Western Digital®

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

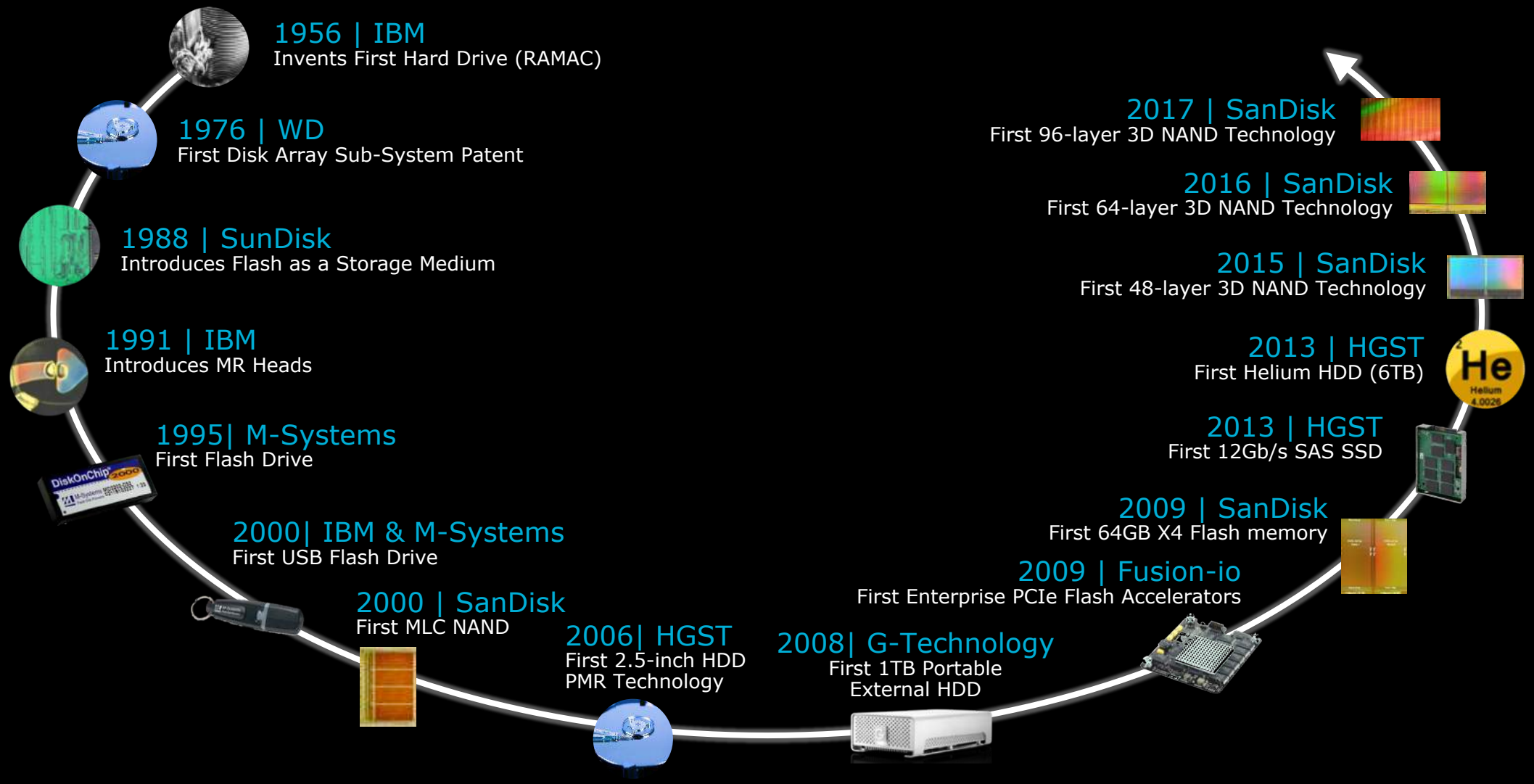
- This presentation may contain product features that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase 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



Portfolio of Leading Brands

Western Digital®

WD

HGST

tegile

G-Technology™

SanDisk®

upthere

Leading Portfolio Breadth and Depth

Solutions to Capture, Preserve, Access and Transform Data

	Client Devices	Client Solutions	Data Center Devices & Solutions
Platforms/Systems/ Solutions			
Devices			

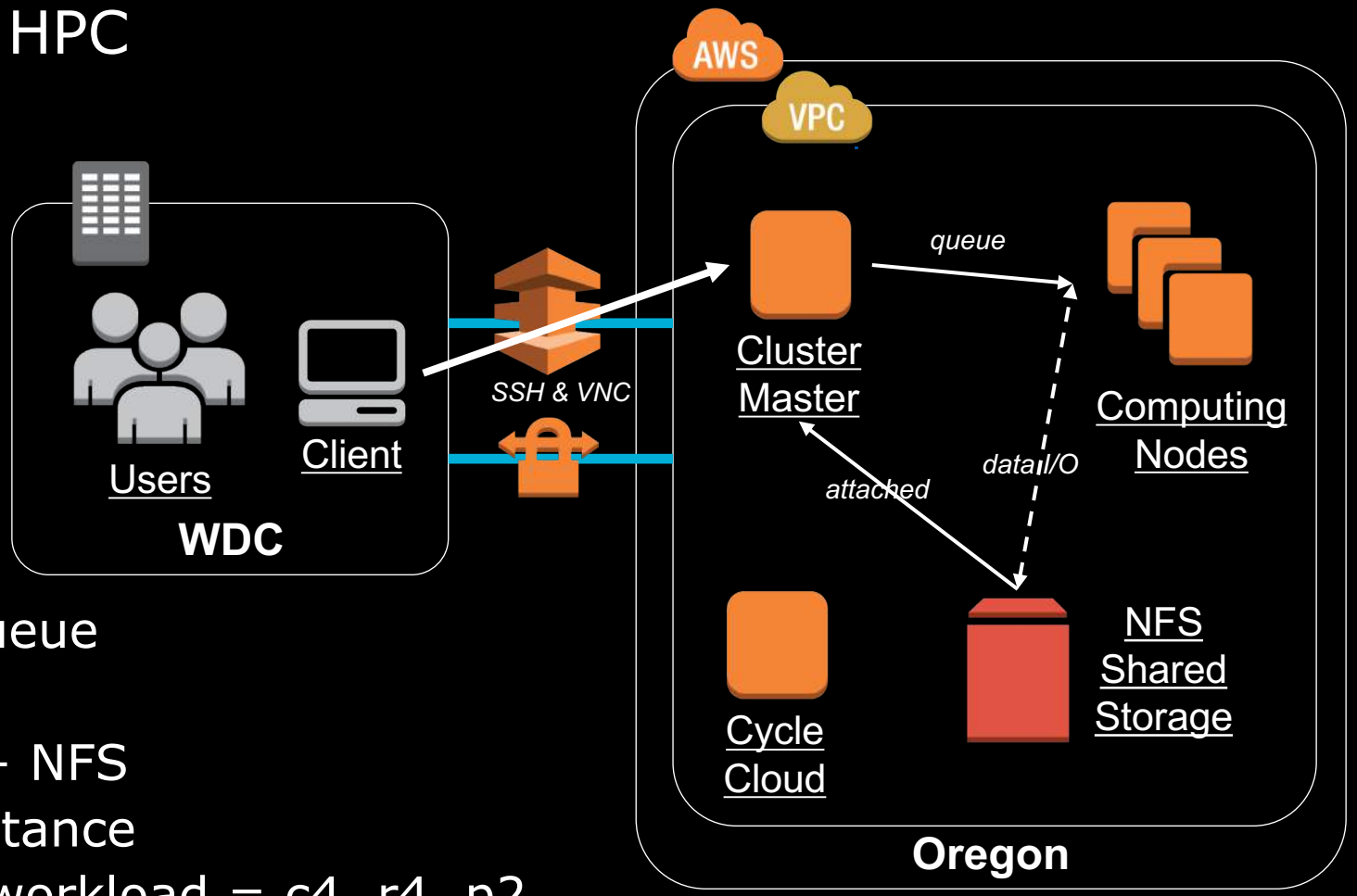
Technologies	Heads	Media	Controllers	Device Software	System Software	RISC-V

