Data Scaling for Efficient Processing on HPC Clusters

Mike Mikailov, Weijie Chen, Weizhe Li, Nicholas Petrick, Fu-Jyh Luo, Stuart Barkley, Dillip Emmanuel, Rusif Eyvazli
Center for Devices and Radiological Health (CDRH), Office of Science and Engineering Laboratories (OSEL), Division of Imaging Diagnostics and Software Reliability (DIDSR)

Background

Exponential growth of data

HPC clusters support a wide array of applications with exponentially growing data in: Bioinformatics analysis; Artificial intelligence/machine learning; Genomics; Next-generation sequence analysis and alignment; Modeling and simulation; Statistical analysis and more.

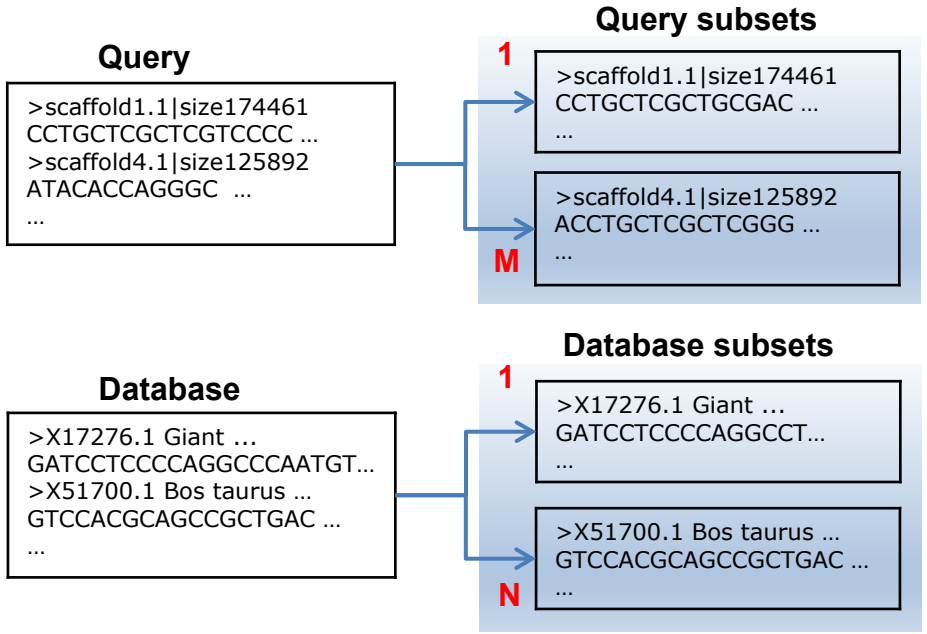
Data scaling

However, applications that do not utilize appropriate scaling techniques have limited ability for taking advantage of the HPC resources and may overwhelm even powerful HPC clusters. Our data scaling techniques overcome the challenges for the application areas.

Materials & Methods

Bioinformatics

For applications such as BLAST large query and reference datasets are partitioned into M and N subsets, which are combined into MxN unique pairs and processed in parallel using MxN tasks.



Divide and Conquer.

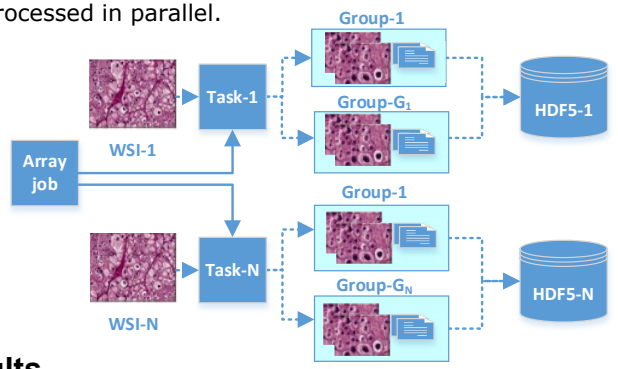
Quickly **scale** your applications on massively parallel HPC clusters!

Join the discussion 9/30/20 11AM-1PM ET:
https://fda1.webex.com/fda1/j.php?MTID=m3498b35080a4ddf4246e24ec33e49ba3

Contact:
Mike.Mikailov@fda.hhs.gov

Digital Pathology Deep Learning AI

Large-size gigapixel whole-slide images (WSI) are partitioned and grouped in hierarchical data format (HDF5) files, designed for concurrent access, and processed in parallel.



Results

- Applying BLAST to one of the FDA Bioinformatics projects reduced run time from 27 days to a single day using 4,104 tasks (where M=152, N=27 subset size 100 MB), each task taking less than 7 minutes to complete.
- CAMELYON datasets of 399 WSIs (>700 GB) were partitioned into 27,280 groups of ~230 MB in 399 HDF5 files for testing a DLNN. With 27,280 tasks this might take upwards of 18 years on a single CPU core, 30 days on a single GPU or less than 45 hours when implemented in parallel on the HPC.

Conclusions

- The scaling techniques presented here are already in use by FDA scientists.
- The techniques enable reduction of the data subset processed by each job task to a size that fits into the memory of the computing nodes where computations are performed. This avoids expensive I/O for swapping and produces excellent results, enabling substantial drops in run times.
- The described methods use only open source code, adds no hardware cost.

Acknowledgments

Kyle J. Myers, PhD, Division Director, CDRH/OSEL/DIDSR and many others at FDA.

References

1. Mikailov, M., Luo, F., Barkley, S. *et al.* Scaling bioinformatics applications on HPC. *BMC Bioinformatics* **18,** 501 (2017). https://doi.org/10.1186/s12859-017-1902-7
2. Mikailov, M., Qiu, J., Luo, F.-J., Whitney, S., & Petrick, N. (2020). Scaling modeling and simulation on high-performance computing clusters. *SIMULATION*, 96(2), 221–232. https://doi.org/10.1177/0037549719878249