Final Technical Report

Grant Number: DE-SC0008455

Title: Partnership For Edge Physics Simulation

Rutgers PI: Manish Parashar

Project PI: C. S. Chang (PPPL)

Period of Performance: 8.1.12 to 7.31.17

Overall Statement of Work: In this effort, we will extend our prior work as part of CPES (i.e., DART and DataSpaces) to support in-situ tight coupling between application codes that exploits data locality and core-level parallelism to maximize on-chip data exchange and reuse. This will be accomplished by mapping coupled simulations so that the data exchanges are more localized within the nodes. Coupled simulation workflows can more effectively utilize the resources available on emerging HEC platforms if they can be mapped and executed to exploit data locality as well as the communication patterns between application components. Scheduling and running such workflows requires an extended framework that should (1) provide a unified hybrid abstraction to enable coordination and data sharing across computation tasks that run on the heterogeneous multi-core-based systems, and (2) develop a data-locality based dynamic tasks scheduling approach to increase on-chip or intra-node data exchanges and in-situ execution. This effort will extend our prior work as part of CPES (i.e., DART and DataSpaces), which provided a simple virtual shared-space abstraction hosted at the staging nodes, to support application coordination, data sharing and active data processing services. Moreover, it will transparently manage the low-level operations associated with the inter-application data exchange, such as data redistributions, and will enable running coupled simulation workflow on multi-cores computing platforms.

Key Activities and Achievements During the Period: During this reporting period, our research activities focused on (1) extending the existing DataSpace framework (within ADIOS) to support large scale in-node data sharing between coupled applications to reduce network data movement, as shown in Fig.1; and (2) utilizing the existing DataSpaces framework (within ADIOS) to enable in-staging XGC1-analysis coupling workflow using f0 data. Specifically, we have successfully utilized DataSpaces to support large-scale in-situ execution of the XGC1 + XGCa coupling workflow, as well as coupling XGC1 with the contour plotting analysis code to support in-staging f0 data visualization (as illustrated in Fig 2). Our experiments demonstrated that both, the XGC1-XGCa and the XGC1-visualization workflows scale to 16K cores on Titan at ORNL, and achieve improvement in IO performance and overall end-to-end execution. Fig 3 compares the reading performance for the XGC1-XGCa workflow using different coupling approaches.

We also designed a framework that co-locates data staging with application execution on the same set of compute nodes and uses node-local storage resource, eg, dynamic random-access memory (DRAM), to cache application data that needs to be shared, exchanged, or accessed. Co-located data staging provides low-latency high-throughput write performance and significantly reduces the volume of data movement over network. It implements location-aware data movement and dynamically selects the appropriate transport mechanism depending on the data locations, eg, local or remote memory. For example, it uses hardware-supported RDMA network operation for fetching data resides on remote compute nodes and uses direct memory access to fetch data in node-local shared memory segment. The framework also uses a data-centric task placement approach to map workflow communications onto physical compute nodes, with the goals of reducing the communication costs. We observed that the communication- and topology-aware task mapping approach effectively reduced the size of network data communication.

We also explored extending DataSpaces staging across multiple memory hierarchy levels, e.g., both DRAM and SSD. Specifically, we explored machine learning based approaches to capture the data access patterns between components of staging-based in-situ application workflows, and to use these learned access patterns to move data between the storage layers of the staging service in an autonomous manner. We used various n-gram models to dynamically manage and optimize data movement across multiple layers of the storage hierarchy. Specifically, incoming read requests to the data staging servers are tracked at runtime to build n-gram models. These models are used to anticipate future requests for prefetching data objects from SSDs to DRAM. This reduces application perceived SSD access overheads and improves the overall data read time. Our preliminary experimental results demonstrated that n-grams models can capture dynamic data access patterns involving multiple applications and can efficiently reduce the data read time.

The project has provided training for graduate students as well as post-docs. The research conducted as part of this project has been a part of student Ph.D. theses.

Our next steps in this project include using DataSpaces to support appropriate in-situ data manipulation (e.g., preferentially sample non-Maxwellian features) on particle data in the coupled XGC1-XGCa workflow, and integrating this feature into ADIOS.

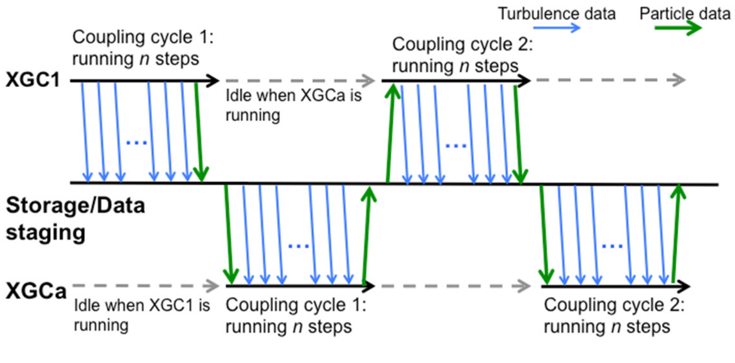
Turbulence data
Particle data
Coupling cycle 1: running *n* steps
XGC1
Idle when XGCa is running
Coupling cycle 2: running *n* steps
...
Storage/Data staging
XGCa
Idle when XGC1 is running
Coupling cycle 1: running *n* steps
Coupling cycle 2: running *n* steps

Figure 1. Data flow between two cycles of the XGC1-XGCa coupled workflow using DataSpaces.