Fueling the Next Decade of Big Data

New technology will enable 40TB by 2025 and even higher beyond

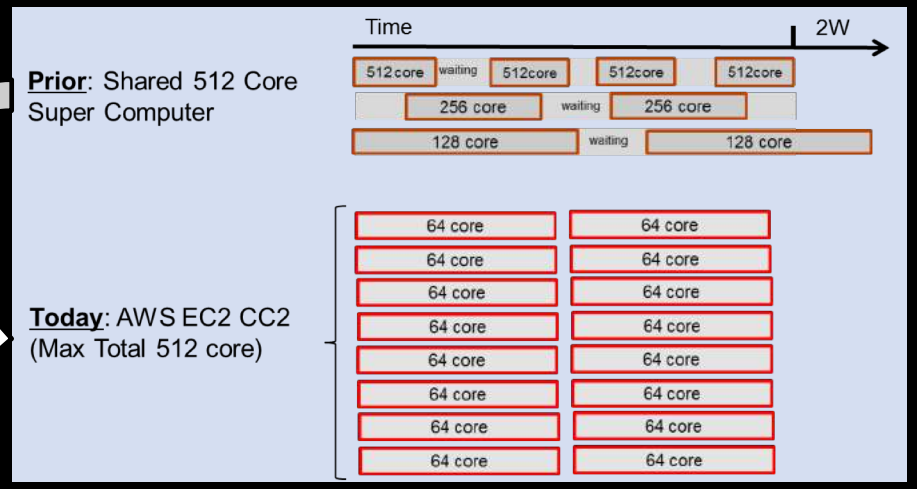
Western Digital Cloud HPC Architecture

- 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

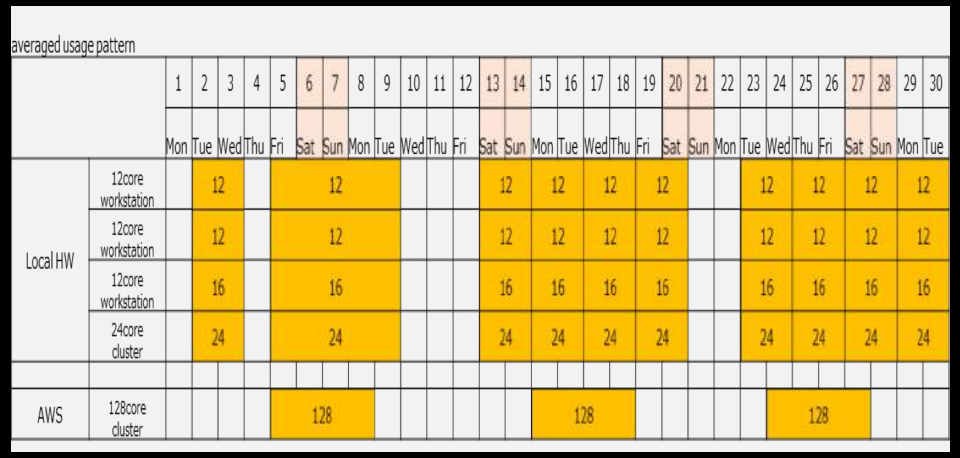


Shape Compute To Match Work To Be Done

Molecular Dynamics
1.67x Initial Overall Throughput Gain



Fluid Dynamics
1.4x Overall Throughput Gain



Molecular Dynamics

Simulation Type	Throughput Increase
Head Drive Interface Vacuum Gaps	1.99x
Vacuum Gap "collection"	4.00x
Media Grains for HAMR (FePt/C)	2.03x
4 Carbon Molecule Clusters	5.67x

MicroMagnetics

Parameter Sweeps	Throughput Gain
Model 1	1.23x - 1.78x
Model 2	1.01x – 1.67x
Model 3	1.23x – 1.69x
Model 4	Up to 2.7x

Western Digital's Innovations + the Cloud Advantage = Faster Business Value

HDD INNOVATIONS

Arial Density

6TB

4TB

Mechanical

Number of Disks

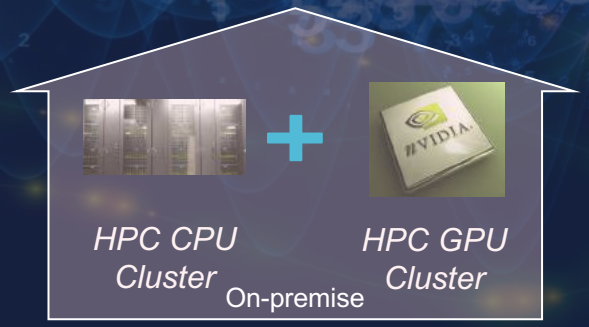
Air/Helium

Speed

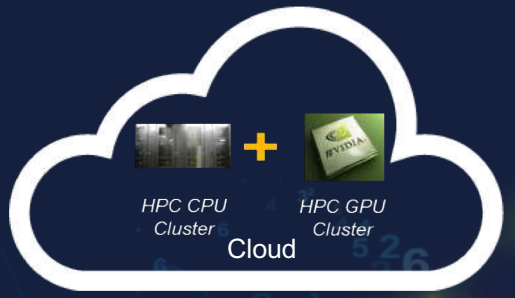
Lubes

HPC Key to Enabling and Driving Innovation

High Performance Computing (HPC)



+



CLOUD ADVANTAGE

Scalability

85,000+ Cores

3+ Data Centers

+6,000 Servers

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

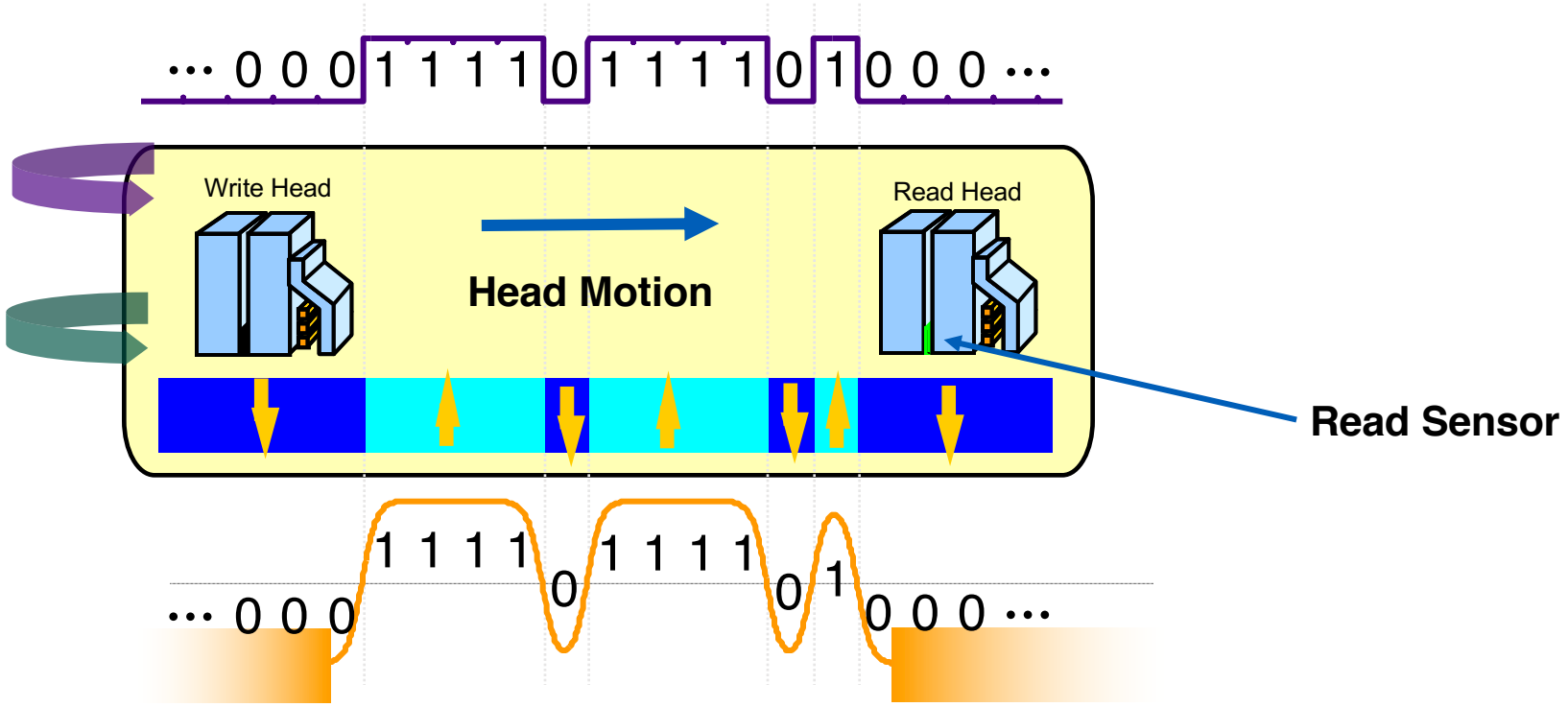
- Western Digital has been a partner with AWS to understand workloads and use cases of HPC GPU computing for 3+ years
- Western Digital tested AWS GPU products and provided detailed performance benchmarks
- Western Digital directly engaged GPU suppliers directly to highlight use cases and requirements
- Actively using AWS GPUs for Machine Learning / Artificial Intelligence and High Performance Computing

