Visualization and Data Analysis

James Ahrens, David Rogers, Becky Springmeyer

Eric Brugger, Cyrus Harrison, Laura Monroe, Dino Pavlakos

Scott Klasky, Kwan-Liu Ma, Hank Childs

Working Group Description

National Nuclear Security Administration
Defense Programs

- The scope of our working group is scientific visualization and data analysis.
  - Scientific visualization
    - refers to the process of transforming scientific simulation and experimental data into images to facilitate visual understanding
  - Data analysis
    - refers to the process of transforming data into an information-rich form via mathematical or computational algorithms to promote better understanding
  - Data management - shared with IONS
    - refers to the process of tracking, organizing and enhancing the use of scientific data
- The purpose of our work is to enable scientific discovery and understanding.
  - Our scope includes an exascale software and hardware infrastructure that effectively supports visualization and data analysis.

Identify current state-of-the-art

Defense Programs

- **Production visualization tools scalably process large data**
  - **Suite of data-parallel visualization, analysis and rendering algorithms**
    - **Open-source tools – ParaView, Visit**
    - **Commercial tool – Ensight**

- **ASC success story**
  - **NNSA/ASC created the large-data scientific visualization tools in use by other agencies (NSF, DOD) and around the world**

Identify Exascale Visualization and Data Analysis Needs

Defense Programs

- **Visualization and Data Analysis (VDA)**

- **Broad range of scope for VDA**
  - VDA as an application
  - VDA as a service
  - VDA as a systems infrastructure

- **Note: like apps, VDA capabilities will require development to exploit opportunities in evolving platforms**

1. Exascale Challenges – storing a full-range of results for later analysis becomes impossible due to technology trends

Defense Programs

- The rate of performance improvement of rotating storage is not keeping pace with compute.
- Provisioning additional disks is a possible mitigation strategy, however, power, cost and reliability issues will become a significant issue.
- A new in-situ exascale visualization and data analysis approach is needed:
  - Slow output to data hierarchy / Data movement = power/cost
  - Where will we process data?
    - Different customer-driven approaches require integration at different HW ‘levels’

2. Exascale Challenges - Exascale simulation results must be distilled with quantifiable data reduction techniques

- **Exascale as massive data**
  - **Defacto data reduction technique**
    - ① **Visualization algorithms**
    - ② **Rendering massive numbers of polygons**
  - **This puts lots of data into a single pixel, combined by the renderer**
    - **This is a workable method but is it what the user wants?**
  - **This approach provides the foundation for our current successes**
    - **brute force approach that requires significant computing resources**
    - **difficult to quantify the bias of this approach**
- **Approaches that quantifiably reduce data as it is generated need to be explored**

3. Exascale Challenges - New exascale-enabled physics approaches require corresponding new visualization and data analysis approaches

- **Implication of exascale as massive compute**
  - **Statistical physics approaches**
    - **Statistical modeling of a physical process**
  - **Parametric studies**
    - **record how a simulation responds in a parameter space of possibilities**
  - **Multi-physics approaches**
    - **simulate a linked model of different related phenomena such as a linked physics and chemistry simulation**
  - **Multi-scale approaches**
    - **simulate phenomena at different spatial and temporal scales**

- **Understanding and presenting both summarized and highlighted results from multiple sources is an important technical challenge that needs to be addressed.**

4. Exascale Challenges - Visualization and data analysis approaches will need to run efficiently on exascale platform architectures

Defense Programs

- Need to take advantage of a very high degree of parallelism
- Technical challenges include achieving portability, efficiency and integration flexibility with simulation codes

1. Path Forward - New visualization and data analysis software infrastructure

- When?: Run-time, Postprocessing
- How?: Interactive, Batch

- **In-situ analysis within the simulation code**
  - Run-time, (batch or interactive)
- Post-processing --- advanced query-based approach

- **Revolutionary approach**
  - **Phase 1**
    - **Prototype approaches in applications**
  - **Phase 2&3**
    - **Develop and deploy**
    - **Continue R&D**

- **Required Partnership**
  - **IONS, tools, systems**
  - **Apps for co-design**
- **Metric**
  - **Our success will be measured by our readiness for applications as machine delivery milestones are met**

- **Risks**
  - **How to do discovery science in an in-situ world?**
  - **Won't find an effective analysis approach for exascale applications**

