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

Report of Contributions

https://events.ecmwf.int/e/143

Contribution ID: **1** Type: **not specified**

Welcome and opening

Wednesday, September 25, 2019 1:30 PM (15 minutes)

Contribution ID: **18** Type: **Oral presentation**

HPC workflows, NVRAM and parallel file systems –Is there a way forward ??

Wednesday, September 25, 2019 3:45 PM (30 minutes)

Improving data intensive workflows in modern supercomputers by any means possible continues to be a focus of both industry and academic research. And while big steps have been made, most have served to move the I/O bottleneck somewhere else and not actually solve it.
Putting NVRAM into computers, semi-shared burst buffers or even into storage servers to improve overall performance and reduce latency is great but can we afford to sacrifice resiliency and flexibility to achieve it. And what about the continued need for parallel file systems (or next gen object storage solutions) in this picture ??
Storage tiering may be the way to go but data movement between tiers is expensive.

This talk is intended to see the future of data intensive workflows from a storage point of view and how an end to end architecture could benefit from NVRAM and NVMe based technologies while still retaining a semblance of traditional storage subsystems.

Keywords

Parallel File systems, workflow acceleration

Author: Dr KLING PETERSEN, Torben (Cray Inc)

Presenter: Dr KLING PETERSEN, Torben (Cray Inc)

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

Contribution ID: **19** Type: **Oral presentation**

Improving the I/O scalability for the next generation of Earth system models: OpenIFS-Cassandra integration as a case study

Thursday, September 26, 2019 2:30 PM (30 minutes)

Earth system models (ESMs) have increased the spatial resolution to achieve more accurate solutions, producing an enormous amount of data. However, some ESMs use inefficient sequential I/O schemes that do not scale well when many parallel resources are used. This is the case of the OpenIFS model.

OpenIFS is a free and simplified version of the Integrated Forecasting System (IFS), available under a license from the European Centre for Medium-Range Weather Forecasts (ECMWF). IFS is a global data assimilation and forecasting system which includes the modelling of the atmospheric composition.

The output in OpenIFS is done through an inefficient serial I/O scheme where the master process gathers all the data and sequentially writes into GRIB files.

When OpenIFS is used for climate modelling, there is a series of necessary post-processing operations to be performed, such as file format conversion to netCDF or regridding. In addition, if users want to query only a subset of the simulated data, they have to post-process entire files, and then filter the data. It is a slow, inefficient process which is difficult to parallelize.

To mitigate these issues, we modified OpenIFS to perform distributed I/O on all processes. Then, adopted Apache Cassandra, a Key-Value distributed database, to store all the information. Also, we tried Intel's modified Cassandra based on persistent memory whose higher throughput and lower latencies promise faster simulations and analysis. As a result, I/O intensive simulations are more efficient while data is managed transparently.

By using a Key-Value distributed database, scientists can query forecasts without post-processing entire files. Information is indexed at run-time, enabling users to access particular subsets without reading large amounts of data. Besides, our proposed system supports sampling to speedup anomaly and pattern detections, which simplifies data exploration greatly.

In this work, it became apparent that Cassandra and its connectors are not designed for HPC environments. They are based on thread-pools, which ultimately compete with MPI for CPU-time, reducing the potential gains of distributed I/O. In future research, we would like to explore a better integration with HPC software and persistent-memory to overcome these limitations.

Keywords

OpenIFS, Cassandra, persistent-memory

Authors: YEPES ARBÓS, Xavier (Barcelona Supercomputing Center); SANTAMARIA, Pol (Barcelona Supercomputing Center); SERRADELL MARONDA, Kim (Barcelona Supercomputing Center); Dr BECERRA, Yolanda (Barcelona Supercomputing Center); Prof. LABARTA, Jesus (Barcelona Supercomputing Center); Dr SERVAT, Harald (Intel Corporation)

Presenter: YEPES ARBÓS, Xavier (Barcelona Supercomputing Center)

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

Contribution ID: **20** Type: **Oral presentation**

Which Memory Abstraction for NVDIMM on Object Storage ?

Wednesday, September 25, 2019 4:15 PM (30 minutes)

The SAGE storage system developed by the SAGE consortium provides a unique paradigm to store, access and process data in the realm of extreme-scale data-intensive computing. The storage system can incorporate multiple types of storage device technologies in a multi-tier I/O hierarchy, including flash, disk, and more importantly NVDIMMs.