HDD Recording System Modeling: Outline

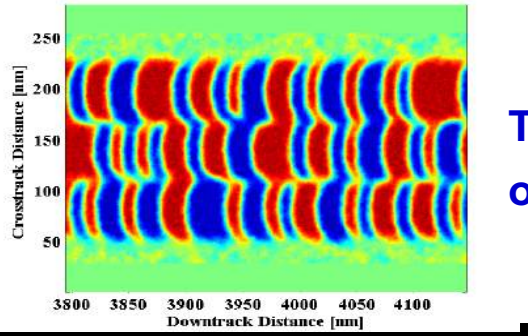
- **Modeling Work Flow**
- **GPU/AWS Realization of the Work Flow**
- **Model Details**
  - **Parallelization yields large gains**
  - **Larger problems: 511 bit patterns, which allows to use a common set of signal processing tools in experiment and in the modeling**
  - **Optimization in a large parameter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

Hard Disk Drive (HDD) Magnetic Recording

1. **Write step**
2. **Read step**
3. **Signal Processing Step**



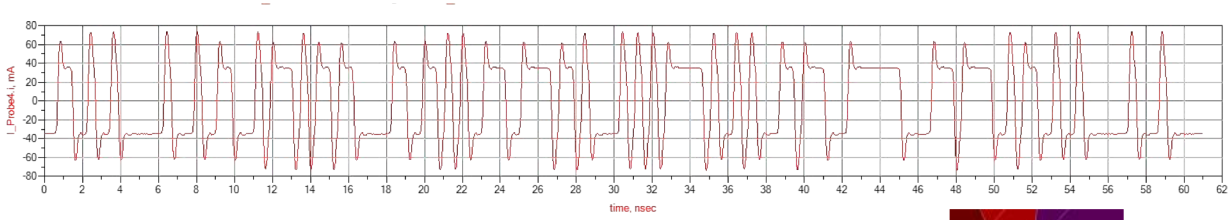
- **Input**: 2-level write current
- **Output**: Noisy readback waveform
- **Goal**: Reliably store data at highest possible areal density



Tracks of data written on the disk

Work Flow

Write Current Waveforms

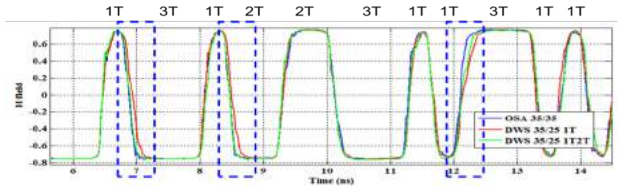


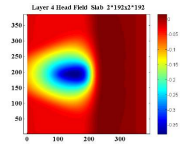
Current waveforms input to head model to produce write field in the media as a function of time (1500, 3D head fields)

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

Writer Model

Micromagnetic or FEM Writer Model for Head Fields



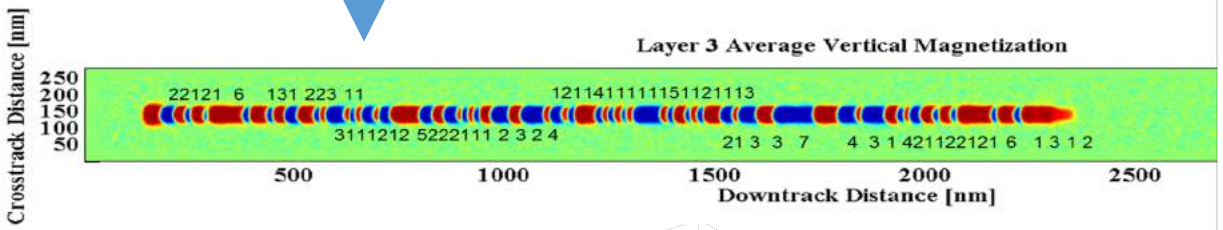


Write Field

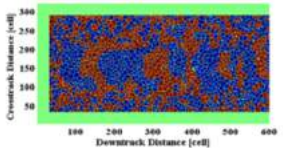
Micromagnetic Write Process:

Multi-Layer Media Model (GPUs)

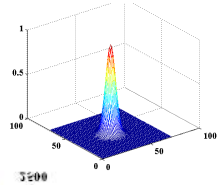
- 3D Media Model
- Many Coupled Layers
- Thermal Field



Multi-layer Granular Magnetic Media

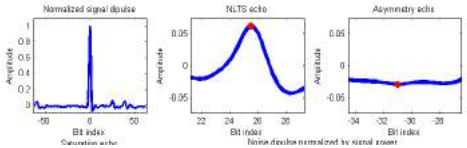


Readback

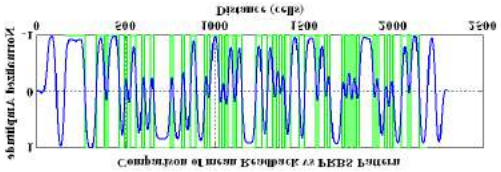


Reader Sensitivity Function

Software Channel (SWC)