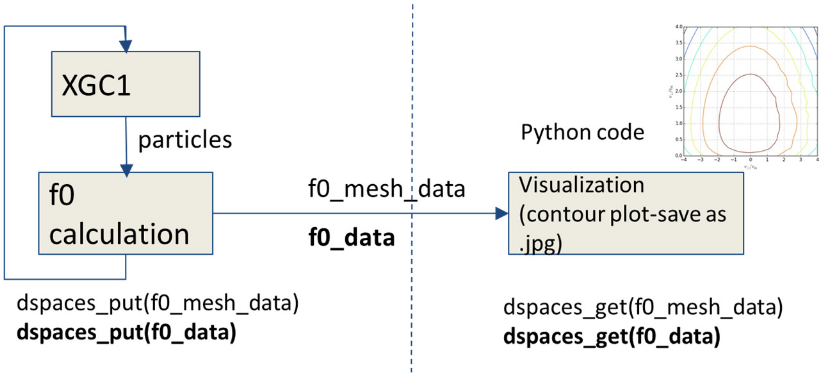

Figure 2. Illustration of XGC1-Visualization coupled workflow using data staging.

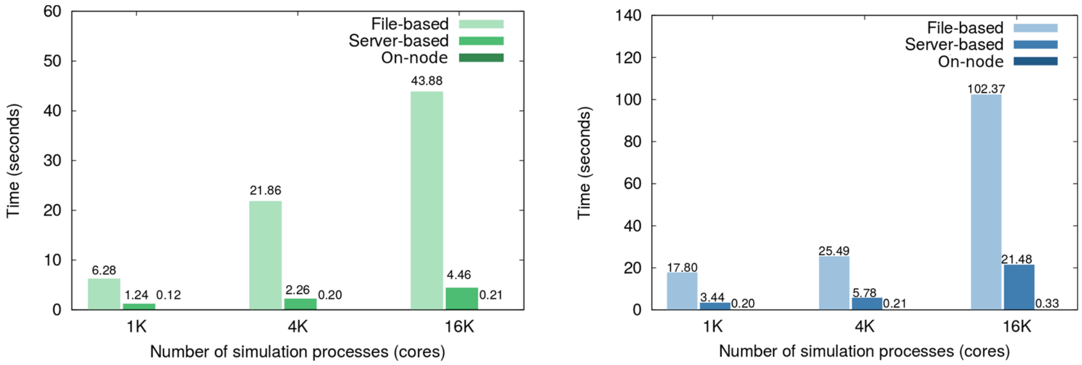

Figure 3. Experimental results of running XGC1-XGCa coupled workflow with three types of coupling methods – file-based coupling, staging-server based coupling, and on-node in memory based coupling. (Left): total time taken reading particle data; (right): total time taken reading turbulence data.

Recent Publications/Products:

- Zhang F, Jin T, Sun Q, Romanus M, Bui H, Klasky S, Parashar M. **In-memory staging and data-centric task placement for coupled scientific simulation workflows**. Concurrency and Computation: Practice and Experience. 2017 Jun 25;29(12).
- Klasky S, Suchyta E, Ainsworth M, Liu Q, Whitney B, Wolf M, Choi J, Foster I, Kim M, Logan J, Mehta K. **Exacution: Enhancing Scientific Data Management for Exascale**. InDistributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on 2017 Jun 5 (pp. 1927-1937). IEEE.
- Sun Q, Romanus M, Jin T, Yu H, Bremer PT, Petruzza S, Klasky S, Parashar M. **In-staging data placement for asynchronous coupling of task-based scientific workflows**. InExtreme Scale Programming Models and Middlewar (ESPM2), International Workshop on 2016 Nov 18 (pp. 2-9). IEEE
- D'Azevedo E, Abbott S, Koskela T, Worley P, Ku SH, Ethier S, Yoon E, Shephard M, Hager R, Lang J, Choi J. **The fusion code XGC: Enabling kinetic study of multi-scale edge turbulent transport in ITER**. Simmetrix Inc., Clifton Park, NY (United States); 2017 Jan 1.
- Melissa Romanus, Fan Zhang, Tong Jin, Qian Sun, Hoang Bui, Ivan Rodero, Jong Choi, Salomon Janhunen, Robert Hager, Scott Klasky, Choong-Seock Chang, Manish Parashar. **"Persistent Data Staging Services for Data Intensive In-Situ Scientific Workflows,"** *In The 7th International Workshop on Data-intensive*

Distributed Computing in conjunction with the 25th International ACM Symposium on High Performance Parallel and Distributed Computing(HPDC'16), Kyoto, Japan

- Jong Choi, Jeremy Logan, George Ostrouchov, Norbert Podhorszki, Scott Klasky, Tong Jin, Qian Sun, Manish Parashar. **"Mitigating I/O Variability with Staging. "** *Submitted to the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'16), Salt Lake City, Utah, USA*
- Qian Sun, Tong Jin, Melissa Romanus, Hoang Bui, Fan Zhang, Hongfeng Yu, Hemanth Kolla, Scott Klasky, Jacqueline Chen, Manish Parashar. **"Adaptive data placement for staging-based coupled scientific workflows."** *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC 15), Austin, TX, USA.*
- Salomon Janhunen, Robert Hager, Seung-Hoe Ku, Choong-Seock Chang, Jan Hesthaven, Jong Choi, Fan Zhang, Manish Parashar. **" Integrated multi-scale simulations of drift-wave turbulence: coupling of two kinetic codes XGC1 and XGCa."** Bulletin of the American Physical Society, Vol. 60, 2015.
- Deployed DataSpaces (dataspaces.org) as part of the ADIOS to support coupling and in-situ workflows at extreme scales.