Multi-GPU Hall MHD

Chris Bard
University of Wisconsin-Madison
ASTRONUM 2013

Collaborators: J. Dorelli (NASA-GSFC),
H. Karimabadi (UCSD/SciberQuest)
R. Townsend (UW-Madison)

Why Hall MHD?

- Ideal MHD breaks down at small length scales (e.g. reconnection)

- Want to preserve simplicity of fluid approach
  - Kinetic codes very computationally intensive

- Motivation:
  - Investigate Hall MHD vs. Kinetic simulations
  - Apply to magnetospheres (planetary and massive star)

Hall MHD algorithm

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0$$

$$\frac{\partial}{\partial t}(\rho \mathbf{v}) + \nabla \cdot \left[\rho \mathbf{v}\mathbf{v} + (p + \frac{\mathbf{B}^2}{2})\mathbf{I} - \mathbf{B}\mathbf{B}\right] = 0 \qquad \mathbf{v}_H = -\delta_i \frac{\mathbf{J}}{\rho} \qquad \mathbf{J} = \nabla \times \mathbf{B}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \cdot [(\mathbf{v} + \mathbf{v}_H)\mathbf{B} - \mathbf{B}(\mathbf{v} + \mathbf{v}_H)] = 0$$

$$\frac{\partial}{\partial t}\left(\frac{\rho v^2}{2} + \rho e + \frac{B^2}{2}\right) + \nabla \cdot \left[(\frac{\rho v^2}{2} + \rho e + p)\mathbf{v} + B^2(\mathbf{v} + \mathbf{v}_H) - [(\mathbf{v} + \mathbf{v}_H) \cdot \mathbf{B}]\mathbf{B}\right] = 0$$

- 2nd order MUSCL-Hanock scheme (van Leer 1985)
- HLL approximate Riemann solver (Harten+ 1983, Toro 1999)
- Hyperbolic divergence cleaning (Dedner 2002)
- Second order differencing of current density + MC slope limiter (Toth+ 2008)

Hall MHD algorithm

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0$$

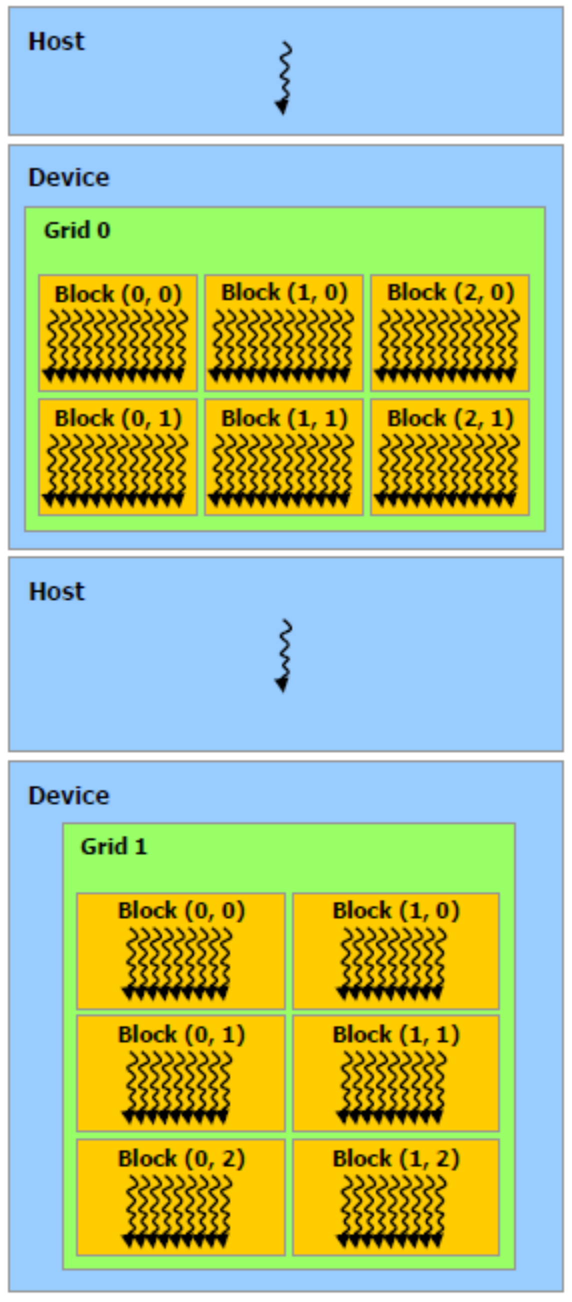
$$\frac{\partial}{\partial t}(\rho \mathbf{v}) + \nabla \cdot \left[\rho \mathbf{v}\mathbf{v} + (p + \frac{\mathbf{B}^2}{2})\mathrm{I} - \mathbf{B}\mathbf{B} \right] = 0 \qquad \mathbf{v}_H = -\delta_i \frac{\mathbf{J}}{\rho} \qquad \mathbf{J} = \nabla \times \mathbf{B}$$