The core software infrastructure driving the storage system is not a parallel file system, but an object storage software platform called "Mero", built from the ground up to cater to extreme-scale HPC. The object storage platform provides a powerful, flexible and open API called Clovis which provides an I/O interface and instrumentation interfaces to get fine grain telemetry data and operation logs from the storage system. These interfaces help to flexibly extend the storage system capabilities with new data management features and provide methods to access data in multiple ways.

In this poster, we present the Global Memory Abstraction Design work being done during the first year of the project, where the main goal is to provide memory like addressing semantics of persistent storage such as NVDIMM, i.e. how to expose the NVRAM technology to applications using Mero. The other goal is also to study the optimal use of NVDIMM in the I/O stack including those in the compute nodes themselves and understand what can bring the interesting properties of NVRAM, i.e. persistence, byte-addressability or its important density, for HPC applications.

Keywords

NVDIMM, Object Storage, Mero

Authors: Dr VAUMOURIN, Grégory (ATOS); Dr VALAT, Sebastien (ATOS); Mr LAFERRIERE, Christophe (ATOS); Mr COUVÉE, Philippe (ATOS); Dr NARASIMHAMURTHY, Sai (Seagate); Mr RIVAS GOMEZ, Sergio (KTH University); Prof. MARKIDIS, Stefano (KTH University); Dr NIKOLERIS, Nikos (ARM); Mr HUANG, Hua (Seagate)

Presenter: Dr VAUMOURIN, Grégory (ATOS)

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

Contribution ID: **21** Type: **Oral presentation**

I/O Challenges for U.S. Navy METOC Modeling – From Data to Decisions

Thursday, September 26, 2019 2:00 PM (30 minutes)

Increased complexity and resolution of modeling systems and increased demand for environmental input to downstream products combine to pose significant challenges for increased volume and velocity of I/O as part of a research or operational workflows, particularly when seeking to exploit distributed computing architectures. While many of the challenges faced are common in the earth system modeling community, I plan to highlight tradeoffs between data locality and task granularity in workflows (and the implications of on-demand storage), particularly when faced with workflows of tasks with highly dissimilar computational resource requirements. This will include a discussion of asynchronous I/O in forecast models, and the possibility of data-driven downstream workflows that are able to fully exploit the use of high-speed persistent storage as a method to increase task parallelism.

Keywords

challenges

Author: WHITCOMB, Timothy (US Naval Research Laboratory)

Presenter: WHITCOMB, Timothy (US Naval Research Laboratory)

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

Contribution ID: **22** Type: **Oral presentation**

Utilizing Heterogeneous Storage Infrastructures via the Earth-System Data Middleware

Thursday, September 26, 2019 3:00 PM (30 minutes)

The Earth-System Data Middleware (ESDM) is a performance-aware middleware that builds upon a data model similar to NetCDF and utilises a self-describing on-disk data format for storing structured data. ESDM allows to employ multiple (shared and local) storage systems concurrently and explicitly supports heterogeneous storage infrastructures. From the user/application perspective, the complexity of multiple-storage tiers is hidden, and data fragmentation controlled by a configuration file.

Performance measurements running on Mistral (DKRZ) show that multiple Lustre file systems can be used efficiently with the current ESDM version. Several configurations were tested generating performance of 200 GB/s. The results were compared to a best-case benchmark using IOR that achieves 150 GB/s. Using tmpfs locally, ESDM was able to achieve 2 TB/s on 500 nodes showing that non-volatile memory could also be well utilized. ESDM also provides a backend to utilise the Kove XPD in-memory shared storage.

While the tooling is not yet completed, ESDM already integrates with NetCDF and can be used as a drop-in replacement for typical use-cases without changing anything from the application perspective. While our current version utilises the manual configuration by data-centre experts, the ultimate long-term goal is to employ machine learning to automatise the decision-making and reduce the burden for users and experts.

Keywords

NetCDF, ESDM, ESiWACE

Authors: Dr KUNKEL, Julian (University of Reading); Dr PEDRO, Luciana (University of Reading); Prof. LAWRENCE, Bryan (NCAS)

Presenter: Dr KUNKEL, Julian (University of Reading)

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

Contribution ID: **23** Type: **Oral presentation**

Accelerating CFD simulations with DCPMM

Thursday, September 26, 2019 10:00 AM (30 minutes)

In this presentation, we will show how CFD simulations can be accelerated using DCPMM. We will show performance results from the NEXTGenIO cluster and explain the different methods employed to improve the end-to-end performance of a real-life CFD simulation.

