UiT / NORGES ARKTISKE UNIVERSITET

April 2019

Leveraging High-Performance Computing (HPC) for Big Data Analytics (BDA)

Presenter: Phuong Ngoc Chau

Overview

- High-Performance Computing (HPC)
- Convergence between HPC and Big Data Analytics
- State of the art Frameworks
- Conclusion

High-Performance Computing (HPC)

Overview

- Convergence between HPC and Big Data Analytics
- State of the art Frameworks
- Conclusion

High-Performance Computing (HPC)

UiT / NORGES ARKTISKE UNIVERSITET

- Tools and systems available to implement and create high performance computing systems
- Used for scientific research and computational science
- Main area of discipline is developing parallel processing algorithms and software so that programs can be divided into small independent parts and can be executed simultaneously by separate processors

High-Performance Computing (HPC)

UiT / NORGES ARKTISKE UNIVERSITET

- When are we using HPC?
  - The problem take too much time.
  - The problem does not fit on personal computer.
- One of the main objective of High Performance Computing(HPC) field is to provide infrastructures to make computing as fast as possible, at the largest possible scale.

Overview

High-Performance Computing (HPC)

Convergence between HPC and Big Data Analytics

State of the art Frameworks

Conclusion

Convergence between HPC and Big Data Analytic

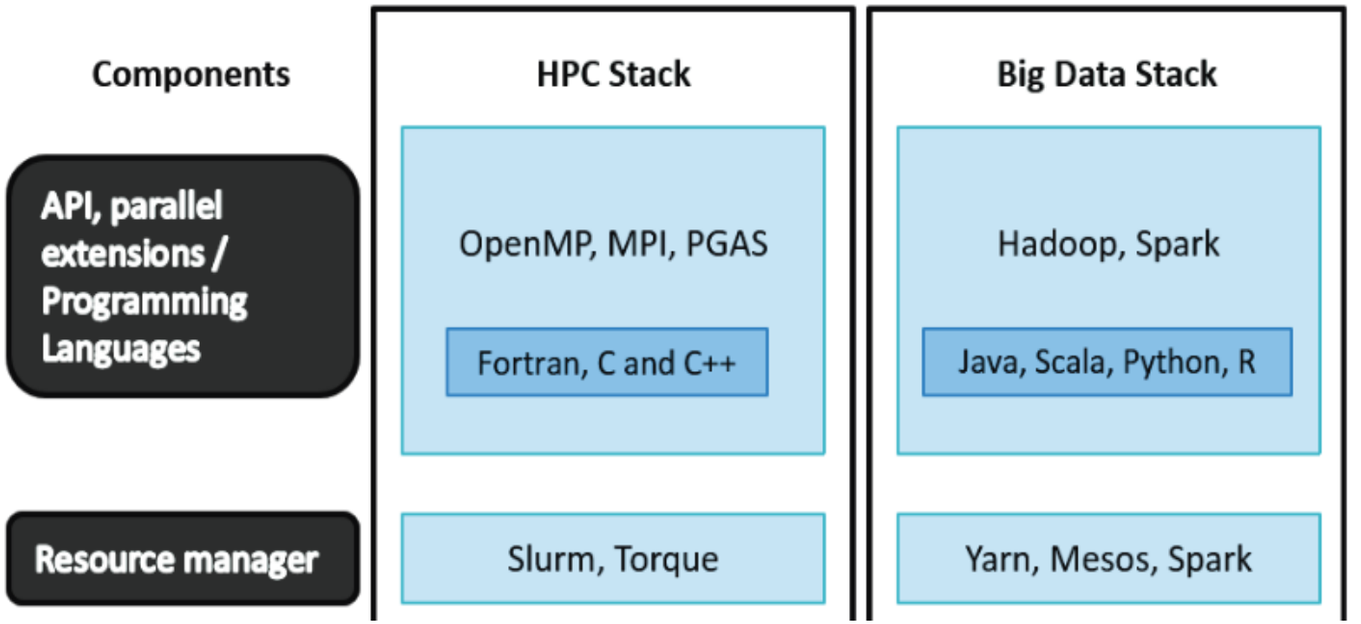
UiT / NORGES ARKTISKE UNIVERSITET

- Big Data targets applications that need to handle very large and complex data-sets.
- Thus, Big Data applications are very demanding in terms of storage, to accommodate such a massive amount of data

→ Convergence between Big Data and HPC frameworks

Convergence between HPC and Big Data Analytic

- Big Data and HPC software stacks [1].



