トップエスイー修了制作

アプリケーションに特化したHadoop タスク分割設計の提案

株式会社インテック　沖田 弘明　okida_hiroaki@intec.co.jp

開発における問題点

大規模データのバッチ処理フレームワークとして、Hadoopが一般的に用いられている。
しかしながら、Hadoopアプリケーション開発において、個々のタスクサイズのバラつきを考慮したMapReduce設計は難しい。その場合、ノード毎の処理時間に差が生じ、結果として、全体のスループットの低下を招いてしまう。

手法・ツールの適用による解決

当社で開発した次世代シーケンサー解析プラットフォーム（NGS解析システム）を対象に、アプリケーションに特化したHadoopタスク分割設計を適用し、問題を解決する。
NGS解析システムはゲノム配列を解析するプラットフォームであることから、“ゲノム”の特徴を適用したHadoopタスク分割設計を行う。

研究課題とアプローチ

課題

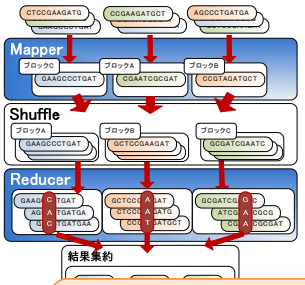
- Hadoopアプリケーション開発において、各ノードで実行するタスクの処理時間を均一にするMapReduce設計を行う

アプローチ

- アプリケーションに特化した情報から、MapReduce設計に有効な特徴を抽出する
- Hadoopアプリケーション開発において、抽出した特徴を適用したMapReduce設計を行う

提案手法の概要

【NGS解析システムの解析フロー】



【ゲノム情報】

※ NCBI Genomeページから転載

① タスク分割設計に有効な特徴を抽出

② 抽出した特徴をタスク分割設計に適用

検証

edubase Cloud上に10台の計算ノードからなるNGSシステムを構築し、1億Readの配列を対象とした検証を実施

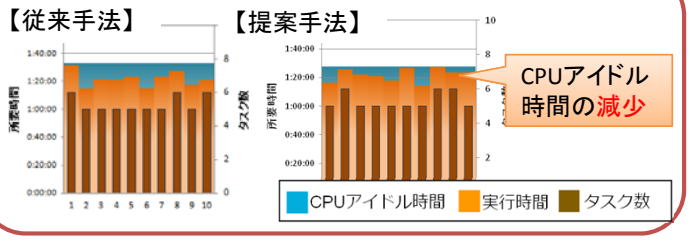
全体の処理時間

	従来手法	提案手法
処理時間合計	3:03:30	2:50:48

全体の処理時間の短縮

各ノードの処理時間

【従来手法】　【提案手法】



CPUアイドル時間の減少

評価と課題

評価

- アプリケーションに特化したHadoopタスク分割設計を適用することで、従来のタスク分割設計に比べ、各ノードにおけるCPUアイドル時間の減少し、全体の処理時間の短縮が期待できることがわかった。
- 各ノードにおけるCPUアイドル時間のバラつきを減少するために、新たな特徴を適用したタスク分割設計を検討する必要がある。

課題

- 大容量の解析データを対象とした場合の検証
- 未適用の特徴を適用したタスク分割設計の検討