Topology-Aware Asynchronous I/O

Advisors

- François Tessier - francois.tessier@inria.fr
- Joshua Bowden - joshua-charles.bowden@inria.fr
- Gabriel Antoniu - gabriel.antoniu@inria.fr

Keywords

HPC, I/O, locality, process placement

Subject

Large-scale simulations running on leadership-class **supercomputers** generate massive amounts of data for subsequent analysis and visualization. Under heavy access, the performance of traditional HPC storage systems shows their limitations and exhibits high variability. Damaris [1] is a middleware system that leverages dedicated cores in multicore nodes to offload data management tasks, including I/O, data compression, scheduling of data movements, in-situ analysis and visualization. Damaris scaled up to 16,000 cores on Oak Ridge's leadership supercomputer 'Titan' (which was first in the Top500 supercomputer list in Nov. 2012) and was tested on other top supercomputers (e.g. University of Tennessee 'Kraken'; Oak Ridge, Tennessee 'Jaguar').

The ongoing increase in the number of computing cores on computational nodes raises new challenges, particularly in terms of locality. In the context of Damaris, the question now arises as to which cores should be dedicated to data management tasks. Proximity to the network card or sharing a cache with the local processes providing the most data are some of the criteria that could influence the I/O performance achievable by the middleware. In addition, depending on requirements, the selected cores could change dynamically during the execution or the I/O process could be spawned on-demand on the most suitable resource [2].

In this internship, we will explore **affinity and topology-aware placement techniques** to determine a **near-optimal placement of the Damaris I/O processes**. To do this, we explore the capabilities of the TopoMatch [3, 4] tool developed at Inria Bordeaux and TAPIOCA, a data aggregation library [5]. TopoMatch is a set of process mapping algorithms capable of dealing with any type of topology. In addition, the work may add capability to the Damaris configuration system to specify the allocation of ranks between the simulation and Damaris.

The student will first become familiar with the HPC domain by manipulating I/O benchmarks on a computing platform like Grid'5000. Secondly, they will have the objective to take Damaris in

hand and run experiments with **Code Saturne** [6, 7], a Damaris-enabled computational fluid dynamics (CFD) simulation developed over the past 25 years at the French energy company EDF. Finally, the goal will be to implement and evaluate different I/O process placement strategies in Damaris. These results could lead to the submission of a scientific paper in a conference of the field.

The selected student will have the opportunity to join a very dynamic international research team at **Inria Rennes** in a stimulating work environment with a lot of active collaborations. This internship comes with an important opportunity to pursue a thesis co-supervised by the CEA and Inria as part of the national NumPEx project [8], the aim of which is to prepare for the arrival of the first French Exascale system in 2025.

Skills and abilities

- Programming skills (Bash, C/C++, Python)
- Knowledge of computer networks and distributed systems
- Familiarity with high-performance computing or cloud computing is an advantage

Bibliography

[1] Dorier, Matthieu & Antoniu, Gabriel & Cappello, Franck & Snir, Marc & Orf, Leigh. (2012). Damaris: How to Efficiently Leverage Multicore Parallelism to Achieve Scalable, Jitter-free I/O. Proceedings - 2012 IEEE International Conference on Cluster Computing, CLUSTER 2012. 155-163. 10.1109/CLUSTER.2012.26.

[2] Estelle Dirand, Laurent Colombet, Bruno Raffin. TINS: A Task-Based Dynamic Helper Core Strategy for In Situ Analytics. SCA18 - Supercomputing Frontiers Asia 2018, Mar 2018, Singapore, Singapore. pp.159-178, ⟨10.1007/978-3-319-69953-0_10⟩. ⟨hal-01730910⟩

[3] E. Jeannot, G. Mercier and F. Tessier, "Process Placement in Multicore Clusters:Algorithmic Issues and Practical Techniques," in IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 4, pp. 993-1002, April 2014, doi: 10.1109/TPDS.2013.104.

[4] Emmanuel Jeannot. Process mapping on any topology with TopoMatch. Journal of Parallel and Distributed Computing, 2022, 170, pp.39-52. ⟨10.1016/j.jpdc.2022.08.002⟩. ⟨hal-03780662⟩

[5] F. Tessier, V. Vishwanath and E. Jeannot, "TAPIOCA: An I/O Library for Optimized Topology-Aware Data Aggregation on Large-Scale Supercomputers," 2017 IEEE International Conference on Cluster Computing (CLUSTER), Honolulu, HI, USA, 2017, pp. 70-80, doi: 10.1109/CLUSTER.2017.80.

[6] Frédéric Archambeau, Namane Méchitoua, Marc Sakiz. Code Saturne: A Finite Volume Code for the computation of turbulent incompressible flows - Industrial Applications. International Journal on Finite Volumes, 2004, 1 (1).

[7] EDF, “Code_saturne website.” https://www.code-saturne.org/cms/web/, 2023.

[8] https://numpex.irisa.fr/