Changing Science through Online Analysis

Kerstin Kleese van Dam, Computational Science Initiative Director
Huub van Dam, Shinjae Yoo, Shantenu Jha, Wei Xu, Sangsoo Ha
Breaking the Paradigm

• Until recently subscribed to the \textit{post-hoc} analysis paradigm
• This meant that Calculations incl. Workflows:
  • Had to be fully planned out at conception
  • Could only adapt to pre-conceived options
  • Were only evaluated in terms of success, scientific outcome etc. after they had completed
• However, \textit{scientific discovery relies on the rare, the unexpected}!

\textit{Online analysis can more effectively support scientific discovery}
Online Adaptive Detection of Events
Initial Use Case - Addressing Data Challenge at Exascale

- 1 Million atom simulations
- Conformational changes take about a microsecond (1e\(^{-6}\))
- Time resolution of the simulation is a femtosecond (1e\(^{-15}\))
- Runtime ~12 days
- Storing everything 8*4*1e\(^6\)*(1e\(^{-6}\)/1e\(^{-15}\))=32 PB
- Typical approach: uniform sampling
  Store 1 out of 1000 structures
- Store too much and data volumes become overwhelming, store too little and you might miss the important transformations

Online Adaptive Approach

- Manifold Learning for MD trajectories
  - Matrix Sketch of MD trajectories
- Important / Interesting event detection
  - Weighted Reservoir Sampling of gradient changes

\[
Y = U \Sigma V^T \leq \varepsilon \|Y\|_F^2
\]

Low dimensional manifold projection of different state of MD trajectories
Streaming Single Value Decomposition

• Our Extension - Why only sampling one stream at a time?

\[
B = U_l \Sigma_l \\
B^{\text{new}} = U_l \Sigma_{l}^{\text{new}}
\]

\[
\ell = \Omega \left( \frac{\sqrt{m} \|Y_{[l]}\|_2^2 \|Y_{[l]} - Y_{[l:(k)]}\|_F}{L^2} \right)
\]

\[
\|w_l - \hat{w}_l\| \leq \frac{k\|Y_{[l]} - Y_{[l:(k)]}\|_F}{\sigma_{t_k}^2 + \alpha} \sqrt{\frac{\Gamma_a \Gamma_b}{\ell - k}} + \frac{\sigma_{t_k}^2 + \alpha \sqrt{L + 8\kappa^2 \|Y_{[l]}\|_2^2 \sqrt{L^2 + 16\kappa^4 \|Y_{[l]}\|_2^4}}}{\sigma_{t_k}^2 + \alpha \sqrt{2L}}
\]
A Case Study - 32 Samples

- Identifying important events when they occur, every time
NWChemEX - Achievement and Next Steps

• Orders of magnitude data reduction
• Able to detect all points of interest - never miss an event!
• Currently applied to classical MD simulations - science requires more accurate treatment of areas of interest
• Online analysis can enable us to fire off these methods when points of interest are detected, rather than after the end of a standard calculation, which would take ~12 days (for 1 microsecond of observation time)
Online Anomalie Detection
Chimbuko - Online Performance Trace Data Analysis at Exascale

• Complex architectures, applications and workflows, require detailed performance analysis

• However, performance analysis at Exascale hits the same data challenges as the science apps.

• Our solution - online analysis and reduction to events of interest
Execution Time Based Anomaly Detection

• Execution time-based detection
  • Statistics approach (e.g., confidence interval)
  • Density-based (e.g., local outlier factor)

• Functions depend on each other
  • Delay of child functions
  • Communication delay of other nodes
Streaming Performance Visualization

1. Streaming Workflow Overview in Scatter Plot
2. Dynamic functions of interest
3. Function Execution in Dynamic Call Stack Tree
4. Function Execution and Message Passing in Zoomable Timeline

- Streaming data reduction and aggregation
- Sliding time window of workflow overview for regular and anomaly function executions
Chimbuko Data Reduction > 100x
Chimbuko - Achievements and Next Steps

• First Online Performance Trace Data Analysis tool, not only for single applications, but also full workflows
• Enables real time identification and analysis or performance anomalies in realistic calculations
• Lightweight version could also support adaptive resource management for complex workflows to ensure optimal resource usage and guarantee of execution time constraints.
In-situ Compression Artifact Removal
Why compression artifact removal?

- Many Exascale Applications will create vast volumes of data, too big to store (I/O constraints)
  - Scientific simulations (fusion, fluid dynamics, climate)
  - Experimental measurement devices (neutron / light sources)
- Lossy image compression is
  - Often used to achieve high data reduction rates
  - However, leads to blocking artifacts and blurring
- If we can remove compression artifacts in-situ, we can
  - Boost scientific accuracy, while maintaining high compression rates
Residual Dense Net (RDN)

- **Speed**: Unlike a traditional approach (compressed sensing), DL approaches can reduce inference time significantly → Enable in-situ inference

- **Quality**: Residual Dense Net (RDN) shows better reconstruction than traditional approaches and other deep learning approaches (e.g., EDSR)

An overview of RDN model: global residual learning (top), and local residual learning between residual dense blocks (bottom)
Incremental Batch Transfer Learning (IBTL)
Case Study - Climate Data

(a) Original

(b) Compressed

(c) Restored RDN with IBTL
Case Study - Kinetic Fusion Data

Original

Restored RDN with IBTL

Compressed
Less is Really More!
Adaptive Evaluation and Steering of Complex Workflows
Adaptive Ensemble Algorithms

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: $O(10^6 - 10^8)$!
  - e.g. Chemistry, Biology, Climate

- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: Intermediate data, determines next stages

- Adaptive Ensemble simulations can easily be 2-3 orders of magnitude faster than vanilla
● Protein Folding using ML driven MD
  ○ ML Model: Convolutional Variational Auto Encoders (CVAE)
  ○ CVAE ingests intermediate data, determines next stages
RADICAL: Managing ML & HPC tasks on Summit
RADICAL & DeepDriveMD - Results

- Manage execution of ensembles
  - Currently: $O(10^3)$ on Summit,
  - @ Exascale: $O(10^6 - 10^8)$!

- Concurrent and adaptive training, simulation and inference tasks (right)

- DeepDriveMD: DL models to adaptively drive ensemble simulations

- Depending upon data volumes involved:
  - Stream simulation data directly to ML
  - Use “in memory” databases

- Using DL shows 20x improvement over non-DL adaptive approaches
RADICAL & Adaptive Ensembles & CANDLE

- Chemical space of drug design in response to mutations very large. 10K -100K mutations; too large for HPC simulations alone!

- Use ML to enhance the **effective performance** of HPC simulations

- Develop methods that use:
  
  (i) Simulations to train ML models to predict therapeutic effectiveness
  
  (ii) Use ML models to determine which drug candidates to simulate
Thank You!
Come and Join Us!
Brookhaven Laboratory is Hiring -
See us at the SC Job Fair,
DOE Booth Event on Thursday or
contact us BrookhavenLabCS@bnl.gov