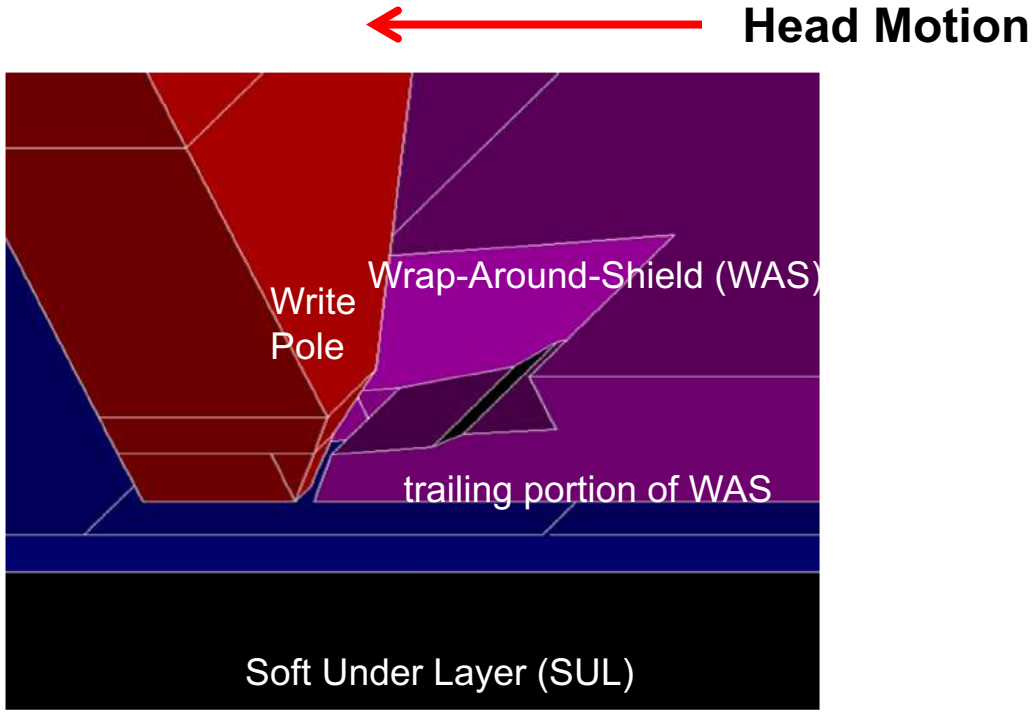


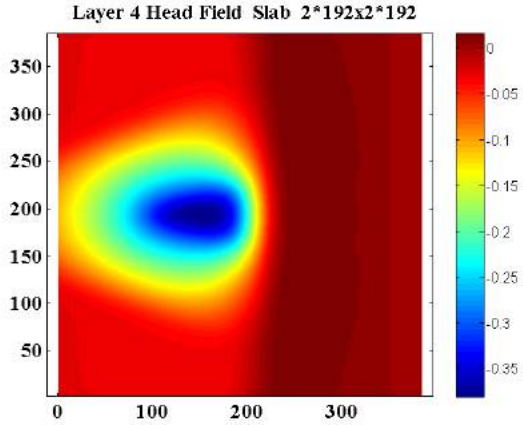
Readback SIgnal



Bit Error Rate (BER) and Analysis:
Signal-to-Noise (SNR), Areal Density Capability (ADC), Distortions

Fields generated from FEM or Micromagnetic Writer Model





Many critical dimensions and angles

Head-Under-Layer Spacing (HUS)

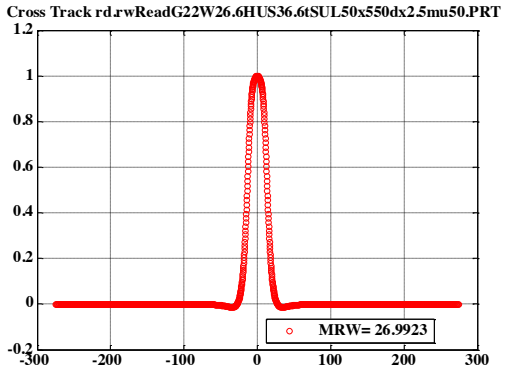
Gap between write pole and trailing portion of WAS

Write field magnitude and angle varies in all three dimensions

Magnetic field produced in each media layer

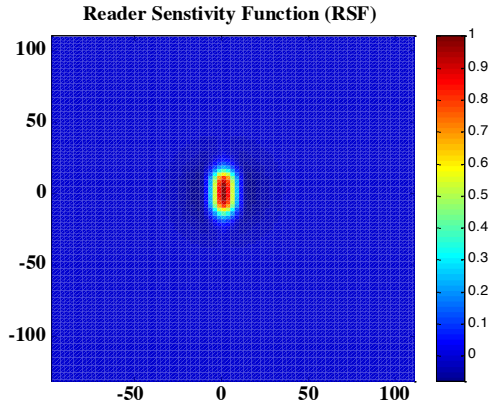
FEM = finite-element-method , magnetostatic model

Readback Process



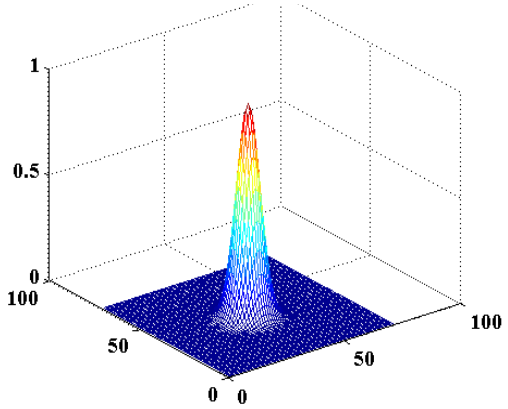
3D Reader Sensitivity Function (RSF):

Convolution of the RSF and magnetization

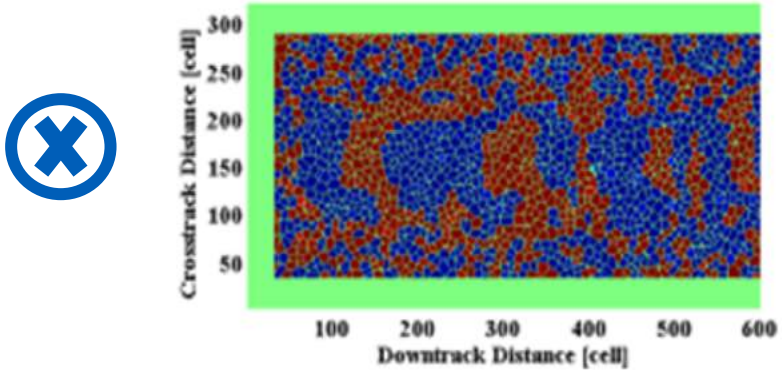


1. Analytic or micromagnetic program to generate RSF
2. Readback is very fast
   - All crosstrack positions are obtained at one time using 2D FFT
   - Post-processing step, all magnetization data is on a separate server.
   - Convolve 3D RSF with the magnetization.
3. Full micromagnetic readback is available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

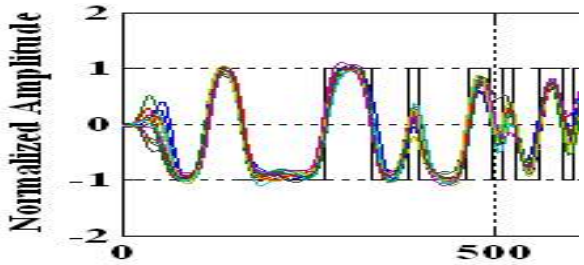
Different RSF for each media layer



RSF

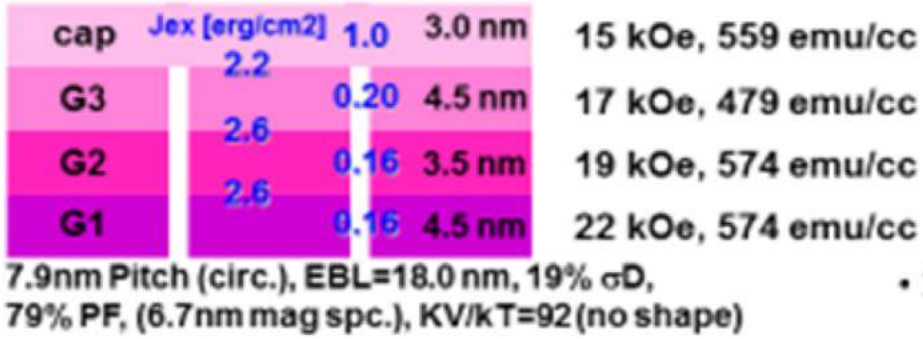


Magnetization

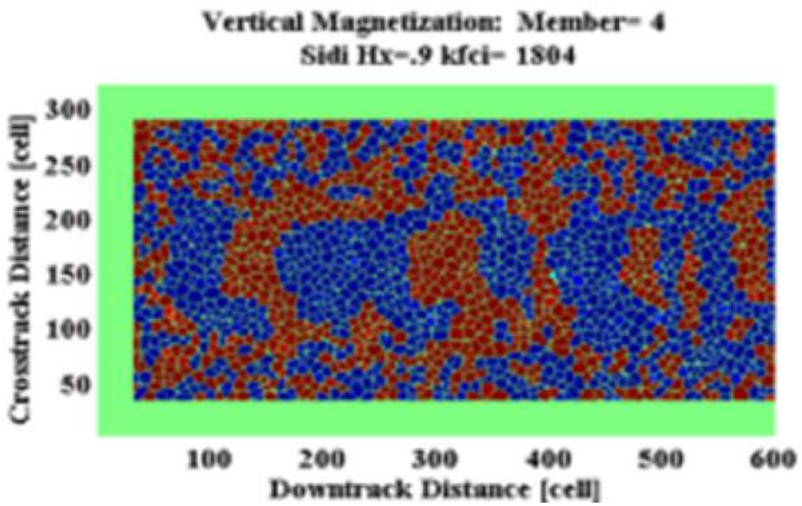


Signal

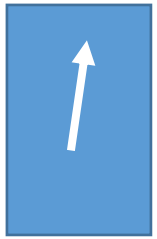
Multi-layer Media Model: Exchange between layers and grains