$$\frac{\partial \mathbf{B}}{\partial t} + \nabla \cdot [(\mathbf{v} + \mathbf{v}_H)\mathbf{B} - \mathbf{B}(\mathbf{v} + \mathbf{v}_H)] = 0$$

$$\frac{\partial}{\partial t}\left(\frac{\rho v^2}{2} + \rho e + \frac{B^2}{2} \right) + \nabla \cdot \left[(\frac{\rho v^2}{2} + \rho e + p)\mathbf{v} + B^2(\mathbf{v} + \mathbf{v}_H) - [(\mathbf{v} + \mathbf{v}_H) \cdot \mathbf{B}]\mathbf{B} \right] = 0$$

This scheme robustly captures MHD discontinuities --

The addition of the Hall effect tends to smear these out due to the physical dispersion induced by whistler waves



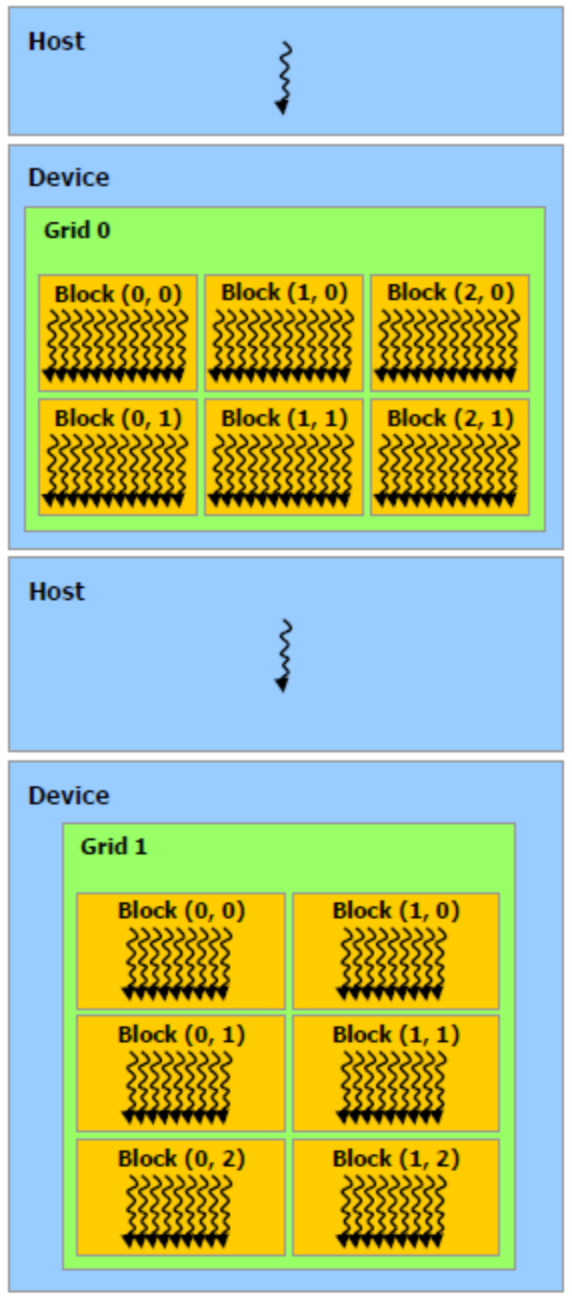
GPUs with CUDA

- Heterogeneous programming model
- High degree of parallelism
  - Thousands of threads executing concurrently
- Latency/Throughput tradeoff

MHD on GPUs

Thread	Thread	Thread	Thread
Thread	Thread	Thread	Thread

For the MHD algorithm, the calculations for each grid cell are independent --> Can be easily parallelized!



GPUs with CUDA

- Types of Memory
  - Global
  - Shared
  - Register
- We utilize a register-heavy approach
- Tradeoff: Memory footprint vs. Speed

Precursor: Single GPU speedups

Proof of viability: Compare timing results for an ideal GPU MHD code vs. a CPU code

CPU: one core of a Intel Nehalem (2.8 GHz)
GPU: NVIDIA GTX480 (Fermi architecture)

Precursor: Single GPU Speedups

Problem Size	Unoptimized C	Optimized C	CUDA
64^2	13.37 s	6.45 s	0.57 s
128^2	73.39	41.80	1.81
256^2	484.33	277.73	5.24
512^2	2366.45	1476.98	18.27
1024^2	11488.6	8029.35	63.84

