Using HPC Computational Physics Tools for Advanced Engineering Simulations and Production Deployment

David Hinz
Sr. Director, IT Engineering Services / HPC

Byron Lengsfield III
Technologist, Recording Subsystems

March 26, 2018
Disclaimer

• This presentation may contain product features that are currently under development.

• This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.

• Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.

• Technical feasibility and market demand will affect final delivery.

• Pricing and packaging for any new technologies or features discussed or presented have not been determined.
Agenda

• Western Digital’s History with AWS and GPUs for HPC

• Current Uses of AWS GPUs

• Recommendations for Success
Western Digital: A Heritage of Innovation

14,000+ active patents worldwide

1956 | IBM
Invents First Hard Drive (RAMAC)

1976 | WD
First Disk Array Sub-System Patent

1988 | SunDisk
Introduces Flash as a Storage Medium

1991 | IBM
Introduces MR Heads

1995 | M-Systems
First Flash Drive

2000 | IBM & M-Systems
First USB Flash Drive

2000 | SanDisk
First MLC NAND

2006 | HGST
First 2.5-inch HDD PMR Technology

2008 | G-Technology
First 1TB Portable External HDD

2009 | SanDisk
First 64GB X4 Flash memory

2009 | Fusion-io
First Enterprise PCIe Flash Accelerators

2013 | HGST
First Helium HDD (6TB)

2013 | HGST
First 12Gb/s SAS SSD

2015 | SanDisk
First 48-layer 3D NAND Technology

2016 | SanDisk
First 64-layer 3D NAND Technology

2017 | SanDisk
First 96-layer 3D NAND Technology

1988 | SunDisk
Introduces Flash as a Storage Medium

2006 | HGST
First 2.5-inch HDD PMR Technology

2009 | Fusion-io
First Enterprise PCIe Flash Accelerators

2009 | Fusion-io
First 12Gb/s SAS SSD

2013 | HGST
First Helium HDD (6TB)

2015 | SanDisk
First 48-layer 3D NAND Technology

2016 | SanDisk
First 64-layer 3D NAND Technology

2017 | SanDisk
First 96-layer 3D NAND Technology
Portfolio of Leading Brands

Western Digital

G-Technology

HGST

tegile

SanDisk

upthere

©2018 Western Digital Corporation or its affiliates. All rights reserved.
Leading Portfolio Breadth and Depth
Solutions to Capture, Preserve, Access and Transform Data

Platforms/Systems/Solutions

Client Devices

Client Solutions

Data Center Devices & Solutions

Devices

Technologies

Heads

Media

Device Software

System Software

RISC-V

©2018 Western Digital Corporation or its affiliates. All rights reserved.
Fueling the Next Decade of Big Data

New technology will enable 40TB by 2025 and even higher beyond
Western Digital Cloud HPC Architecture

• Dedicated AWS account for HPC
  – DirectConnect = Main
  – VPN = Backup

• Clusters for the usage
  – By Organizations
  – By Applications
  – By Users

• Autoscaling Cluster
  – Autoscale based on the user queue
  – CycleCloud by CycleComputing
  – Master node = Job Scheduler + NFS
  – Instance Store for NFS = i3 instance
  – Optimal Instance Type for the workload = c4, r4, p2
Shape Compute To Match Work To Be Done

Molecular Dynamics
1.67x Initial Overall Throughput Gain

Fluid Dynamics
1.4x Overall Throughput Gain

### Molecular Dynamics

<table>
<thead>
<tr>
<th>Simulation Type</th>
<th>Throughput Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Drive Interface Vacuum Gaps</td>
<td>1.99x</td>
</tr>
<tr>
<td>Vacuum Gap &quot;collection&quot;</td>
<td>4.00x</td>
</tr>
<tr>
<td>Media Grains for HAMR (FePt/C)</td>
<td>2.03x</td>
</tr>
<tr>
<td>4 Carbon Molecule Clusters</td>
<td>5.67x</td>
</tr>
</tbody>
</table>

### MicroMagnetics

<table>
<thead>
<tr>
<th>Parameter Sweeps</th>
<th>Throughput Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1.23x - 1.78x</td>
</tr>
<tr>
<td>Model 2</td>
<td>1.01x – 1.67x</td>
</tr>
<tr>
<td>Model 3</td>
<td>1.23x – 1.69x</td>
</tr>
<tr>
<td>Model 4</td>
<td>Up to 2.7x</td>
</tr>
</tbody>
</table>
Western Digital’s Innovations + the Cloud Advantage = Faster Business Value