7.9nm Pitch (circ.), EBL=18.0 nm, 19% σD, 79% PF, (6.7nm mag spc.), KV/kT=92 (no shape)



Top surface of the media



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

$E_a = K \langle m|k \rangle^2$

- Four layer media
- Each grain is built from ~.9nm x .9nm elements, core is ~ 64 elements
- Each layer is a macro-spin.
- Field averaged over a layer and propagated via LLG+Temperature
- Anisotropy, volume and exchange distribution

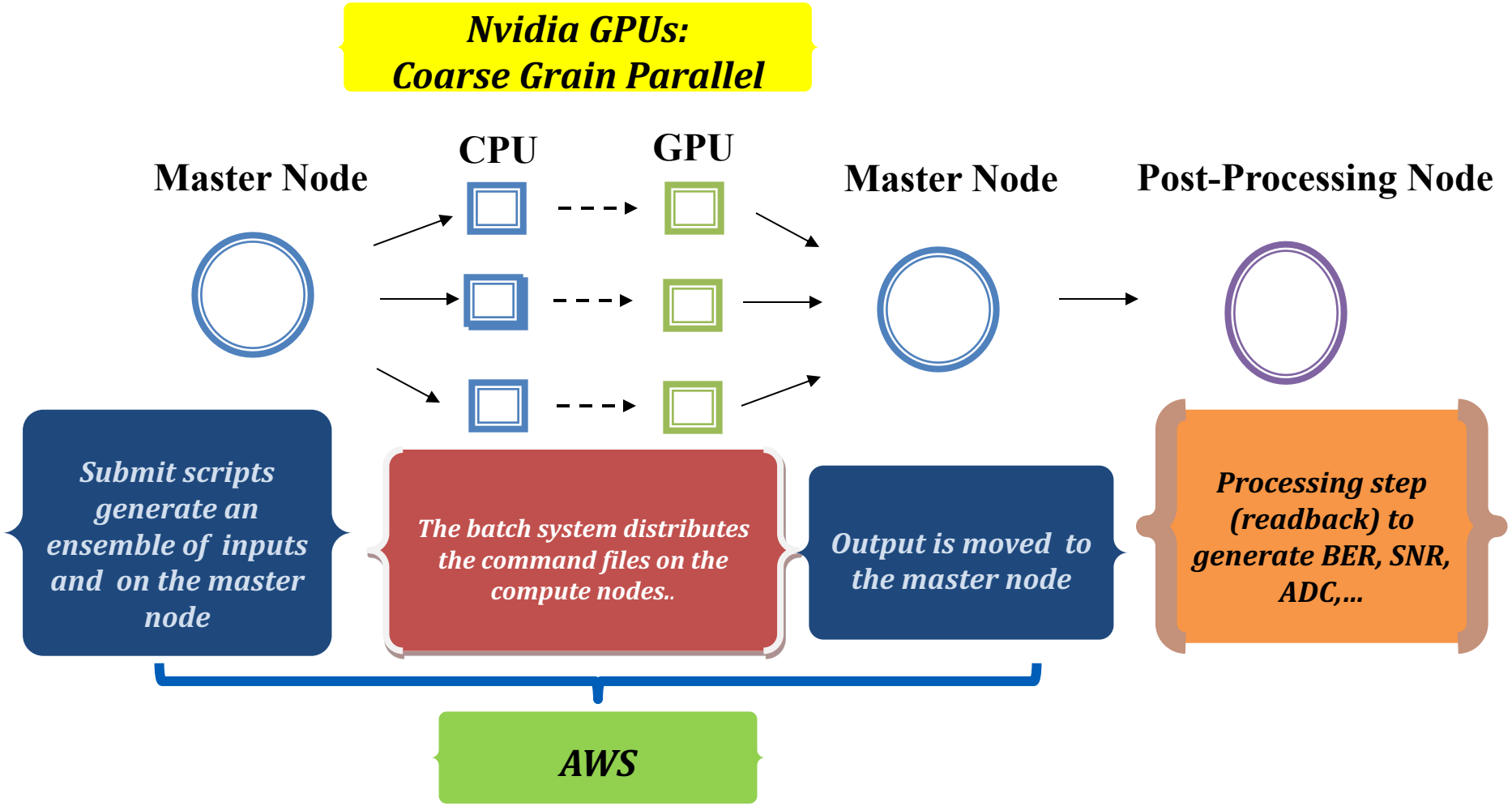
Macro-spin Model (each layer is a macro-spin)

- Long range magnetic field computed using FFT CUDA library
- Exchange interaction is parallelized over 3D regions

