nextgenio newsletter

Welcome to the sixth issue of the NEXTGenIO newsletter!

NEXTGenIO at SC 2018 Partner booths

BSC	FUJITSU	intel	TECHNISCHE UNIVERSITÄT DRESDEN
2038	1226	3223	827

arm	epcc
2639	2800

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

Demos at the BSC booth (#2038):

- **dataClay**: Tuesday 13 Nov at 11:40 and Wednesday 14 Nov at 16:00
- **COMPSs**: Tuesday 13 Nov at 16:40, Wednesday 14 Nov at 14:00, Thursday 15 Nov at 11:20
- **Tiramisú**, one of the NEXTgenIO use cases, using COMPSs and dataClay – date and time to be confirmed, look out for details.

SC 2018 Birds-of-a-Feather Session

Multi-Level Memory and Storage for HPC and Data Analytics

Tuesday 13th November, 5:15pm – 6:45pm, C147/148/154

Join NEXTGenIO members Hans-Christian Hoppe (Intel Corporation) and Michèle Weiland (EPCC, University of Edinburgh), along with Kathryn Mohror of Lawrence Livermore National Laboratory, for a BoF discussing the evolving storage-class memories (SCM) technologies landscape in the context of specific use cases, emerging technologies, and actual success stories.

More information at: http://bit.ly/2Jvc6xj

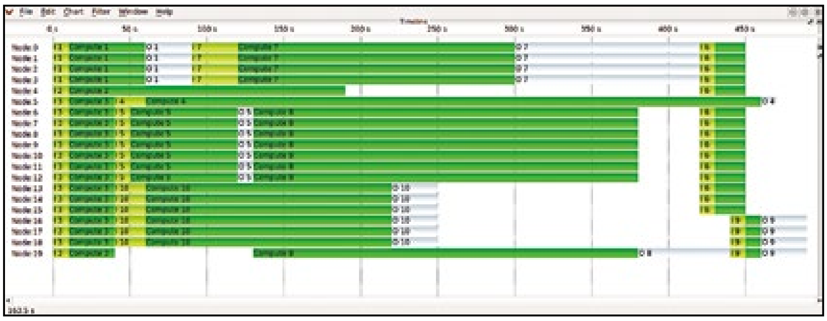
SC18
Dallas, TX | hpc inspires.

NEXTGenSim

NEXTGenSim, a workflow and storage aware scheduling simulator

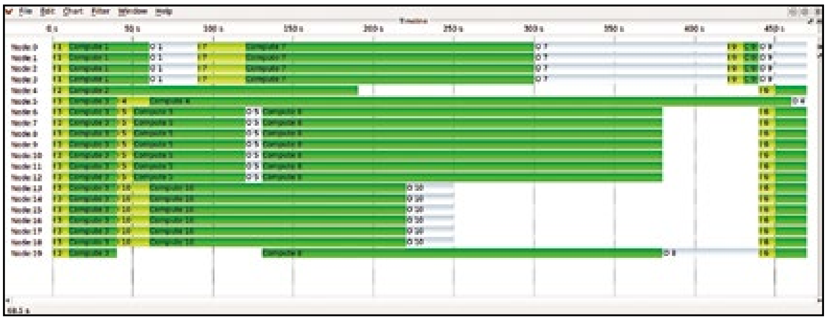
Nick Johnson, EPCC (University of Edinburgh)

As systems bearing on-node persistent storage become available, how to efficiently manage these resources in a supercomputer becomes a pressing question. To help answer this, researchers at EPCC, University of Edinburgh, are developing a work-flow and data aware scheduling simulator, NEXTGenSim. The aim of NEXTGenSim is to provide fast design-space exploration of scheduling algorithms and system parameters to sysadmins and developers based on real-world traces. This will reduce the time required to configure real systems for best performance.