HGST runs worlds largest F500 cloud HPC simulation run
Reduces simulation time from 30 days to 8 hours!!!
Western Digital : AWS and GPUs for HPC

• Western Digital has been a partner with AWS to understand workloads and use cases of HPC GPU computing for 3+ years

• Western Digital tested AWS GPU products and provided detailed performance benchmarks

• Western Digital directly engaged GPU suppliers directly to highlight use cases and requirements

• Actively using AWS GPUs for Machine Learning / Artificial Intelligence and High Performance Computing
HDD Recording System Modeling: Outline

- Modeling Work Flow
- GPU/AWS Realization of the Work Flow
- Model Details
  - Parallelization yields large gains
  - Larger problems: 511 bit patterns, which allows to use a common set of signal processing tools in experiment and in the modeling
  - Optimization in a large parameter space
  - Describe recording physics at a smaller length scale
- Conclusions
  - Seamless migration to AWS
  - Realization of the capabilities of a new generation of GPUs
1. **Write step**
2. **Read step**
3. **Signal Processing Step**

- **Input**: 2-level write current
- **Output**: Noisy readback waveform
- **Goal**: Reliably store data at highest possible areal density

**Tracks of data written on the disk**

---

**Hard Disk Drive (HDD) Magnetic Recording**

\[
\ldots 0 0 0 1 1 1 1 0 1 1 1 1 0 1 0 0 0 \ldots
\]

Write Head \quad \rightarrow \quad Read Head

Write step
Read step
Signal Processing Step

Head Motion

Read Sensor
Work Flow

Current waveforms input to head model to produce write field in the media as a function of time (1500, 3D head fields)

Multilayer micromagnetic media models used to generate 400 repeats of Pseudo-Random Bit Pattern (PRBS) which is then written by the write field, read and fed into a SWC for BER and other metrics
Fields generated from FEM or Micromagnetic Writer Model

Write Pole

Soft Under Layer (SUL)

Wrap-Around-Shield (WAS)

trailing portion of WAS

Many critical dimensions and angles

Head-Under-Layer Spacing (HUS)

Gap between write pole and trailing portion of WAS

Write field magnitude and angle varies in all three dimensions

Magnetic field produced in each media layer

FEM = finite-element-method, magnetostatic model
**Readback Process**

**3D Reader Sensitivity Function (RSF):**
Convolution of the RSF and magnetization

1. Analytic or micromagnetic program to generate RSF
2. Readback is very fast
   - All crosstrack positions are obtained at one time using 2D FFT
   - Post-processing step, all magnetization data is on a separate server.
   - Convolve 3D RSF with the magnetization.
3. Full micromagnetic readback is available but more compute intensive

**Different RSF for each media layer**
Multi-layer Media Model: Exchange between layers and grains

Top surface of the media

Magnetization aligned perpendicular to the surface of the media along anisotropy axis $|k>$.  

$\mathbf{E_a} = K <m|k>^2$

Macro-spin Model (each layer is a macro-spin)

- Long range magnetic field computed using FFT CUDA library
- Exchange interaction is parallelized over 3D regions
Linux Scripts Generate Input from Templates which are Submitted to the Batch System
- The input template contains hardwired input values and variables which are specified in the Submit Script.

**Nvidia GPUs:**
*Coarse Grain Parallel*

- Submit scripts generate an ensemble of inputs and on the master node
- The batch system distributes the command files on the compute nodes.
- Output is moved to the master node

**Master Node**

**CPU**

**GPU**

**Master Node**

**Post-Processing Node**

**Processing step (readback) to generate BER, SNR, ADC,...**

**AWS**
GPU code implementation

• Example: Anisotropy field (one body: \( \langle M|K \rangle |K\rangle \) field term)
  ✓ Grain-wise anisotropy fields computation are assigned to different cores on a GPU
  ✓ Each layer in a grain runs independently

• CUDA Random Numbers for Thermal Field
• CUDA FFT for Magnetic Field
• Parallel 3D Group Exchange

Media
Analysis of Small Bit Pattern ~ 620 Runs Needed

- Jitter noise power is dominant
- Non-transition noise is pattern dependent
  - increases as length of the bit increases
  - noise can be seen in average magnetization and smearing in autocorrelation matrix

**Noise Terms**

1. Jitter from STD of the transition position
2. DCsnr from STD of signal at bit center
3. Head noise: HEsnr from amplitude and head noise
   - Size of smallest signal relative to head noise
4. T50 (signal width) from the readback