Keywords

application, performance

Author: WEILAND, Michele (EPCC, The University of Edinburgh)

Presenter: WEILAND, Michele (EPCC, The University of Edinburgh)

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

Contribution ID: **24** Type: **Oral presentation**

NEXTGenIO Prototype Architecture

Wednesday, September 25, 2019 2:45 PM (30 minutes)

Gives an overview about the architecture of the NEXTGenIO prototype.
Outlines the different platform modes such as 1LM and 2LM and their use cases.
Provides an abstract overview of the basic NEXTGenIO software stack.
Shows the fundametal differences between the use of non-volatile main memory as storage and classic storage as we use it today.

Keywords

Architecture Prototype, non volatile memory modes, storage

Author: HOMOELLE, Bernhard (Fujitsu SVA)

Presenter: HOMOELLE, Bernhard (Fujitsu SVA)

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

Contribution ID: **25** Type: **Oral presentation**

Profiling and Debugging HPC Applications on Emerging Memory Technologies

Thursday, September 26, 2019 12:00 PM (30 minutes)

The emergence of non-volatile memory technologies in high performance servers has resulted in multiple initiatives to exploit these new technologies for scientific computing. Traditionally challenges for adopting new technologies in HPC has focused on CPU technologies –however the paradigm shift of new ways of exploiting memory technologies presents new challenges.

As part of the NextGenIO project Arm, formerly Allinea, have been looking at how to support application developers adopt non-volatile memory DIMMs (NVDIMMs) for HPC and scientific computing applications. By adapting the existing Arm Forge product line, we have introduced new capabilities for debugging and profiling applications which make use of NVDIMM technologies.

In this presentation we will present of existing methodologies for debugging and profiling applications using existing memory technologies and detail the development of the new capabilities and how they can be utilized for real applications.

Specifically, we will focus on debugging NVDIMMs when accessed through the Persistent Memory Development Kit (using the pmemobj library). For profiling we will present on a newly developed custom metrics plugin for MAP which captures additional detail of NVDIMM usage. In both cases we will illustrate the features with real world code examples.

Keywords

Debugging Profiling Arm Tools

Authors: PERKS, Oliver (Arm); MOONEY, Kevin

Presenters: PERKS, Oliver (Arm); MOONEY, Kevin

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

Contribution ID: **26** Type: **Oral presentation**

Running ECMWF workflow on the NextGenIO prototype

Thursday, September 26, 2019 11:00 AM (30 minutes)

The architecture of the NextGenIO prototype featuring Intel's Optane DCPMMs presents both challenges and opportunities as a platform for ECMWF's time-critical operational workflow.

This workflow generates massive amounts of I/O in short bursts, accumulating to tens of TiB in hourly windows. From this output, millions of user-defined daily products are generated and disseminated to member states and commercial clients all over the world. These products are processed from the raw output of the IFS model, within the time critical path and under a strict delivery schedule. Upcoming improvements in resolution and growing popularity will increase both the size and number of these products. Based on expected model upgrades, by 2025 we estimate the operational model will output over 335 TiB per forecast.

We will present our last 4 years work, refactoring ECMWF's software stack to deliver an operational system that handles these workflows, and demonstrate that these new storage class memories will help significantly in tackling our I/O issues as we approach Exascale.

Keywords

Authors: QUINTINO, Tiago (ECMWF); SMART, Simon (ECMWF); IFFRIG, Olivier (ECMWF); BONANNI, Antonino (ECMWF); RAOULT, Baudouin (ECMWF)

Presenters: QUINTINO, Tiago (ECMWF); SMART, Simon (ECMWF)

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

Contribution ID: **27** Type: **Oral presentation**

"Accelerating Time to insight through advanced memory centric system architectures"

Wednesday, September 25, 2019 1:45 PM (1 hour)

Advancements in memory technologies open the door to the development of novel domain specific system architectures delivering not only improved performance but reduced energy consumption as well.

Keywords

Author: Mr FLEISCHER, Balint (Micron Technologies)

Presenter: Mr FLEISCHER, Balint (Micron Technologies)

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

Contribution ID: **28** Type: **Oral presentation**

Supporting NVRAM storage with active systemware

Thursday, September 26, 2019 11:30 AM (30 minutes)

NVRAM technology used for storage and memory within compute nodes of a HPC system has the potential to bring large performance benefits and enable new functionality, especially for computing scientific workflows and processing scientific data. However, safely and efficiently utilising such technology on a large scale, multi-user, HPC system presents challenges using existing applications and systemware. Within the NEXTGenIO project we have developed a number of systemware component, including a multi-node filesystem that utilises in-node NVRAM, data schedulers to move data on and off compute nodes, and batch system integration to ensure user workflows and jobs can safely exploit compute node NVRAM. This presentation will outline these components and discuss some of the performance and functionality benefits such components can bring to HPC systems and users.