2. Advanced quantifiable data reduction algorithms

- Data triage
  - How do we significantly reduce the data as it is generated?
    - Statistical sampling
    - Compression
    - Multi-resolution
    - Science-based feature extraction
- Revolutionary approach
  - Phase 1
    - Prototype approaches independently and with applications
  - Phase 2&3
    - Develop and deploy
    - Continue R&D
- Required Partnership
  - Applied math
  - Apps for co-design
- Metric
  - Measure of amount of data reduced and quality of result, time
- Risks
  - Won't find an effective analysis approach for exascale applications

3. Visualization and data analysis techniques to help understand advanced exascale physics

Defense Programs

- Visualization and Data Analysis for:
  - Statistical physics approaches
  - Parametric studies
  - Multi-physics approaches
  - Multi-scale approaches
- how results from different aspects of a simulation suite relate to each other

- Evolutionary/Revolutionary approach
  - Phase 1
    - Prototype approaches independently and with applications
  - Phase 2&3
    - Develop and deploy
    - Continue R&D

- Required Partnership
  - Applications for co-design
- Metric
  - Ties to appropriate application milestones

- Risks
  - Won't understand output of exascale applications

4. Implement core visualization and data-analysis capability using a scalable parallel infrastructure

- – Our visualization and data analysis solutions need to work on the exascale supercomputers on both swim lanes.
- Evolutionary approach
  - Phase 1
    - Prototype approaches independently and with applications
  - Phase 2&3
    - Develop and deploy
    - Continue R&D
- Required Partnership
  - Programming models, tools and applications groups
- Metric
  - Our success will be measured by our readiness for applications as machine delivery milestones are met
- Risks
  - Not running on the machines

5. Exascale visualization and data analysis hardware infrastructure

Defense Programs

- **Data-intensive hardware infrastructure for the exascale platform**
  - Memory buffers for staged analysis and storage
  - Analysis-enabled storage
  - Large node memory portion of the supercomputer
- **Evolutionary/Revolutionary approach**
  - Phase 1
    - Prototype approaches independently and with applications
  - Phase 2&3
    - Develop and deploy
    - Continue R&D
- **Required Partnership**
  - HW, Systems, I/O
- **Metric**
  - Ties to appropriate machine milestones
- **Risks**
  - HW platform that makes data analysis difficult

6. Tracking and using knowledge about the scientific goals makes visualization and data analysis more effective

Defense Programs

- Provenance
  - Where the data comes from?
  - How was it generated?
  - Uncertainty quantification
  - Workflow management and beyond
- Monitoring, debugging
  - Supercomputer “situational awareness”
- Revolutionary approach
  - Phase 1
    - Prototype approaches independently and with applications
  - Phase 2&3
    - Develop and deploy
    - Continue R&D
- Required Partnership
  - Tools, Systems, Applications for co-design
- Metric
  - Time to solution, cost, improved quality
- Risks
  - Won’t know where data came from
  - Inefficient use of supercomputer

Recommended Co-Design Strategy

- **Critical steps/activities**
  - **Data Analysis and Movement are critical cross-cutting issues**
  - **Develop well-defined HW needs for VDA**
    - **In-situ and interactive data extraction could benefit from co-design**
      - **(e.g. resource management, time series analysis)**
  - **Pilot co-design projects to help develop teams and define co-design processes**
    - **These can be subsets – VDA and IO working with a defined app**
- **Working with vendors**
  - **Well-defined communication with partners and vendors**
  - **Visualization software vendors (e.g. Kitware, CEI), platform vendors**
  - **Path Forward investments**

Recommended Co-Design Strategy

Defense Programs

- **Role of skeleton/compact apps**
  - **VDA mini-apps**
    - **Investigate coupling of in-situ capabilities with applications at scale**
  - **Inform Architecture decisions by providing information on use patterns**
- **Concerns/suggestions**
  - **Organization effects outcomes**
    - **Partnering with Verification/Validation, applied math**
    - **Overlap with other groups**
  - **Influence we can have on HW vendors w/in timeframe**
    - **Particularly with respect to data**
  - **Co-design should have significant investment and sustained commitment**

Big Picture Issues

Defense Programs

- **Coordination**
  - **VDA will be more tightly coupled with apps than in the past**
  - **Data movement and provenance are shared responsibilities**
- **Test beds**
  - **Common test beds crucial to creating Exascale community**
  - **Specialized test beds may be necessary**
- **Simulators**
  - **Requirement: include adequate characterization of data movement**
    - **Unclear if there are requirements for VDA-only simulation components**
  - **Requirement: system-level simulation, to model various VDA use cases**
- **Remaining gaps**
  - **Interactive data analysis (e.g. querying and extraction) for Exascale is coupled with in-situ (what you can compute at runtime) and computation on data (what you want to understand after the simulation)**

Conclusions

A data-oriented perspective is critical to exascale success

End