Numbers in () are speedups compared to Optimized C timings

Problem Size	Register 16	Register 8	Register 4	Register 2
64^2	0.8 s (8.1)	0.57 s (11)	0.67 s (9.7)	1.2 s (5.4)
128^2	2.25 (19)	1.81 (23)	2.04 (21)	4.34 (9.6)
256^2	6.52 (43)	5.24 (53)	6.71 (41)	16.13 (17)
512^2	22.58 (65)	18.27 (81)	25.19 (59)	68.12 (22)
1024^2	84.62 (95)	63.84 (126)	90.61 (89)	253.88 (32)

Bard+Dorelli 2013, JCP, submitted

Precursor: Single GPU Speedups

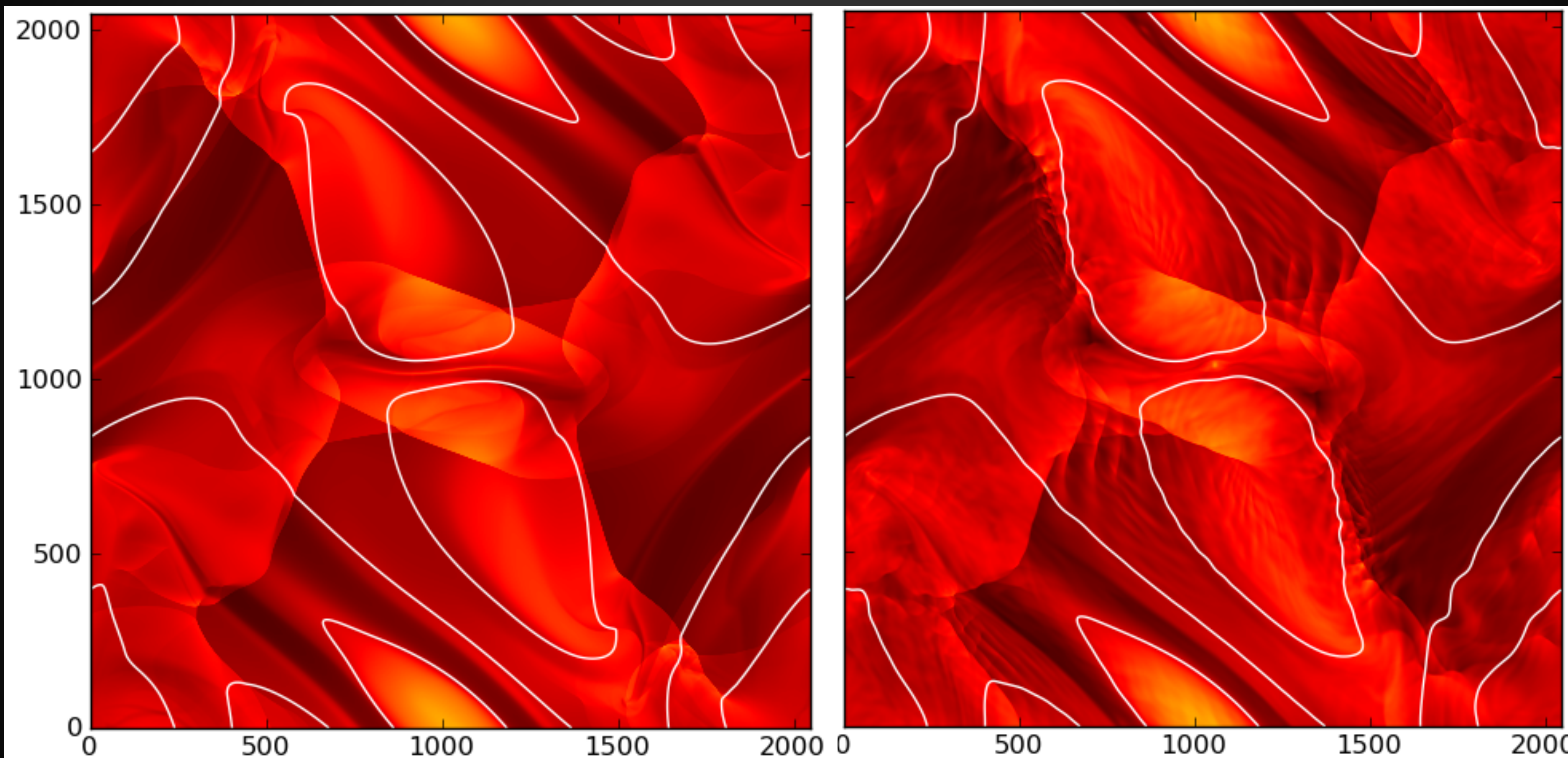
Problem Size	Unoptimized C	Optimized C	CUDA
64^2	13.37 s	6.45 s	0.57 s
128^2	73.39	41.80	1.81
256^2	484.33	277.73	5.24
512^2	2366.45	1476.98	18.27
1024^2	11488.6	8029.35	63.84

