インタークラウドの活用事例

Overlay Cloudによるバイオインフォマティクス パイプライン再現環境の構築

横山 重俊　政谷 好伸（国立情報学研究所）小笠原 理（国立遺伝学研究所）
大田 達郎（ライフサイエンス統合データベースセンター）吉岡 信和　合田 憲人（国立情報学研究所）

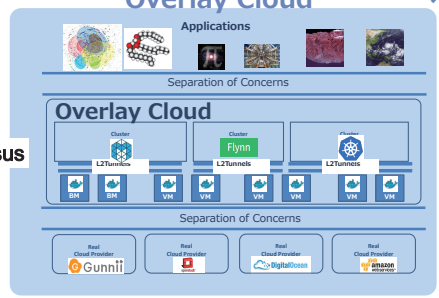
バイオインフォマティクス分野では、論文の**再現性確保要請**に対応するため、DNA塩基配列の公共データベース構築等による**データ共有**やデータ解析ソフトウェアのオープンソース化により**処理プログラムの共有**が進んでいる。再現性確保には以下の**課題**がある。
1. データ処理ソフトウェアの複雑化・多様化、2. 次世代シーケンサー普及による**大量データ分散発生**　1) 公共データベースの巨大化、2) 発生データの分散化、3. **データ解析量の増大**。
本研究では、以下の解決策をインタークラウド上でバイオインフォマティクスパイプライン再現環境に適用することで、**パイプラインの可搬性**および**処理時間短縮**を目指している。**1. データ解析ソフトコンテナ化**によるクラウドを跨る可搬性確保、2. **コンテナ分散配置**によるデータとデータ解析プログラム間遅延削減、3. **分散処理基盤利用**によるデータ解析ソフトウェアの処理性能向上

Overlay Cloudによるインタークラウド

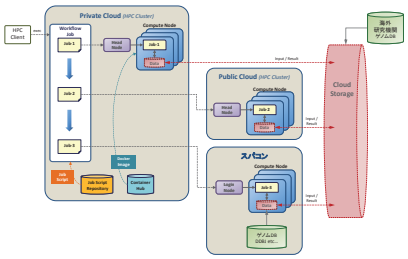
運輸業界におけるコンテナ革命

versus

IT業界におけるコンテナ革命? Overlay Cloud



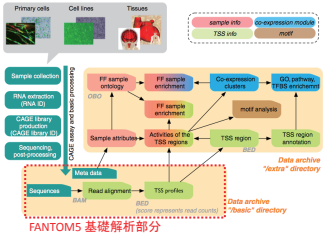
Overlay Cloudのバイオインフォマティクスへの適用



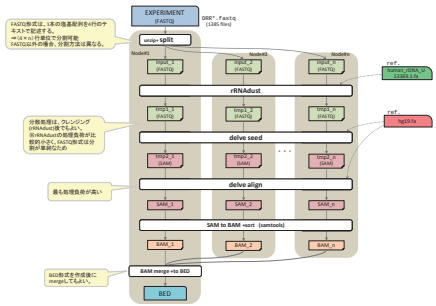
バイオインフォマティクス パイプライン再現環境構築例

FANTOM 5プロジェクト

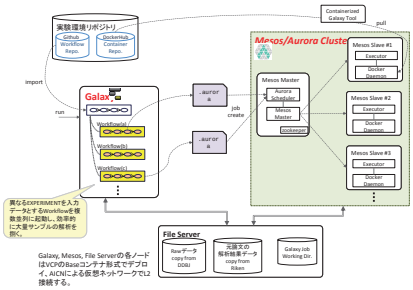
- 理化学研究所が主導、2000年にFANTOMプロジェクト発足
- 20カ国、114研究機関が参加する国際研究コンソーシアム
- ゲノムDNAから転写されているRNAの機能をカタログ化
- 成果は公共データとして公開され、世界中で利用可能



FANTOM5 基礎解析パイプライン



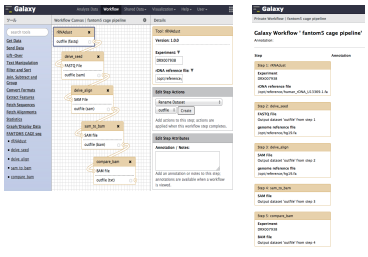
FANTOM5 基礎解析パイプライン実行環境



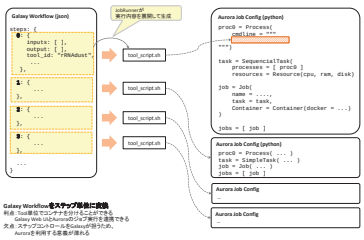
Galaxyについて

生命科学分野におけるデータ解析作業をWeb UIで実行
- 複数のツールを組み合わせたデータ解析ワークフロー構築
- 異なるサンプルに対する繰り返し処理
- 研究者間でワークフロー、データの共有

Galaxy Workflow 作成



Galaxy Workflow ⇒ Aurora Job Config



現状と今後の展開

FANTOM5基礎解析部分パイプライン再現性について適用したアーキテクチャで実現できることを確認した。また、Overlay Cloudの上に分散処理基盤を動的に構築することで、FANTOM5基礎解析部分パイプライン分散基盤上での走行確認ができた。今後は、FANTOM5基礎解析部分パイプライン再現性確認を複数クラウドをターゲットに実施し、さらにその環境を使ってFANTOM5基礎解析部分パイプライン分散化評価を行う。その後、広域分散時のコンテナ配置最適化およびFANTOM5以外の事例への横展開を実施する予定である。

NII　連絡先：横山重俊/国立情報学研究所　アーキテクチャ科学研究系/yoko@nii.ac.jp