**Run Time**
- 2-3 hours per CPU core
- 1-3 minutes on a GPU (K80, Single Precision)
Windowing Algorithm for 511 bit PRBS

- Cuda FFT is used in each Window
- Head moves in the Window
- Then Window moves
  - Head is always near the center
  - Problem re-started in each Window

Motion of the Head
Local iteration

Motion of the Window
Global iteration

Triple track 511 bit PRBS now practical with GPUs
Discretized Media Model

GPUs give us the ability to simulate recording physics at a smaller length scale

Many variables:
- $J_{sii}$ is the surface exchange between cells in the internal regions
- $J_{isis}/J_{sisi}$ is the surface exchange between cells at the boundary of the internal and shell regions
- $J_{s}$ is the exchange between boundary cells
- $J_{sb}$ is the exchange between the shell and the magnetic boundary (lateral exchange)

Grain properties are partition into regions

‘7-Layer’ Model Capped-Soft-Hard Media

- ECL (exchange coupling layer)
- thickness and vertical exchange are varied
- The limit of “metallic” vertical exchange between the layers can be studied (thin ECL and ~“metallic” vertical exchange)
- Thus this Model can represent a range of physical (sputtered) layers

Magnetization in each cell is independently propagated in the LLG+T simulation
Macro-spin model vs. Discretized models

- Closure domain formation effects resolution performance
  - Independent spins in the discretized model can describe the gradual change of magnetization in the cap (top) layer
  - Closure domains can form in the discretized model
  - Closure domains lead to poorer resolution than found in macro-spin model in agreement with experiment

In plane magnetization shows the presence or absence of closure domains
Improved Domain Wall (DW) Media Performance Under Squeeze:

- Optimization algorithms are implemented at the script level (outer loop)
- DW ‘optimization’ yielded much improved squeeze performance

GPUs make multi-dimensional optimization practical in these systems
Technical Conclusions

Large throughput gains seen and new capabilities realized by moving models to GPUs

- Enhanced performance
  - Rapid turn-around
  - Optimization in a larger parameter space
  - Minimize CPU-GPU communications

- Ability to handle larger problems and use the same tools employed in experiments
- Ability to treat recording physics on a smaller length scale

AWS Realization

- Seamless migration from existing GPU clusters
- Realization of the capabilities of a new generation of GPUs: we saw a large improvement in performance in going from K80s to P3s
Recommendations For Success

• Workload Expertise is Imperative
  – All workloads are not the same relative to GPU vs. CPU dependence and parallelization

• Complete Proof of Concepts / Trials Before Committing Large Deployments
  – 30 day PoCs allow different configurations and workflows to be evaluated

• Work with AWS and NVidia if the results are not meeting expectations
  – Remove blocking problems through strong partnerships that are developed in meetings not in email exchanges
Realistic Media Distributions

Exchange distribution

Bin 2D distribution: $(\sigma_K, \sigma_V)$ matrix

Mean Field Exchange Strength and Volume Dependence

- Large number of parameters require many simulations to converge Signal-to-Noise (SNR) and other performance metrics.

- Typically 600 runs to converge transition position jitter and estimate Areal Density Capability (ADC).

- Ensemble members are run in parallel.
Noise Sources

1. Transition position jitter
2. DCsnr: noise at the center of the bit
   ✓ poor writing
   ✓ erasure due to writing an adjacent track
3.Electronic noise
   ✓ Poor resolution
   ✓ Head and Electronics noise (HEsnr)

Average Readback Signal

Noise in ensemble members

Good Resolution

Poor Resolution
Media Model: Discretization

Media Model Resolves
Grain Boundaries ~ 1 nm

Macro-spin model:
Each layer is a spin
Many opportunities for parallelization

1. One-body terms (Anisotropy)
2. Two-body terms (Exchange in 3D)
3. Long Range magnetic interactions (CUDA FFT)
4. Stochastic (Thermal) Field (CUDA Random numbers)

Magnetization propagates via LLG + Thermal Field equation

\[
\frac{dM}{dt} = -\left(\frac{\gamma}{(1+\alpha^2)}\right) M \times \text{Heff} + = -\left(\frac{\gamma}{((1+\alpha^2)M_s)}\right) M \times (M \times \text{Heff})
\]

\(\alpha = \text{damping constant}\)
\(\gamma = \text{gyromagnetic ratio}\)