The source of this picture is from "A comparative survey of the HPC and big data paradigms: Analysis and experiments"

Overview

High-Performance Computing (HPC)

Convergence between HPC and Big Data Analytics

State of the art Frameworks

Conclusion

Previous Methods

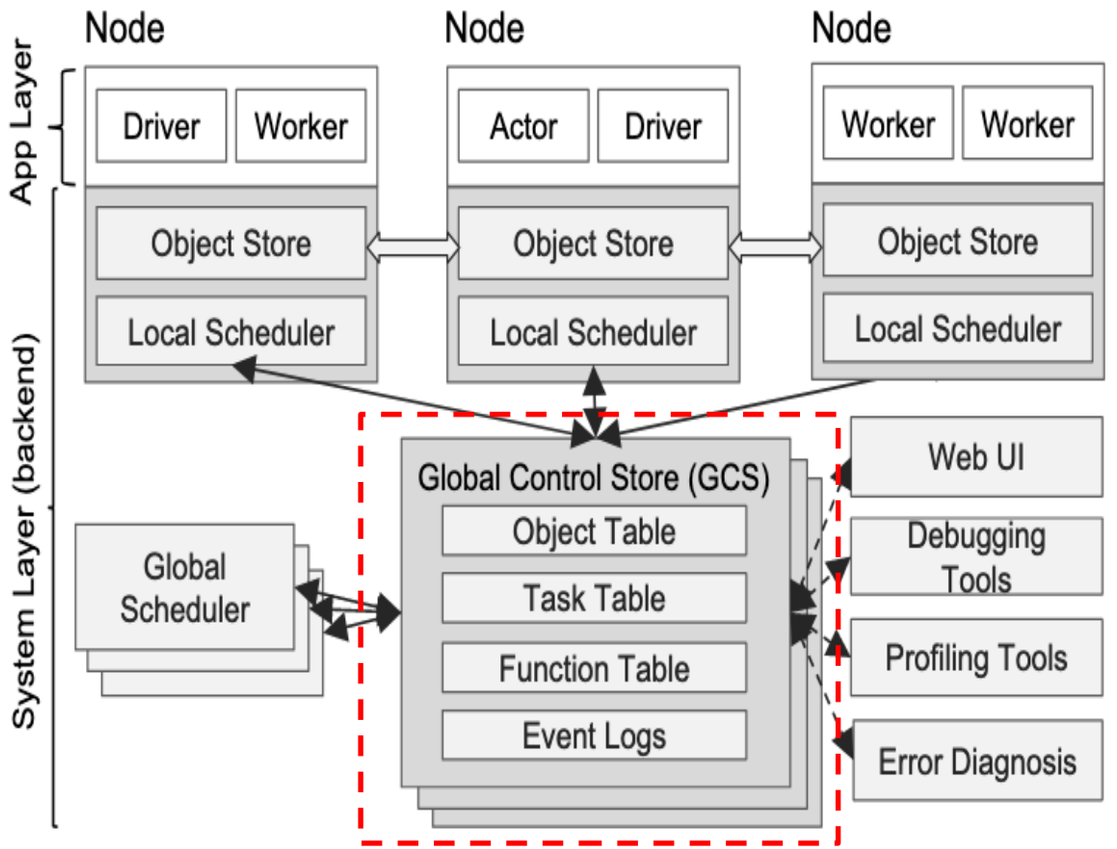
RAY:

- Framework for more efficient and Expressive Distributed computing.
- A general-purpose cluster-computing frame work → enables simulation, training and serving for RL application.
- Implementing a unified interface to express task parallel and actor-based computational.

→ supported by a single dynamic execution engine.

Previous Methods

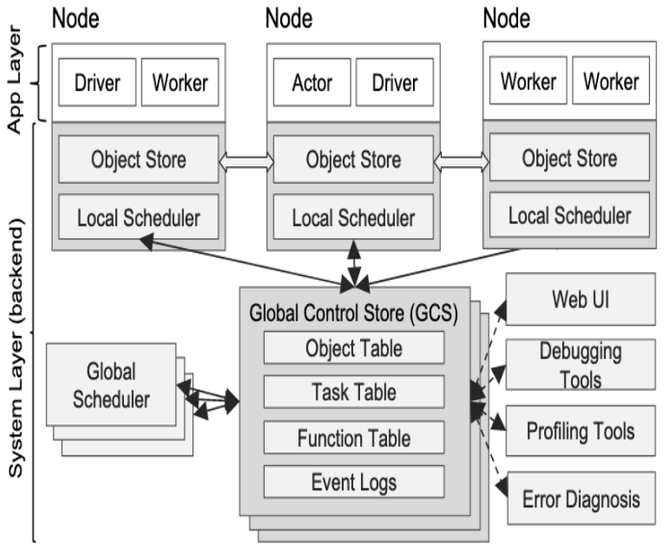
- RAY



The source of this picture is from "Ray: A distributed framework for emerging {AI} applications"

Previous Methods

- RAY



The source of this picture is from "Ray: A distributed framework for emerging {AI} applications"