Batch System Job Flow

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.**

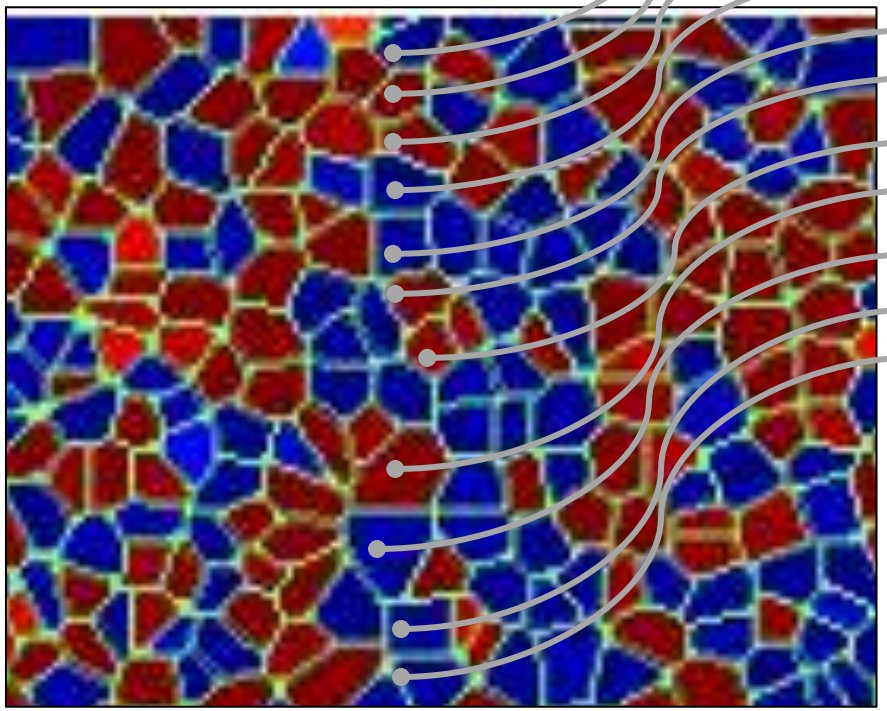


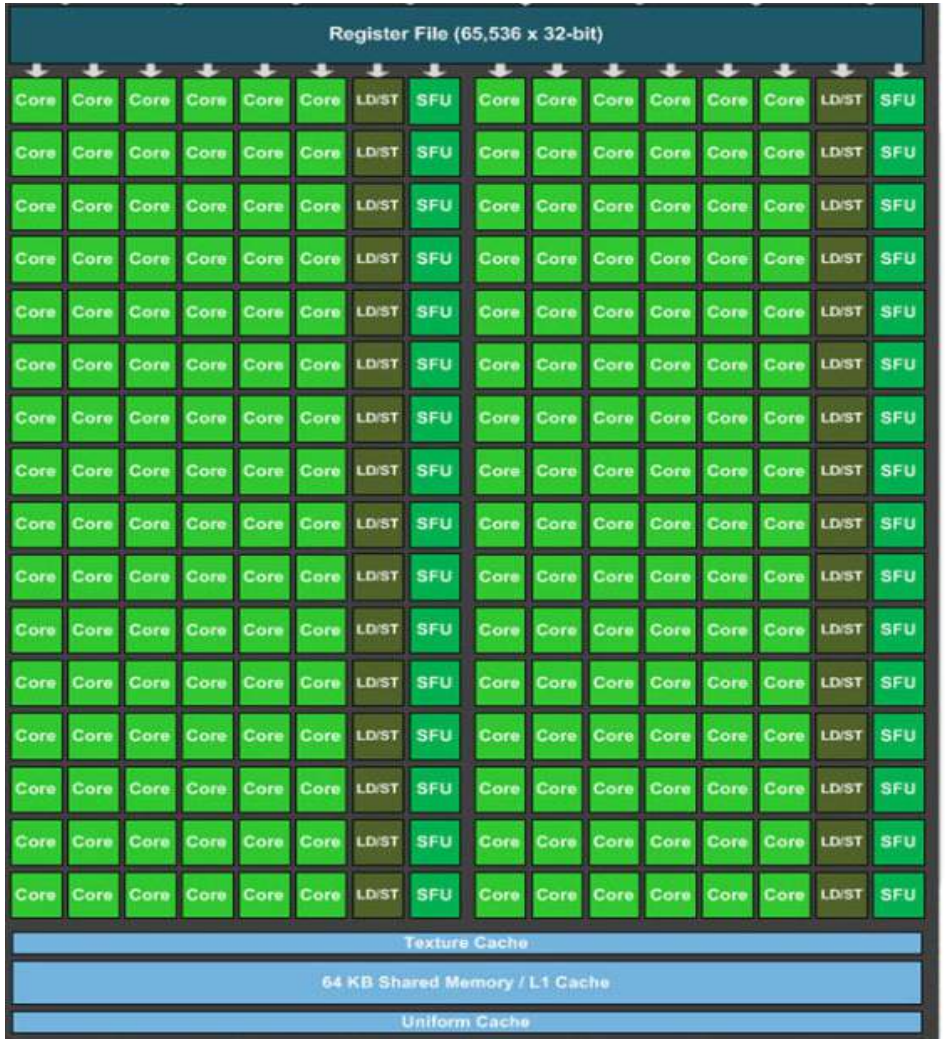
GPU code implementation

- Example: Anisotropy field (one body: <M|K> |K> 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

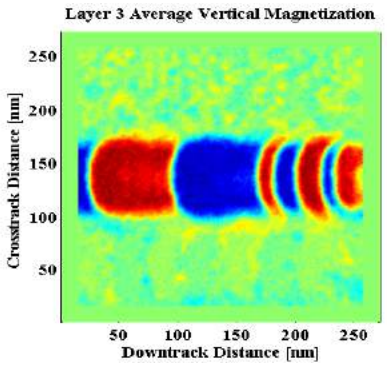
Media



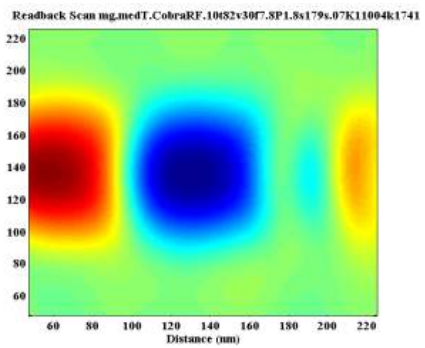


GPU

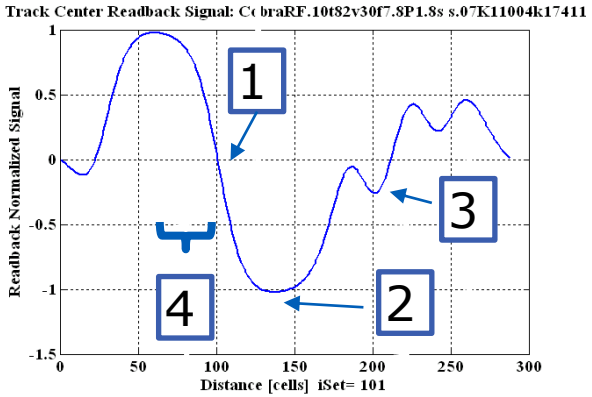
Analysis of Small Bit Pattern ~ 620 Runs Needed



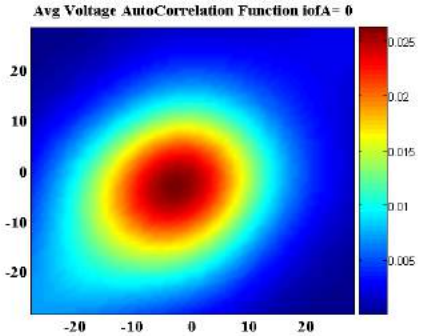
Avg. Magnetization



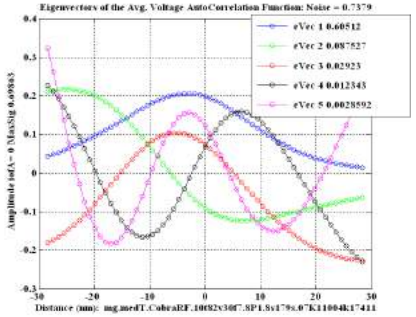
Avg. Readback
via a Linear Model



Center: Avg. Readback



Noise Autocorrelation
at LF Transition



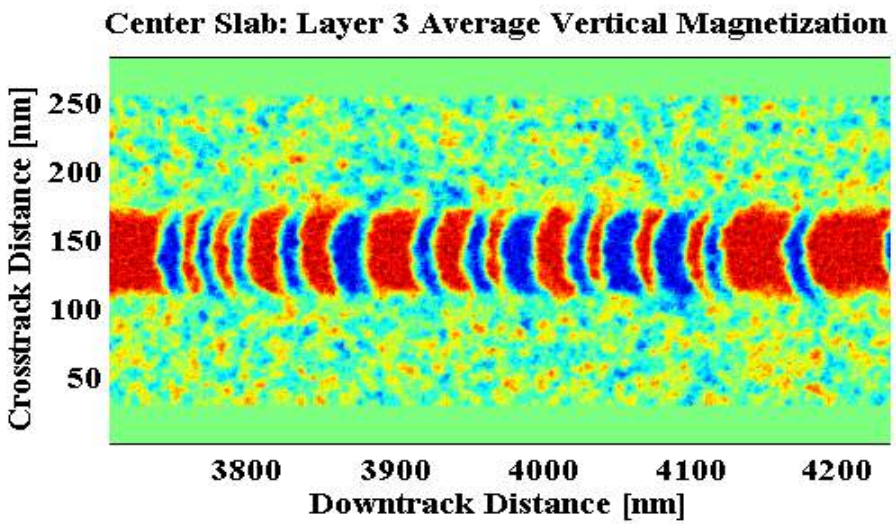
Noise Terms

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

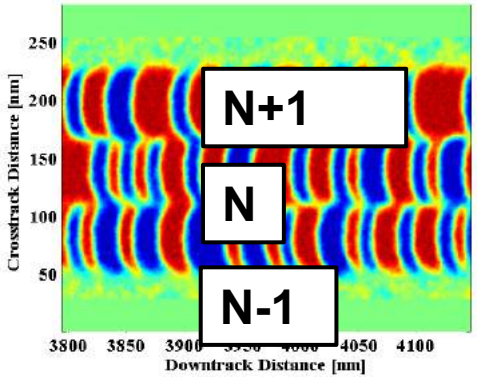
- Jitter noise power is dominant
- Non-transition noise is pattern dependent
  - increases as length of the bit increases
  - noise can be seen in average magnetization and smearing in autocorrelation matrix

Run Time ~ 2-3 hours per CPU core
1-3 minutes on a GPU
(K80, Single Precision)

