IN-SITU VIS IN THE AGE OF HPC+AI

Peter Messmer, WOIV - ISC 2018, Frankfurt, 6/28/18
EXCITING TIMES FOR VISUALIZATION

Huge dataset everywhere

Novel workflows for HPC * AI

Rendering technology leap
DEEP LEARNING COMES TO HPC
Accelerates Scientific Discovery

UIUC & NCSA: ASTROPHYSICS
5,000X LIGO Signal Processing

U. FLORIDA & UNC: DRUG DISCOVERY
300,000X Molecular Energetics Prediction

LBNL: High Energy Particle PHYSICS
100,000X faster simulated output by GAN

PRINCETON & ITER: CLEAN ENERGY
50% Higher Accuracy for Fusion Sustainment

U.S. DoE: PARTICLE PHYSICS
33% More Accurate Neutrino Detection

Univ. of Illinois: DRUG DISCOVERY
2X Accuracy for Protein Structure Prediction
The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka!” but “That’s funny ...”

— Isaac Asimov
ANOMALY DETECTION
Look at all the data and learn from it!

Vast amount of data produced by simulations and experiments

I/O limitations force throw-away

Human in the loop infeasible

“In-situ vis 2.0”

Dataset: Visualization of historic cyclones from JWTC hurricane report from 1979 to 2016

Kim et al, 2017

Run inference on the simulation data to detect anomalies
ANOMALY DETECTION IN CLIMATE DATA

Identifying “extreme” weather events in multi-decadal datasets with 5-layered Convolutional Neural Network. Reaching 99.98% of detection accuracy. (Kim et al, 2017)

Dataset: Visualization of historic cyclones from JWTC hurricane report from 1979 to 2016

MNIST structure

Systemic framework for detection and localization of extreme climate event
OPTIMIZING PARAMETERIZED MODELS
Learning from simulations to accelerate simulations

Heuristics and semi-empirical models throughout simulation codes

Often expensive, complex control flow

Use trained approximate models instead

Trained with past simulations

Examples:

- **Parameterized models:**
  - Atmospheric radiation transport, collisional cross-sections, chemical reaction chains, ...

- **Simulation parameters**
  - Grid refinement, load balancing, scheduling, ...

Use simulations to train approximate models
HPC SIMULATION OF THE FUTURE

Concurrent simulation, training, inference

Leads to
- Improved approximate models
- Better utilization of data
- Faster results

Coupling of HPC and DL software stack is needed
DATA SCIENCE DRIVES SOFTWARE-STACK

Data science mission-critical to non-traditional HPC organizations

Deep learning, graph analytics, in-core databases, ...

Sustainability and performance by scale

Frameworks supported by big corps, large communities

Big market, big support by all vendors

⇒ Economics drive performance portability and sustainability
⇒ Happened in the past, but different situation today

Cast HPC algorithms in DL terms

Trilinos: 274 stars, PyTorch: 16’615 stars
EXAMPLE: WAVE EQUATION VIA CONV NEURAL NETWORK

Applying stencils = Inference in Conv Network

\[
dv \frac{dt}{dt} = c \Delta u
\]

\[
du \frac{dt}{dt} = v
\]

\[
\Delta u_0 = u_{x-1} - 2u + u_{x+1}
\]

\[
V_{3/2} = c dt \Delta u_0 + v_{1/2}
\]

\[
u_1 = dt v_{1/2} + u_0
\]

Simplified 1D example cartoon
Approach works for higher dimensions

Other examples: Stencils, Spectral transforms, spectral elements, ..
EXAMPLE: MARCHING CUBES

(well.. squares)
HPC CAN CONTRIBUTE TO EMERGING DATA SCIENCE NEEDS

HPC solved lots of “new” problems in the past

DL will need distributed memory parallelism

New challenges for DL algorithms

HPC has probably hit those challenges in the past

Better implementation, better algorithms

Collaborate with DL framework developers, contribute to DL frameworks

First step: speak a common language

https://www.hpcwire.com/2017/02/21/hpc-technique-benefits-deep-learning/
RENDERING TECHNOLOGIES
VISUALIZATION IN THE DATACENTER

Benefits of Rendering on Supercomputer

Scale with Simulation
No Need to Scale Separate Vis Cluster

Cheaper Infrastructure
All Heavy Lifting Performed on the Server

Interactive High-Fidelity Rendering
Improves Perception and Scientific Insight
VISUALIZATION REQUIREMENTS

