SAGE: Percipient Storage for Exascale Data Centric Computing

(An Update)

Malcolm Muggeridge and Sai B. Narasimhamurthy, Seagate Systems UK

Abstract: We briefly describe a "Percipient" Storage system designed for data centric computing at Exascale, along with its ecosystem components and applications. Percipient Storage will be one of the first Blue Prints of potential storage system architectures that can be realised for BDEC as we move towards the era of Exascale compute, with the need to integrate the analysis of massive data sets.

Exascale computing is characterised by the availability of infrastructure to support computational capability in the order of an Exaflop. The definition is now more broadly understood to include the storage and processing of massive data volumes in storage pools of around an Exabyte as part of a scientific, social, or engineering workflow, or created through simulation. We envision Exascale capable infrastructures to be available in the 2022 timeframe, that will be exploited by applications and workflows for science and technological innovation. There has been innovation in computing infrastructure driven by Moore's law and the development of heterogeneity with multi-core and many-core processing which will be supported by increasing levels of parallelism in Exascale class codes. However I/O and storage have lagged far behind computing. As an example, compute core concurrencies (billion-way concurrency on some machines!) at Exascale will have increased about 4000 times compared to early Petaflop machines, but the storage performance in the same time period is predicted to only go up by about 100 times. In fact, the performance of disk drives per unit capacity is actually decreasing with continuing growth of high capacity disk drives. Flash based devices equally improve both in capacity and performance but have limitations in their endurance inherent within their formulation. Simultaneously the landscape for storage is expected to change with the emergence of new storage device technologies such as Non Volatile Memory. The optimal use of these devices in the I/O hierarchy, combined with existing disk technology, is only beginning to be explored in HPC.

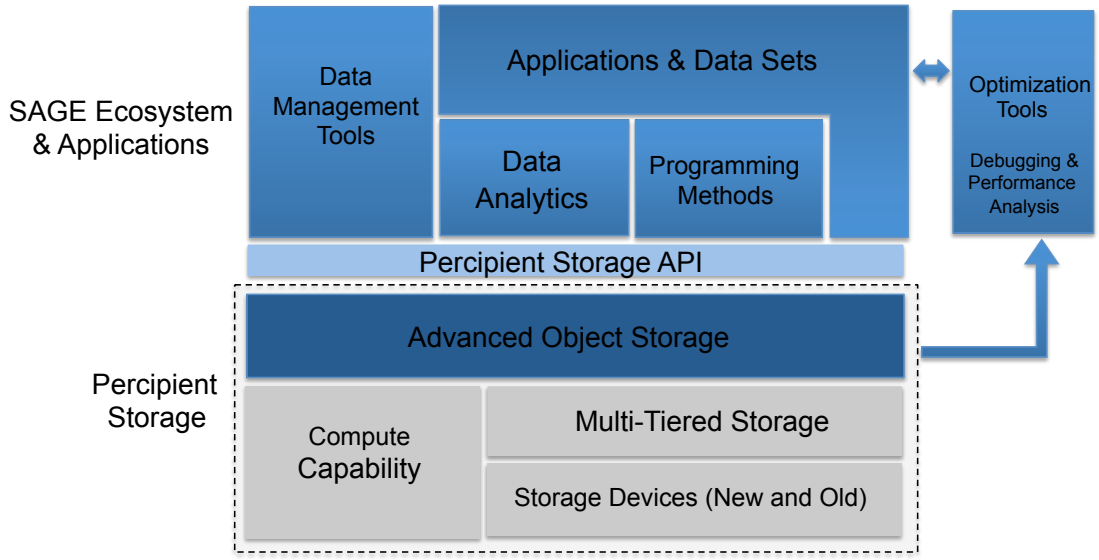
SAGE [1] proposes hardware to support multiple tiers of I/O devices and the associated intelligent management software, providing a demonstrable path towards Exascale. SAGE proposes a novel approach in extreme scale HPC, of moving traditional computations typically done in the compute cluster, to the storage system, dramatically reducing the energy footprint of the overall system. This helps in meeting the overall system Performance/Watt goals required to meet the needs of Exascale class systems. Further this helps to support BDEC workflows that will consist of not just extreme I/O workloads, but also Big Data Analytics workflows, thanks to data generated from scientific instruments and sensors - insights from which need to be fed back to running simulations. Percipient Storage in SAGE can be used as a "pure-play" I/O system dealing with very high data rates and volumes, a system that accepts partial computations from applications, and, as a system that runs full-fledged pre/post processing functions in parallel with running simulations.

---

[1] www.sagestorage.eu

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 671500 . SAGE consists of a consortium of 10 partners in Europe led by Seagate Systems, UK.

The figure below shows the architecture of the SAGE system based on the vision of "Percipient Storage" that we previously proposed[2].



The SAGE storage system consists of a multi-tiered storage system consisting of multiple storage device technologies (including NVRAM and Flash) – with provision for in-built compute close to the tiers. At the core of the system will be an Advanced Object Store with many Exascale components co-designed and built on top of a basic Object I/O and Key Value Store including the capability to accept compute functions. The Percipient Storage API (Termed "Clovis") is exposed to data management tools such as *Advanced HSM* and *Exascale data Integrity checking* added as "plug-ins". Apache Flink works on top of the API to deal with pre/post processing data analytics workflows. Programming models such as MPI (& MPI-IO) and PGAS can work on top of the Clovis API exploiting the Exascale features of the Object Store and the tiers of NVRAM. Co-Design is pursued with applications and data sets from Space Weather, Bio-informatics, Nuclear Fusion, Synchrotron light sources, Brain Simulation, Climate & weather forecasting which all have a extreme scaling needs in terms of I/O and needs for data analysis. Further, structured telemetry data from the Percipient Storage system is continuously fed to performance analysis and debugging tools generating new insights into storage and I/O performance and flexible methods to analyse failures like never before.

We have completed the process of architecture and design of many of the components for SAGE, and have now started to work through the implementations. The SAGE hardware will be hosted at Juelich supercomputing center in mid-2017 where the value of SAGE will be demonstrated for the use cases. More details of the architecture and the details of the individual components of SAGE will be described in upcoming workshops.

---

[2] "Percipient Storage: A Storage Centric Approach to BDEC", Malcolm Muggeridge and Dr. Sai Narasimhamurthy, Presented at BDEC Barcelona Workshop, 2016