Original work-load, each job having separate input, computation and output phases

NEXTGenSim is a discrete event simulator capable of simulating thousands of nodes and millions of jobs. It is able to ingest logs from PBS Pro and provides output in OTF2 format suitable for visualisation in VAMPIR. This eases the burden on developers and sysadmins by limiting the required tools and inputs to a set already available on our systems. Current development is in mimicking the operation of the scheduling algorithms in SLURM and validating against real clusters to give confidence in the predicted parameter choices for future systems.

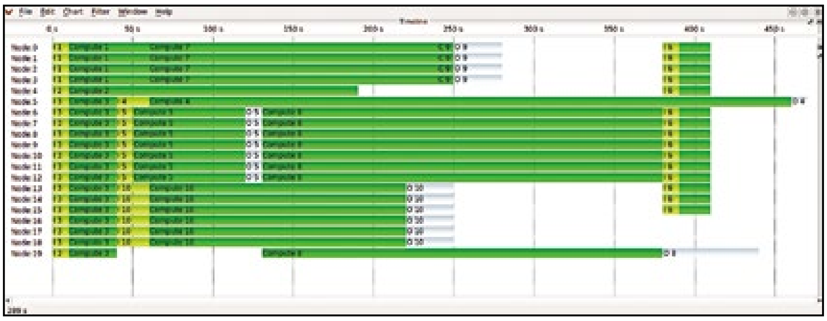


Intermediate work-flow with jobs scheduled to re-use nodes from preceeding jobs

Our figures show a simple test-case with a workflow. To begin, each step of the workflow has an input, computation and output phase. Each step can run anywhere as it has no dependence on the location of a predecessor job, so long as that job finishes before it starts. We then schedule to ensure, where possible, jobs in a workflow run on the same sets of nodes. Finally, we remove extraneous input and output phases where data can be held on-node in persistent storage between jobs rather than waiting for expensive (NFS) disk IO operations.

This simple case raises further questions: How long should data persist between jobs? When does it make more sense to move the data to make scheduling computation easier? Can a user “game” the system by calling a set of jobs a workflow and hogging nodes? Can jobs run between jobs in a workflow on the nodes primarily used by that workflow? How do job priorities work in workflow scenarios?

In the upcoming months, we hope to tackle these questions and provide advice to project colleagues developing SLURM on design and parameter choices before deployment.



Optimised work-flow, input and output phases eliminated by using persistent storage on-node between jobs in a work-flow.