Keywords

NVRAM, Systemware, Scheduler

Authors: JACKSON, Adrian (EPCC, The University of Edinburgh); Mr PANOURGIAS , Iakovos (EPCC); Dr NOU, Ramon (BSC); Mr MIRANDA , Alberto

Presenter: JACKSON, Adrian (EPCC, The University of Edinburgh)

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

Contribution ID: **29** Type: **Oral presentation**

Performance results of MONC using NVRAM

The I/O bottleneck is still a challenge to overcome when an HPC system is built. The NEXTGenIO project has investigated this issue over the past four years. A prototype hardware platform based around new non-volatile memory (NVRAM) technology has been designed and built. NVRAM is used to bridging the latency gap between memory and disk. In addition to the hardware, a full software stack has also been developed and deployed on the prototype.

One of the applications used to evaluate the platform's effectiveness regarding I/O performance and throughput is MONC, the Met Office/NERC Cloud simulator. This application is a large-eddy simulation code for cloud and atmospheric modelling. This application generates large diagnostic files at regular intervals as well as a checkpoint file at the end of the execution, allowing to continue the simulation from that point.

This poster will present and discuss the preliminary results of evaluating MONC on the NEXTGenIO platform. Several usage scenarios will be discussed.

Keywords

MONC, NVRAM, IO

Authors: HERRERA, Juan (EPCC); WEILAND, Michele (EPCC, The University of Edinburgh)

Presenter: WEILAND, Michele (EPCC, The University of Edinburgh)

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

Contribution ID: **30** Type: **Oral presentation**

A top-down performance analysis methodology for workflows

Wednesday, September 25, 2019 4:45 PM (30 minutes)

Scientific workflows are well established in parallel computing. A workflow represents a conceptual description of work items and their dependencies. Researchers can use workflows to abstract away implementation details or resources to focus on the high-level behaviour of their work items. Due to these abstractions and the complexity of scientific workflows, finding performance bottlenecks along with their root causes can quickly become involved. This work presents a top-down methodology for performance analysis of workflows to support users in this challenging task.

Our work provides summarized performance metrics covering different workflow perspectives, from general overview to individual jobs and their job steps. These summaries allow to identify inefficiencies and determine the responsible job steps.

In addition, we record detailed performance data about job steps, enabling a fine-grained analysis of the associated execution to exactly pinpoint performance issues. The introduced methodology provides a powerful tool for comprehensive performance analysis of complex workflows.

Keywords

Presenter: WILLIAMS, Bill (GWT-TUD GmbH)

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

Contribution ID: **31** Type: **Oral presentation**

Disrupting High Performance Storage with Intel DC Persistent Memory & DAOS

Thursday, September 26, 2019 9:00 AM (1 hour)

With an exponential growth of data, distributed storage systems have become not only the heart, but also the bottleneck of datacenters. High-latency data access, poor scalability, impracticability to manage large datasets, and lack of query capabilities are just a few examples of common hurdles. With ultra-low latency and fine-grained access to persistent storage, Intel Optane DC Persistent Memory Modules (DCPMM) represents a real opportunity to transform the industry and overcome many of those limitations. But existing distributed storage software was not built for this new technology, and completely masks the value DCPMM could provide. One needs to rethink the software storage stack from the ground up, to throw off irrelevant optimizations designed for disk drives and to embrace fine-grained and low-latency storage access, in order to unlock the potential of these revolutionary technologies for distributed storage.This presentation will introduce the ground-breaking technology from Intel in relation to Storage Class Memory and NVMe SSD, and the libraries being developed to take advantage of that technology. We introduce the architecture of the Distributed Asynchronous Object Storage (DAOS), which is an open-source software-defined multi-tenant scale-out object store designed from the ground up to take advantage of DCPMM and NVMe SSDs.

Keywords

persistent memory, DAOS, software-defined storage

Author: CHAARAWI, Mohamad (Intel)

Presenter: CHAARAWI, Mohamad (Intel)

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

Contribution ID: **32** Type: **not specified**

Open discussion: The role of Storage Class Memories for HPC

Thursday, September 26, 2019 4:00 PM (45 minutes)