議事録

会議名：国立遺伝学研究所スーパーコンピュータ　ユーザ会　三島開催
日時：2014 年 7 月 22 日
会場：国立遺伝学研究所　本館 2 階　会議室
概要：
1. ユーザ会趣旨説明
2. 挨拶
3. 遺伝研スパコンシステム概要
4. INSD の構築
5. 検索・解析サービス
6. 活用事例
DBCLS
大田様
公開 NGS 全体に対しバッチ処理を実行しデータベースを作成している。その際の課題としてデータ解析のパイプライン処理の共有と再実行を行うために仮想化技術(バーチャルマシン、コンテナ)の開発を始めている。ローカル環境で解析したいた際には、データ転送速度がネックとなっていた。スパコン上(遺伝研の中)で転送すると 8 倍位の速さが出た。更にスパコンでの並列化により研究上で大きな恩恵を受けている。
データ解析再現性を担保するため、ソフトウェアの version 管理が必要である。解析環境を VM のイメージファイルとして用意し共有すれば良いのでは？イメージを共有する事で環境依存の問題が減り、計算の選択肢が増える。NIG スパコンは柔軟性、永続性、秘匿性が少し劣る。
遺伝研スパコンは DDBJ のデータベースと接続されていることから、それを活かした運用を望む。
＜質疑応答＞
・スパコンからの DDBJ データベースには InfiniBand 経由でアクセスしているか？(小笠原先生)
⇒している。(大田様)

国立遺伝学研究所 生命情報研究センター 比較ゲノム解析研究室
辰本様
スパコン利用失敗事例として、UGE エラーファイルに掛かれないジョブは qacct で常に確認が必要。データ転送は Aspera が圧縮方法は gzip が高速である。Fat 計算ノードはメモリアクセスが高速でないため既存アセンブルツールには不向きである。
＜質疑応答＞
・データの増加量はどの程度を見込んでいるか？
⇒1 年間に少なくとも 100TB 以上の容量は必要。(辰本様)

名古屋大学 大学院工学研究科　計算理工学専攻　計算生物物理グループ(笹井研究室)
徳田様
出芽酵母の分子動力学シミュレーションを NIG スパコンで実施している。研究室の計算機と比べメモリが大きい点、無償であり挑戦的な課題にも取り組みやすい点、pipeline のツールがインストールされている点が評価できる。

＜質疑応答＞

・全体でどの程度の計算時間が掛かるか？(小笠原先生)
⇒今回の事例では thin ノードで数日程度。(徳田様)

・出芽酵母でなく、マウス等ゲノムサイズが大きいモデルの場合、計算時間はどの程度か？(藤山先生)
⇒すぐには正確に見積もれない。モデル化する際の可視レベルによっても異なる。(徳田様)

・スライドの「遺伝研ならでは」の意味は？
⇒無償のため、挑戦的な研究テーマも取り組めるという点がならではと考えている。(徳田様)

大阪大学蛋白質研究所

笠原様

転写サイクルのシミュレーションに関して TSUBAME と NIG スパコンを併用している。NIG スパコンは制限も少なく使いやすい。その分、モラルを守って使用する必要があり、毎年の継続申請も必要である。NIG スパコンでもユーザが公開 Web サーバを立てられれば良いと思う。

＜質疑応答＞

・DDBJ での Web サーバリソース貸出は今後の検討課題である。(小笠原先生)

・シミュレーション結果はずっと保存しておきたいものか？(高木先生)
⇒出来れば全て残しておきたいが、ディスク容量の問題もあるため、要素数を減らした形で保存している。(笠原様)
⇒必要であれば大規模申請をすることで 1TB 以上可能である。(高木先生)
⇒現状、大規模ユーザによってほとんどのディスクを利用しており、1 人あたりの容量を減らした方が良いかとも考えている。しかし、現状では必要であれば申請して利用して頂いて問題ない。(小笠原先生)

国立遺伝学研究所 生命情報研究センター 大量遺伝研究室

長崎様

多型検出を GATK で実施している。GATK は Medium ノードや Fat ノードで実施している。
ジョブが落ちた際はエラー内容を確認可能である。

大柳様

遺伝研・植物遺伝研究室

植物オミックスの DB(PODC)を作成している。NIG スパコンへの要望としてジョブとディスクが逼迫している状況の改善を望んでいる。また、NIG ユーザ以外とのデータ授受が出来る仕組みがあれば利便性が高いと考える。仮想化および特権ユーザの貸与サービスも実施して欲しい。

＜質疑応答＞

・海外ユーザとのデータ交換は自前で SSH サーバを立てて交換するなどが必要でコストも掛かるため、データ交換し易い仕組みがあれば良いと思う。(大柳様)