NEW White Papers

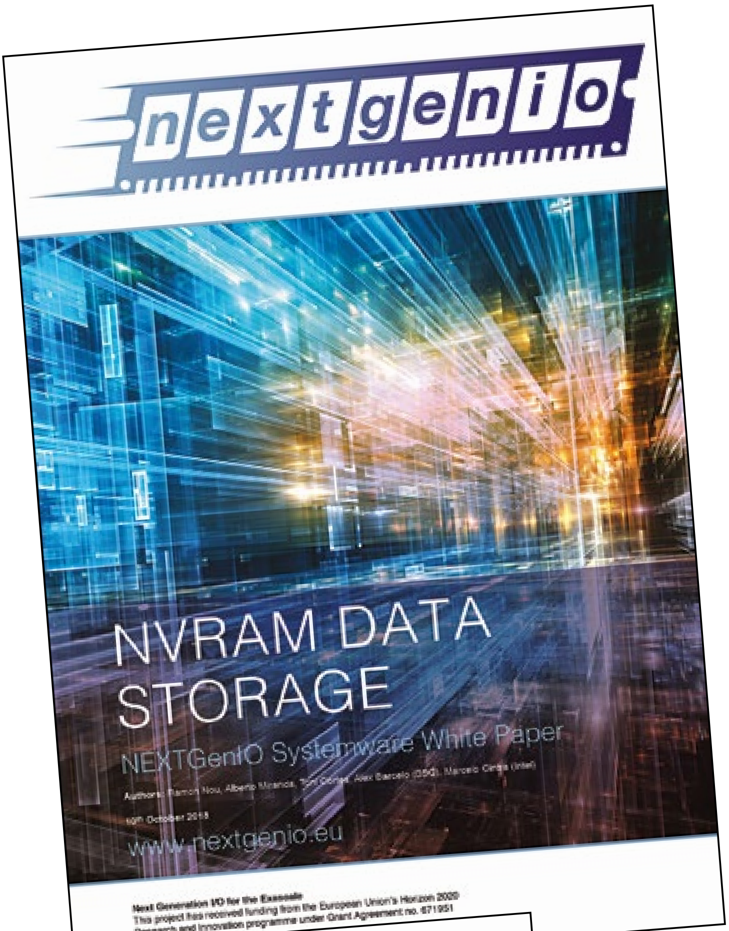
New White Papers from NEXTGenIO

We have recently published 3 new White Papers, which are now available to download from **http://www.nextgenio.eu**

NVRAM Data Storage describing the systemware components that allow applications to store data in the persistent memory layer.

Hardware Architecture describing the hardware architecture that was developed based on a detailed requirements capture exercise and that is implemented in the NEXTGenIO prototype.

Systemware Architecture and Usage Scenarios describing the systemware architecture that was developed based on a detailed requirements capture exercise informed by HPC and data analytics usage scenarios.



nextgenio

NVRAM DATA STORAGE

NEXTGenIO Systemware White Paper

Authors: Ramon Nou, Alberto Miranda, Toni Cortes, Alex Barcelo (BSC), Marcelo Cintra (Intel)

10th October 2018

www.nextgenio.eu

Next Generation I/O for the Exascale
This project has received funding from the European Union's Horizon 2020
Research and Innovation programme under Grant Agreement no. 671951

nextgenio

HARDWARE ARCHITECTURE

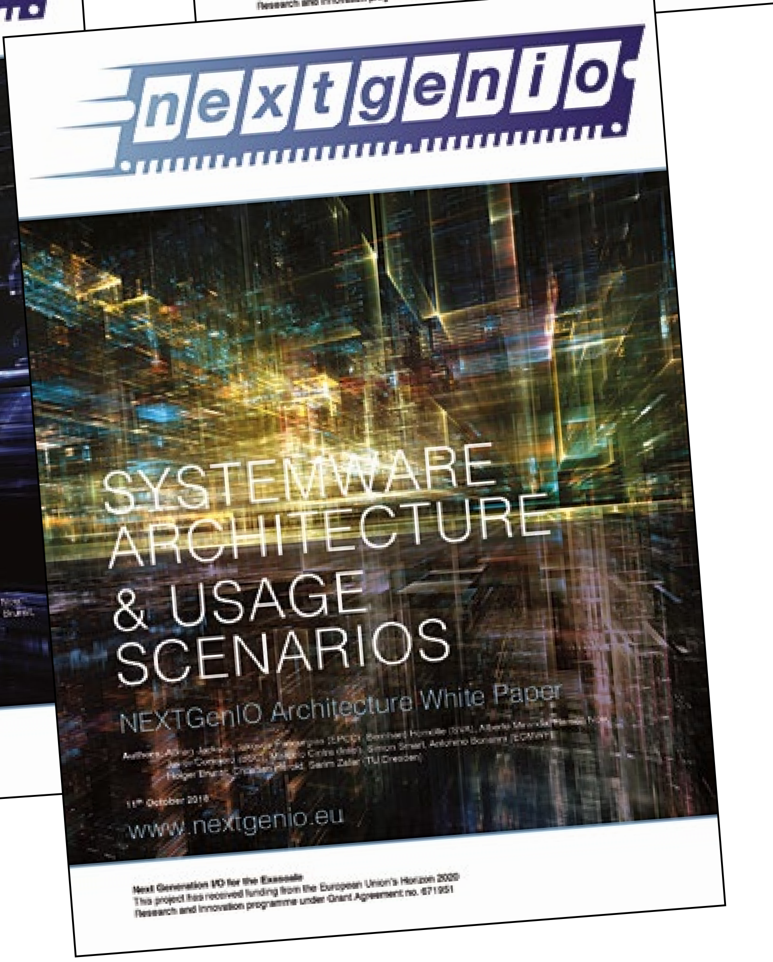
NEXTGenIO Architecture White Paper

Authors: Adrian Jackson, Iakovos Panourgias (EPCC), Bernhard Homölle (SVA), Alberto Miranda, Ramon Nou, Javier Conejero (BSC), Marcelo Cintra (Intel), Simon Smart, Antonino Bonanni (ECMWF), Holger Brunst, Christian Herold, Sarim Zafar (TU Dresden)