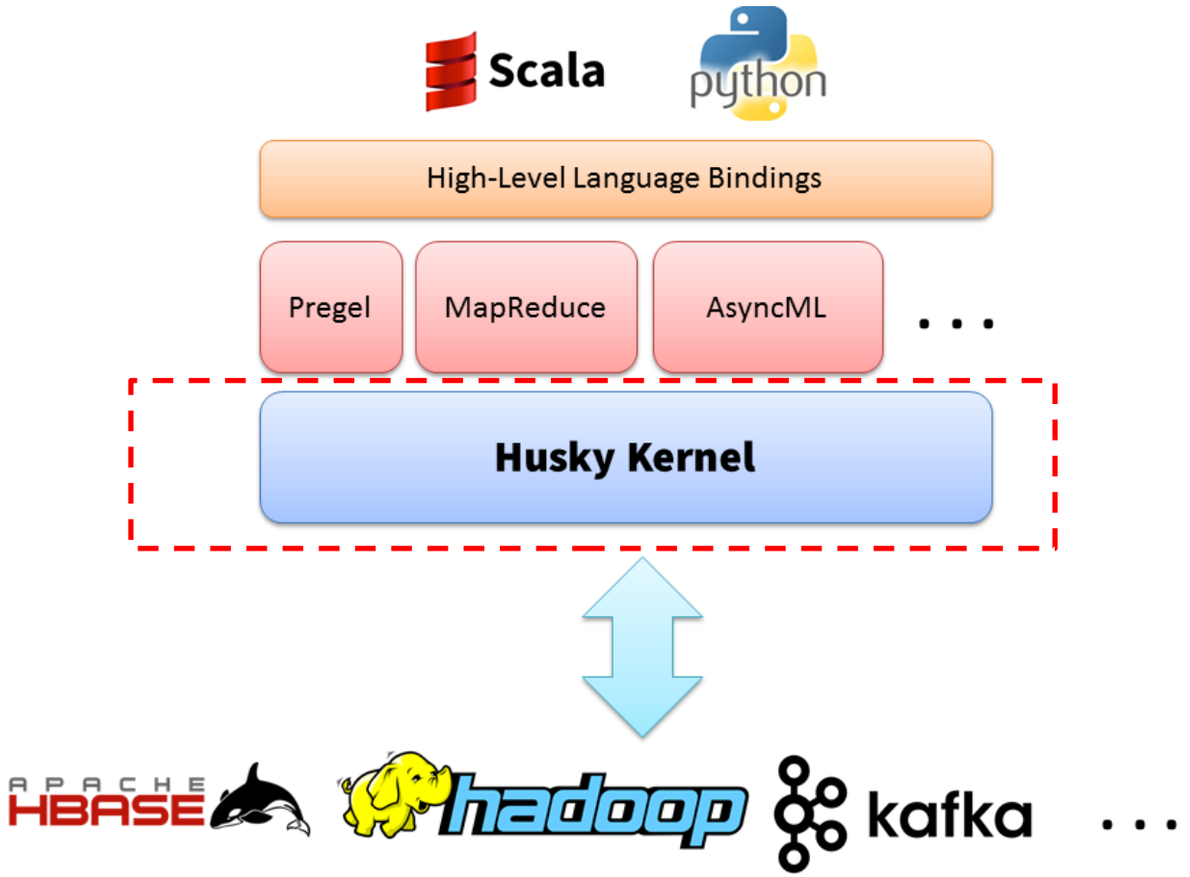
Previous Methods

UiT / NORGES ARKTISKE UNIVERSITET

- HUSKY
  - Distributed Framework for Emerging Application
  - More low-level tools for bug-prone programming.
  - Balance between high performance and low development cost.
  - Develop for memory large scale data mining and efficient distributed algorithms.
  - Existing frameworks can be easily implemented and bridged together inside Husky.

Previous Methods

- HUSKY



The source of this picture is from "http://www.husky-project.com/"

Overview

High-Performance Computing (HPC)

Convergence between HPC and Big Data Analytics

State of the art Frameworks

Conclusion

Conclusion

UiT / NORGES ARKTISKE UNIVERSITET

- My project is tended to improve state of the art HPC frameworks for Big Data Analytics.
- There is still plenty of room for improvement in terms of communication and synchronization overheads, and performance of applications.
- In addition, in today's digital data, large data services make power consumption a large part of the total cost. Therefore, the convergence between HPC and big data need to be handled.

Conclusion

UiT / NORGES ARKTISKE UNIVERSITET

- Firstly, I will test HUSKY and RAY with some fundamental algorithms in HPC like PageRank, K-Means, etc.
- Secondly, Basing on these results, I try to improve performance, synchronization, fault tolerance, etc.
- Finally, I will try to implement an efficient framework that leveraging HPC to improve performance for BDA.

References

[1] Asaadi, HamidReza, Dounia Khaldi, and Barbara Chapman. "*A comparative survey of the HPC and big data paradigms: Analysis and experiments*." 2016 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2016.

[2] M. Mercier, D. Glesser, Y. Georgiou, and O. Richard, “Big Data and HPC collocation : *Using HPC idle resources for Big Data Analytics*,” pp. 347–352, 2017

[3] Moritz, Philipp, et al. "*Ray: A distributed framework for emerging {AI} applications*." 13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18). 2018.

[4] Uta, Alexandru, et al. "*Exploring hpc and big data convergence: A graph processing study on intel knights landing*." 2018 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, 2018.

[5]Yang, Fan, Jinfeng Li, and James Cheng. "*Husky: Towards a more efficient and expressive distributed computing framework*." Proceedings of the VLDB Endowment 9.5 (2016): 420-431.

UiT / NORGES ARKTISKE UNIVERSITET

THANK YOU FOR LISTENING!

UiT / NORGES ARKTISKE UNIVERSITET

Q&A