東京工業大学 学術国際情報センター

三浦様

HPC と BigData の融合について進めている。HPCI 共用ストレージがあり、DDBJ DB をこのストレー

ジにおいては如何か。それによって HPCI 参加サイトから DDBJ DB に直接アクセス可能となる。まずはその前段階として GSIC と DDBJ で連携しては如何か。

＜質疑応答＞

・現状 HPCI 共用ストレージはどの程度使用されているか？(藤山先生)

⇒ユーザによる利用は課題採択制となっているが、DDBJ DB の場合では別枠となるのではないかと思う。ユーザによる利用は毎年更新されるため、年度末以外は比較的空き容量がある。またバックエンドにテープも用意している。(三浦様)

7. 意見交換

・slot 数を指定するのみでは 1 ノード占有利用できないのでは?(佐々木様)

⇒1 ノード占有利用したい場合は def_slot オプションで指定すれば可能。(石川)

・殆どがシングルプロセスで動くが、部分的に複数プロセスで動くようなジョブの場合、qsub で最大リソース量を確保する必要があるが、動的にリソースを管理するような仕組みはないか？(大田様、瀬々様)

⇒現状の UGE オプションでは無い。(川越)

⇒ジョブの中で qsub することで、工夫すれば可能かと思われる。(小笠原先生)

・他機関において、ユーザ審査が通らなかった事例はあるのか？(城石先生)

⇒HPCI での採択は少なくとも 50%以下と思われる。(三浦様)

・非居住者のユーザに対する審査が厳しいのでは？時間が掛かり、海外居住者には敷居が高い。(北住様)

⇒経産省のルールに則っており、厳しくなっている。(高木先生)

⇒東工大や HPCI でも同様である。(三浦様)

⇒以前は審査の手順が決まっておらず時間が掛かったが、現在は手順が整っており審査も以前よりはスムーズである。(小笠原先生、高木先生)

・Pipeline 使用時に acknowledgement への記載方法が出ないが良いのか？(西井様)

⇒用意する予定である。(神沼先生)

⇒スーパーコンピュータ自身にはサイトに掲載している。(高木先生)

・課金制を導入する際の遺伝研のコスト増が問題とならないか？(大田様)

⇒遺伝研ではすでに課金システムが整っており、導入は難しく無い。課金のみでなく審査もきちんとやる等のハイブリッドな対応が必要ではないか？(城石先生)

・HPCI にリソースを提供し、そちらにユーザ審査を依頼しては？(桂先生)

⇒その場合、HPCI 基準での審査となるため、NIG としてのルールを定めなければならない。(三浦様)

・千単位で流れているジョブを審査するのは現実的では無い。一部の必要以上にリソースを占有しているジョブ投入ユーザのみ注意すれば良いのでは？（倉田先生）

・ユーザが効率的にジョブを投入すればより活用できるはずである。(内藤様)

・実行ジョブ数に対して課金しては？(北住様)

・ヘビーユーザに対する教育を実施すれば良いのでは？(赤間様)

⇒すでに Fat,Mediuim ノードへのジョブ投入者でリソース要求量と実使用量の乖離が大きいユーザについては、メールを送っている。(川越)

⇒ルール化して対応している訳ではないため、今後対応を検討する。(高木先生)

・解析のスナップショットを保存するサービスを提供しては？(佐々木様)

・企業ユーザに対するポリシーを明確にした方が良い。(有田先生)
・ユーザがデータを公開しても良いという意思があるかも確認した方が良い。(有田先生)
・スパコンのユーザ数が増えることへのインセンティブが無い。解析データの元手が自前データの研究者は予算があると予想でき、課金されても使うだろう。一方、publicDB を解析している研究者は予算がつきにくく、課金されると困るのでは？(大久保先生)
・ソフトウェアのバージョン管理をすることを検討する。(高木先生)
・リピート配列等があるとメモリ使用量が読めない。なるべく乖離が少なくなるよう努力はしているが要求リソースが大きくなってしまうのはある程度仕方のない事である。(辰本様)
⇒その点を意識して使用しているユーザであれば問題ないと考える。意識せずに大きなリソースを要求しているユーザが問題である。
・課金されると困る。課金は最終手段として欲しい。(数名)
・自身がヘビーユーザなのかが分からない。他のユーザと比べ自分はどの程度使用しているか分かるような仕組みがあれば良いのでは？(寺田様)
・スパコン内のデータベースの更新した日時とバージョンをどこかに記録して残していただけると助かる。論文なり、書き物においてはバージョンの記載も必要となることがあるが、自分で記録を忘れると確認が困難である。(長崎様)
⇒技術的に問題ないので対応可能である。(川越)