11th October 2018

www.nextgenio.eu

Next Generation I/O for the Exascale
This project has received funding from the European Union's Horizon 2020
Research and Innovation programme under Grant Agreement no. 671951



nextgenio

SYSTEMWARE ARCHITECTURE & USAGE SCENARIOS

NEXTGenIO Architecture White Paper

Authors: Adrian Jackson, Iakovos Panourgias (EPCC), Bernhard Homölle (SVA), Alberto Miranda, Ramon Nou, Javier Conejero (BSC), Marcelo Cintra (Intel), Simon Smart, Antonino Bonanni (ECMWF), Holger Brunst, Christian Herold, Sarim Zafar (TU Dresden)

11th October 2018

www.nextgenio.eu

Next Generation I/O for the Exascale
This project has received funding from the European Union's Horizon 2020
Research and Innovation programme under Grant Agreement no. 671951

Recent publications & presentations

NEXTGenIO work has recently been published in two papers, both as part of the proceedings of international conferences.

echofs: A Scheduler-guided Temporary Filesystem to leverage Node-local NVMs, *Alberto Miranda, Ramon Nou and Toni Cortes*, 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2018) Lyon, France, September 24-27, 2018

Abstract—The growth in data-intensive scientfic applications poses strong demands on the HPC storage subsystem, as data needs to be copied from compute nodes to I/O nodes and vice versa for jobs to run. The emerging trend of adding denser, NVM-based burst buffers to compute nodes, however, offers the possibility of using these resources to build temporary filesystems with specific I/O optimizations for a batch job. In this work, we present echofs, a temporary filesystem that coordinates with the job scheduler to preload a job's input files into node-local burst buffers. We present the results measured with NVM emulation, and different FS backends with DAX/FUSE on a local node, to show the benefits of our proposal and such coordination.

GekkoFS – A temporary distributed file system for HPC applications, *Marc- André Vef, Nafiseh Moti, Tim Süß, Tommaso Tocci, Ramon Nou, Alberto Miranda, Toni Cortes, and André Brinkmann*, IEEE Cluster, Belfast, UK, September 2018.

Abstract—We present GekkoFS, a temporary, highly-scalable burst buffer file system which has been specifically optimized for new access patterns of data- intensive High-Performance Computing (HPC) applications. The file system provides relaxed POSIX semantics, only offering features which are actually required by most (not all) applications. It is able to provide scalable I/O performance and reaches millions of metadata operations already for a small number of nodes, significantly outperforming the capabilities of general-purpose parallel file systems.

Work done in NEXTGenIO has also been presented at conferences in Europe and the USA.

ECMWF's Next-Generation IO and Product Generation for the IFS Model, *Tiago Quintino, Baoudoin Raoult, Pedro Maciel, and Simon Smart*, AMS 2018, 98th Meeting of the American Meteorological Society Fourth Symposium on High Performance Computing for Weather, Water, and Climate, Austin, Texas, USA, 9 January 2018.

ECMWF's Extreme Data Challenges on the HPC and Cloud systems, *Tiago Quintino, Baudouin Raoult, Simon Smart, James Hawkes and Peter Bauer*, Extreme Data Workshop, Jülich, Germany, 18 September 2018.

Development of a Domain Specific Distributed Object-Store For Numerical Weather Prediction and Climate Data, *Simon Smart, Tiago Quintino, Baudouin Raoult and Peter Bauer*, HPC-IODC: HPC I/O in the Data Center Workshop, ISC, Frankfurt, Germany, 28 June 2018.

The publications and slides from the presentations can be downloaded from our website.