Accuracy

Responsiveness

Interpretability
REAL-TIME RAY-TRACING
Enabled by Volta V100

OptiX Ray-Tracing SDK
State-of-the-art performance: 500M+ rays/sec
Algorithm and hardware agnostic
Shaders with single-ray programming model, recursion

Free for private and commercial use
https://developer.nvidia.com/optix
OptiX 5 with the CNN-based AI Denoiser

“Interactive Reconstruction of Monte Carlo Image Sequences Using a Recurrent Denoising Autoencoder.”
SIGGRAPH 2017
OPTIX FOR SCIENCE
Ray-Tracing for extra visual cues

Enhanced visual perception
Single workflow for science and outreach
Fast rendering with AI denoiser
Integrated into leading vis tools

VMD
VTK
ParaView
WILL HPC TRANSFORM AI OR WILL AI TRANSFORM HPC?

Combination of HPC and AI will lead to better exploitation of data, faster models

Both communities need each other -
AI will need distributed memory technologies at one point
HPC needs AI for analysis and parameterized models

Economic interest in AI drives software and hardware development

Expressing HPC algorithms in AI terms helps leveraging AI features

Applying HPC techniques to scale AI

=> Continue with joint events!

Simulation and streamline data courtesy of Christoph Garth
OptiX AI Denoiser in ParaView

Without Denoiser

With Denoiser
NVIDIA IndeX SDK

Large scale and distributed data rendering
Scene management with volume data
Transparent support for NVLink
Higher-order filtering, advanced lighting & transfer functions

https://developer.nvidia.com/index
ParaView Plugin

NVIDIA IndeX ParaView Plugin

Distributed with ParaView (pvNvidiaIndeX)
Mixing with ParaView primitives
Support for multiple slice planes
Structured and unstructured meshes
In-situ capable via Catalyst

https://developer.nvidia.com/index
NVIDIA INDEX PARAVIEW PLUGIN
Growing user base

- Single GPU version bundled with ParaView
  - Over 20,000 downloads in 2 months.
- Cluster version
  - 10+ labs and institutions actively working with us.

# NVIDIA INDEX PARAVIEW PLUGIN

## Licensing and Availability

<table>
<thead>
<tr>
<th>SINGLE GPU</th>
<th>ACADEMIC</th>
<th>COMMERCIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free license</td>
<td>Free license</td>
<td>License fee</td>
</tr>
<tr>
<td>Supports single GPU</td>
<td>Multi-node Multi-GPU</td>
<td>Multi-node Multi-GPU</td>
</tr>
<tr>
<td>Plugin binaries part of ParaView v5.5 download</td>
<td>Plugin source part of ParaView v5.5 download</td>
<td>Plugin source with customer specific features</td>
</tr>
<tr>
<td>New releases synced with ParaView release cycle</td>
<td>Latest features in ParaView-Master</td>
<td>Premium support</td>
</tr>
</tbody>
</table>
FLEXIBLE GPU ACCELERATION ARCHITECTURE

Independent CUDA Cores & Video Engines

Decode HW*

Formats:
- MPEG-2
- VC1
- VP8
- VP9
- H.264
- H.265
- Lossless

Bit depth:
- 8 bit
- 10/12 bit

Color**
- YUV 4:2:0

Resolution
- Up to 8K***

* See support diagram for previous NVIDIA HW generations

** 4:2:2 is not natively supported on HW

*** Support is codec dependent

Encode HW*

Formats:
- H.264
- H.265
- Lossless

Bit depth:
- 8 bit
- 10 bit

Color**
- YUV 4:4:4
- YUV 4:2:0

Resolution
- Up to 8K***
VIDEO CODEC SDK

Hardware Accelerated Video Encoding/Decoding

Encode and Decode API

Industry-standard codecs: H.265(HEVC), H.264(AVCHD), VP9, VP8, VC-1 & MPEG-2

Supports up to 8K x 8K encoding

Lossy & lossless compression

Streaming of Large Tile Counts

Frame rates / Latency / Bandwidth

Synchronization

Comparison against CPU-based Compressors

Strong Scaling (Direct-Send Sort-First Compositing)

Hardware-Accelerated Multi-Tile Streaming for Realtime Remote Visualization, Biedert et al, EGPGV 2018, https://doi.org/10.2312/pgv.20181093
CONCEPTUAL OVERVIEW

Asynchronous Pipelines

HPC System

Clients

Render
Buffer (Device)
Encode
Buffer (Host)
Send
Receive
Buffer (Host)
Decode
Buffer (Device)
Display

