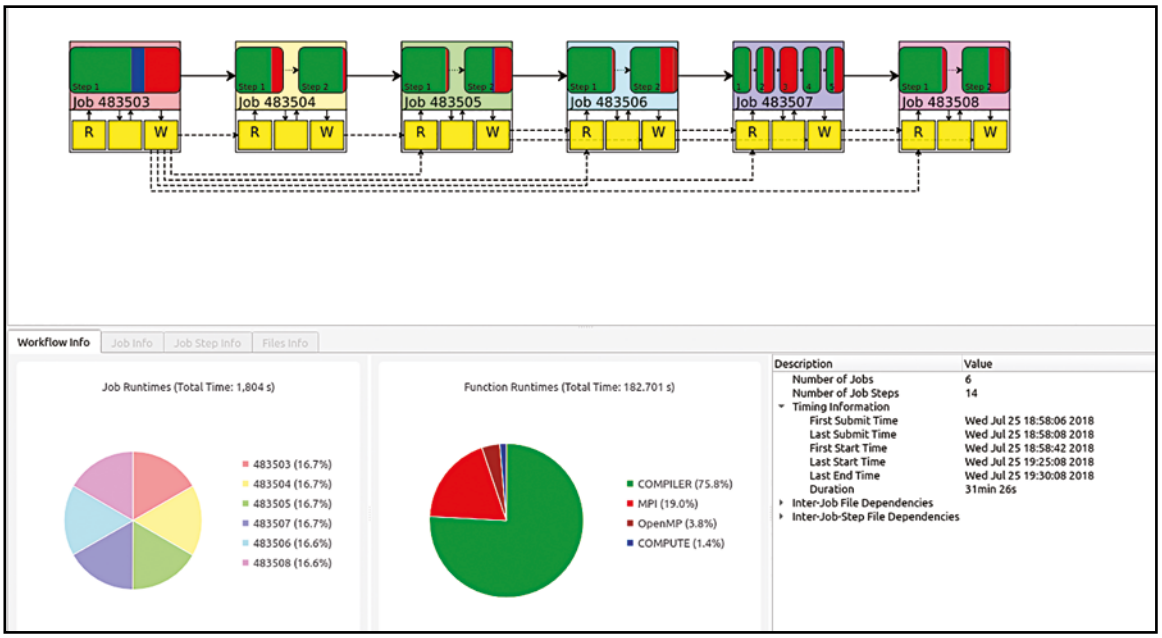

Figure 2 Main window of the Workflow Visualizer that displays the Workflow Graph, the Job Summary, function runtimes and the Info Table containing job details from the scheduler.

Figure 2 illustrates the design of our approach using the example from Figure 1. We use the Score-P measurement infrastructure to collect performance metrics of the workflow and store individual trace logs (OTF2 formatted) per job step.

We use OTF2-profile to extract summaries from each log for the top-level representations. To determine the actual execution order of the jobs along with their steps, we use the tool JobLog to query and store job timings and dependency information. JobLog is called once per job execution after all steps are completed.

Finally, the workflow visualizer combines all profiles and job log files in a hierarchical manner. It offers summary charts for the first (workflow), second (job) and third (job step) hierarchy levels. The workflow visualizer provides details of job structure and dependencies between jobs (bold arrows), I/O dependencies between jobs (dashed lines), and job steps. A summary of the job runtimes and their distribution (bottom left) allows costly jobs to be identified. It also provides a summary of the runtime share (bottom centre), and details from clicking on the job/step box (bottom right).

In the near future, the workflow visualizer will include I/O (file) dependencies between job steps. Furthermore, resource utilization metrics per job step will be added.

Score-P
Scalable performance measurement infrastructure for parallel codes

VAMPIR

PARTNER PROFILE

Vampir implements optimized event analysis algorithms and customizable displays which enable fast and interactive rendering of very complex performance monitoring data.

The product has been developed at the Center for Applied Mathematics of Research Center Jülich and the Center for High Performance Computing of the Technische Universität Dresden. The development is continued by Center for Information Services and High Performance Computing (ZIH) of Technische Universität Dresden.

Vampir has been available as a commercial product since 1996, and has been enhanced over time within the scope of many research and development projects.

Workshop: NVRAM storage for exascale I/O

NEXTGenIO Workshop on applications of NVRAM storage to exascale I/O

ECMWF, Reading (UK), 25-27 September 2019

The NEXTGenIO project partners have designed, built and delivered a prototype hardware platform based around new non-volatile memory (NVRAM) technology. In addition to the hardware, the project has also developed a full software stack which is deployed on the prototype and which explores its novel capabilities. This workshop will discuss the usage scenarios for byte-addressable persistent memory and the impact it will have on high-performance computing and data intensive applications.

With the Exascale gradually becoming a reality for the supercomputing community, memory and I/O bottlenecks remain two of the key challenges that need to be solved in order for applications to be able to fully exploit the capabilities of these systems. The NEXTGenIO project partners have been collaborating closely over the past four years to design, build and deliver a prototype hardware platform based around new non-volatile memory (NVRAM) technology. In addition to the hardware, the project has also developed a full software stack which is deployed on the prototype and which explores its novel capabilities.

This workshop will discuss the usage scenarios for byte-addressable persistent memory and the impact it will have on high-performance computing and data intensive applications. Speakers at the workshop will represent the user community, hardware vendors and software tools providers.

Contributions to the workshop presentations and posters are invited. These should focus on the Exascale I/O and Memory challenges and experience with Storage Class Memory and byte-addressable Non-Volatile storage, from the level of single applications to full integrated workflows.

The workshop starts at 14:00 on Wednesday 25 September and continues on Thursday 26 September. On Friday 27 September there will be a NEXTGenIO hackathon (maximum 20 places available).

Deadline for registration and submission of abstracts: 2 August 2019.
More information: http://bit.ly/NVRAM-Workshop

Recent presentations

NEXTGenIO work has recently been presented at workshops in the UK and the USA.

ODB and the Development of a Domain Specific Distributed Object-Store, *Simon Smart, Tiago Quintino, Baudouin Raoult (ECMWF)* - Workshop on the Interface for Observation Data Access (IODA), 12 February 2019, Monterey, California.

Development of High-Performance distributed object-store for Exascale numerical weather prediction and climate model data, *Tiago Quintino (ECMWF)* - presentation and live demo at Workshop on Storage Challenges in the UK, 6 March 2019, Reading University, UK.

Modelling HPC Workloads with (some) Machine Learning, *Antonino Bonanni, Simon Smart and Tiago Quintino (ECMWF)* - poster presented at 1st Workshop on Leveraging AI in the Exploitation of Satellite Earth Observations and Numerical Weather Prediction, 23-25 April 2019, NOAA Center for Weather and Climate Prediction, College Park, MD, USA.

Where available, publications and slides from presentations can be downloaded from our website.