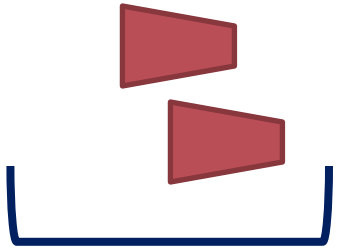
Windowing Algorithm for 511 bit PRBS



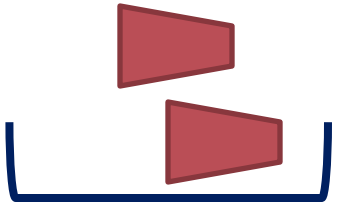
- Cuda FFT is used in each Window
- Head moves in the Window
- Then Window moves
  - Head is always near the center
  - Problem re-started in each Window



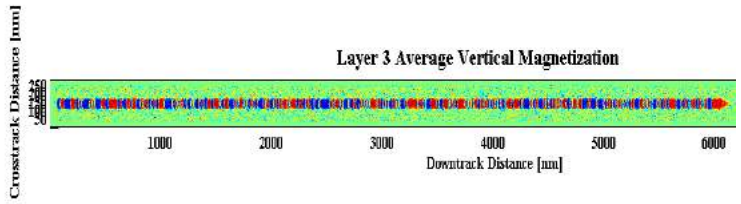
Motion of the Head
Local iteration



Motion of the Window
Global iteration



Triple track 511 bit PRBS now practical with GPUs



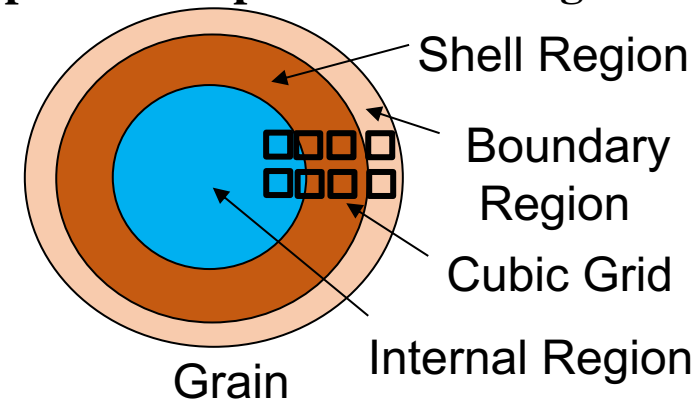
Discretized Media Model

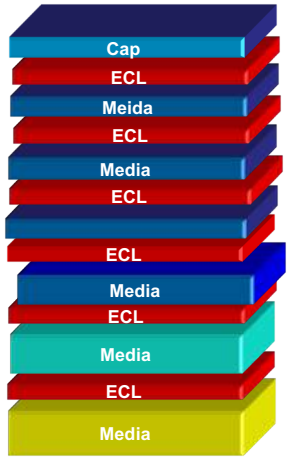
GPUs give us the ability to simulate recording physics at a smaller length scale

Many variables:

Jsii is the surface exchange between cells in the internal regions
Jsis/Jssi is the surface exchange between cells at the boundary of the internal and shell regions
Jss is the exchange between boundary cells
Jsb 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

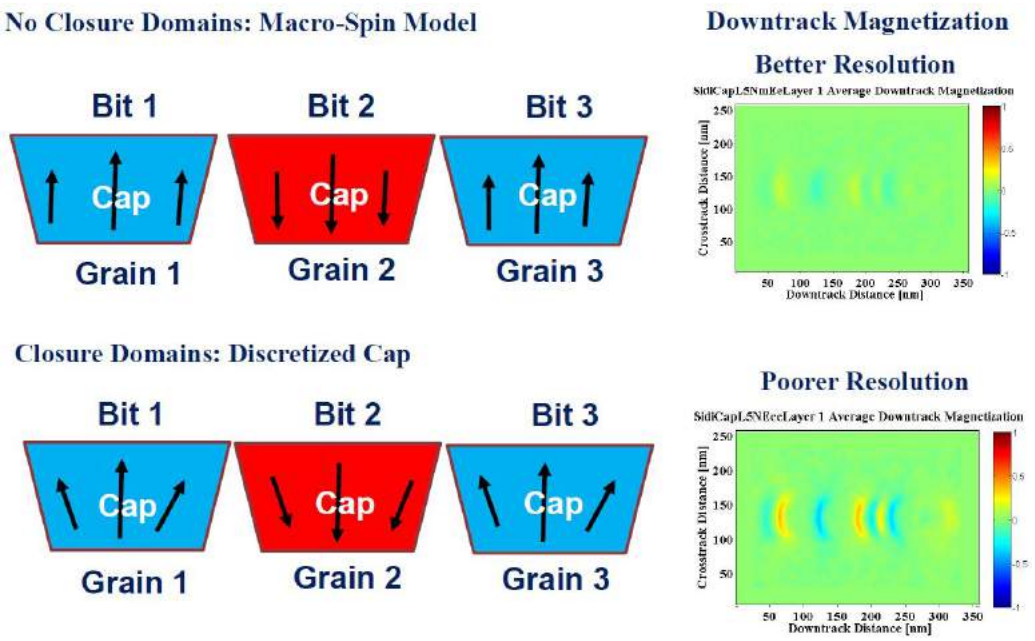
The limit of "metallic" vertical exchange between the layers can be studied (thin ECL and ~ "metallic" vertical exchange)

Thus this Model can represent a range of physical (sputtered) layers

Magnetization in each cell is independently propagated in the LLG+T simulation

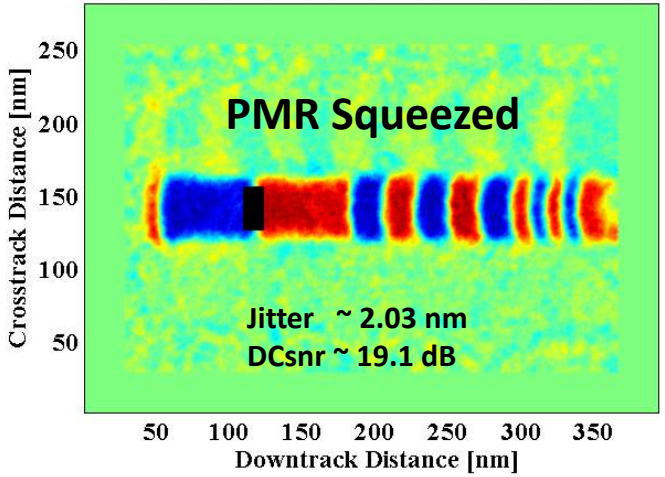
Macro-spin model vs. Discretized models

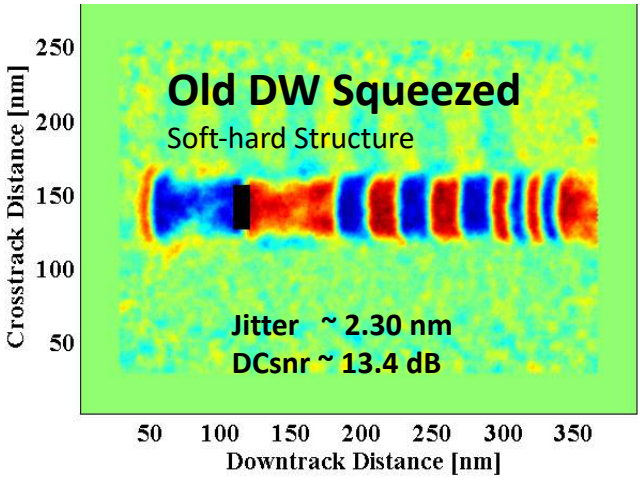
- **Closure domain formation effects resolution performance**
  - **Independent spins in the discretized model can describe the gradual change of magnetization in the cap (top) layer**
  - **Closure domains can form in the discretized model**
  - **Closure domains lead to poorer resolution than found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