Tiled Display
VR
BENCHMARK SCENES

NASA Synthesis 4K

Space
Low Complexity

Orbit
Medium Complexity

Ice
High Complexity

Streamlines
Extreme Complexity
HARDWARE LATENCY

Decreases with Resolution

![Graph showing the relationship between resolution and latency for HEVC and H.264 encoding and decoding. The graph indicates that latency decreases as resolution increases.]
**N:N STREAMING**

**Pipeline Latencies**

- **Server:** Piz Daint
- **Client:** Piz Daint
- Ice 4K
- **H.264:** 32 Mbps
- **MPI-based synchronization**

---

**Diagram Details:**
- **Tiles:** 1, 2, 4, 8, 16, 32, 64, 128, 256
- **Latency [ms]:**
  - **Synchronize (Servers):** Various levels
  - **Encode:** Various levels
  - **Network:** Various levels
  - **Decode:** Various levels
  - **Synchronize (Clients):** Various levels

---

**Legend:**
- Blue: Synchronize (Servers)
- Orange: Encode
- Grey: Network
- Yellow: Decode
- Blue: Synchronize (Clients)
N:N STREAMING
Client-Side Frame Rates and Bandwidths

Server: Piz Daint
Client: Piz Daint
Ice 4K
H.264: 32 Mbps
HEVC: 16 Mbps
**N:1 STREAMING**

Client-Side Frame Rate

Server: Piz Daint
Clients: Site A (5 ms)  
         Site B (25 ms)

Ice 4K  
H.264: 32 Mbps
N:1 STRONG SCALING

Client-Side Frame Rate

Server: Piz Daint
Clients: Site A (5 ms)
         Site B (25 ms)

Ice 4K
H.264: 32 Mbps

Frame Rate [Hz]

Tiles

Site A (1x GP100) Site B (1x GP100) Site B (2x GP100)
EXAMPLE: OPTIX PATHTRACER

Rendering Highly Sensitive to Tile Size

Server: Piz Daint
Client: Site A (5 ms)
H.264

Frame Rate [Hz]
Tiles

0
10
20
30
1 2 4 8 16 32 64 128 256

Regular Tiling
Auto-Tuned Tiling
STANDARD-COMPLIANT BITSTREAM
Web Browser Streaming Example

EGL-based shim GLUT

Stream unmodified simpleGL example from headless node to web browser (with interaction!)

JavaScript client

WebSocket-based bidirectional communication

On-the-fly MP4 wrapping of H.264
In Situ Volume Compression using NVENC

Use NVENC + VPU (Titan and later) to compress volume data

In situ encoding runs overlapped with computation: essentially “free”.

At least 100:1 compression (over 1000:1 in many cases)

Similar application of NvPipe for EM data at Cvlab at EPFL

N. Leaf, B Miller and K-L Ma. In Situ Video Encoding of Floating-Point Volume Data Using Special-Purpose Hardware for a Posteriori Rendering and Analysis. IEEE LDAV 2017
CONTAINERS
NVIDIA GPU CLOUD FOR HPC VISUALIZATION

ngc.nvidia.com
CLIENT-SERVER CONTAINERS

Client Container

User Interface
Pre-packaged Plugins
GPU accelerated Encoding
GLX and EGL

Server-Side Container
NVIDIA GPU Cloud Containers

ParaView with NVIDIA OptiX

Provides GPU accelerated ray-tracing technology within ParaView offering enhanced visual cues and high performance rendering for large scale scenes.

NOTE: The HPC visualization containers require NVIDIA Docker 2.0 to be installed on your system. For installation instructions, see NVIDIA Docker 2.0 github.

For general information on HPC visualization containers, configuration of the client and server container in both GLX and EGL configuration, see the NGC User's guide.

See here for a document describing the steps to pull NGC containers.

1. Single-Node Containers with GLX

<table>
<thead>
<tr>
<th>TAG</th>
<th>SIZE</th>
<th>USER</th>
<th>LAST MODIFIED</th>
<th>PULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>glx-17.11.13-beta</td>
<td>2.44 GB</td>
<td><a href="mailto:camm@nvidia.com">camm@nvidia.com</a></td>
<td>April 3, 2018</td>
<td></td>
</tr>
<tr>
<td>egi-17.11.13-beta</td>
<td>1.86 GB</td>
<td><a href="mailto:camm@nvidia.com">camm@nvidia.com</a></td>
<td>November 21, 2017</td>
<td></td>
</tr>
</tbody>
</table>

http://ngc.nvidia.com
LEADING EDGE CONTAINERS

Novel features

Latest version of ParaView IndeX plugin

ParaView+OptiX container, including AI denoiser and OptiX raytracing

Rapid release cycle