In plane magnetization shows the presence or absence of closure domains

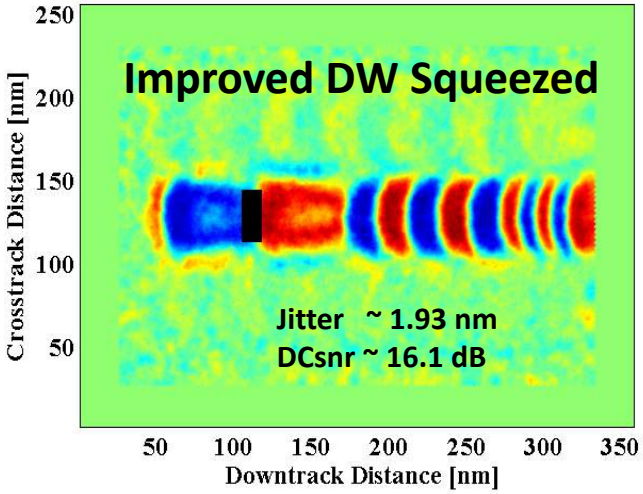
Improved Domain Wall (DW) Media Performance Under Squeeze:





GPUs make multi-dimensional optimization practical in these systems

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



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

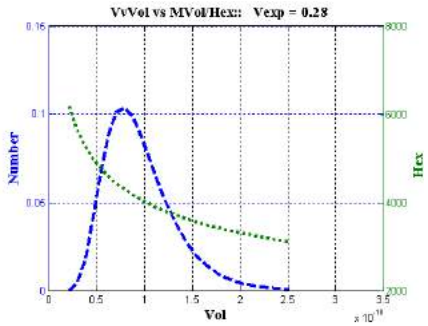
- Workload Expertise is Imperative
  - All workloads are not the same relative to GPU vs. CPU dependence and parallelization
- Complete Proof of Concepts / Trials Before Committing Large Deployments
  - 30 day PoCs allow different configurations and workflows to be evaluated
- Work with AWS and NVidia if the results are not meeting expectations
  - Remove blocking problems through strong partnerships that are developed in meetings not in email exchanges

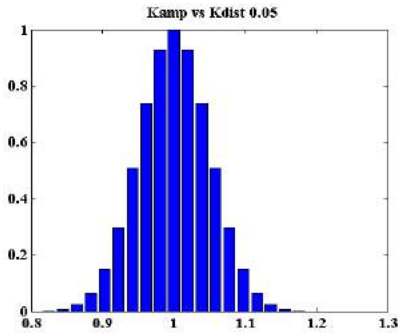
Western Digital®

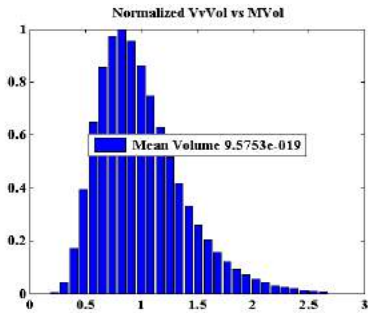
Realistic Media Distributions

Exchange distribution

Bin 2D distribution: (σK, σV) matrix







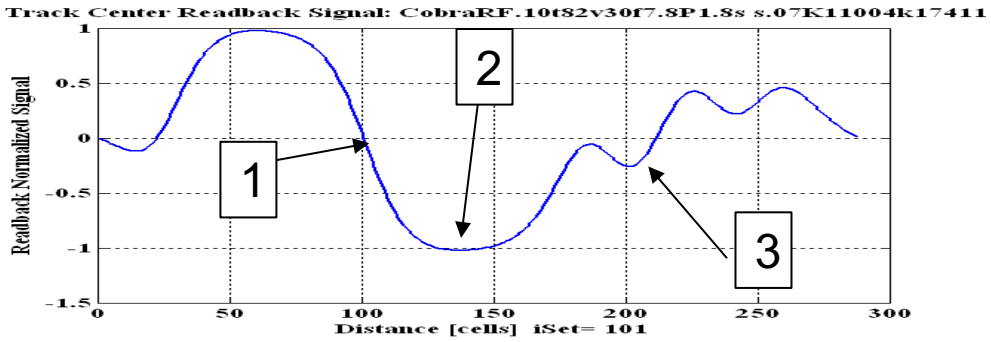
Mean Field Exchange Strength and Volume Dependence

Anisotropy Distribution

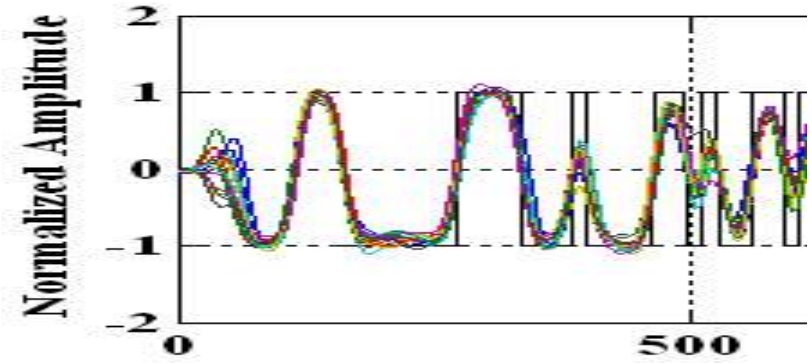
Volume Distribution

- **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

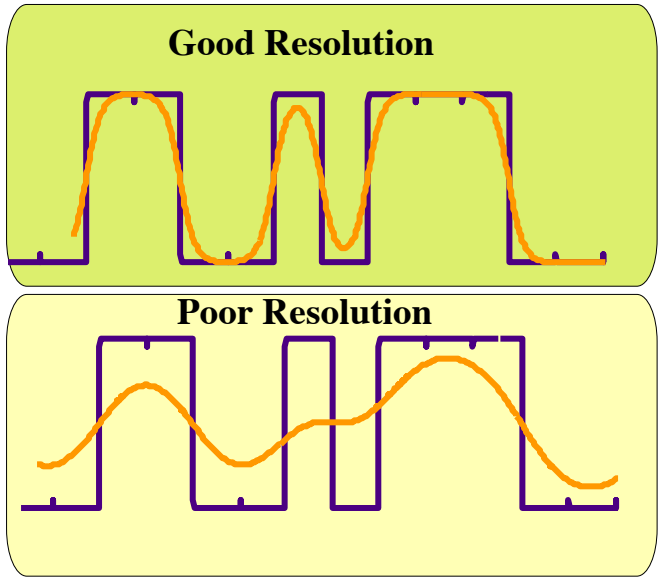


Average Readback Signal



Noise in ensemble members

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)

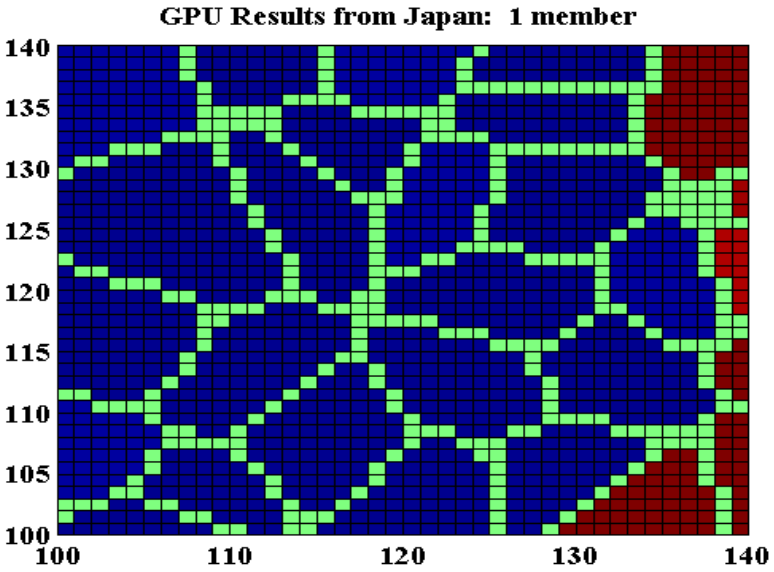


Media Model: Discretization

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

$$dM/dt = -(\gamma/(1+\alpha^2))\ M \times Heff\ \ + = -(\gamma/((1+\alpha^2)Ms))\ M \times (\ M \times Heff\ )$$

α = damping constant **γ = gyromagnetic ratio**