Maximum speedup: 126x

Problem Size	Register 16	Register 8	Register 4	Register 2
64^2	0.8 s (8.1)	0.57 s (11)	0.67 s (9.7)	1.2 s (5.4)
128^2	2.25 (19)	1.81 (23)	2.04 (21)	4.34 (9.6)
256^2	6.52 (43)	5.24 (53)	6.71 (41)	16.13 (17)
512^2	22.58 (65)	18.27 (81)	25.19 (59)	68.12 (22)
1024^2	84.62 (95)	63.84 (126)	90.61 (89)	253.88 (32)

Bard+Dorelli 2013, JCP, submitted

Benchmark: Orszag-Tang Vortex



Ideal

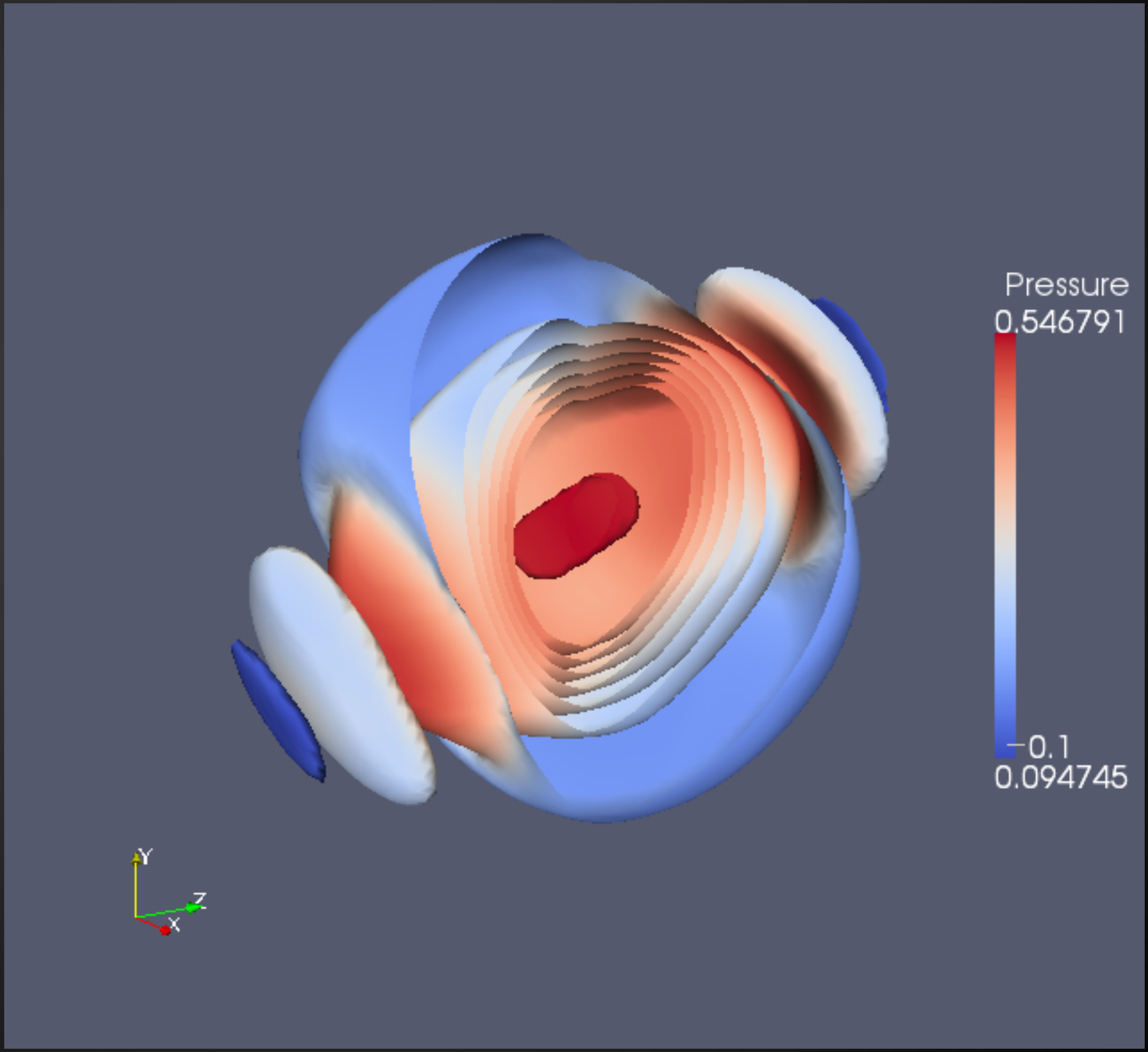
2048x2048, 64 GPUs

Hall

Benchmark: Magnetized Blast Wave

Contours: Density

Color: Pressure



Benchmark: Whistler Wave

Following Toth+ (2008)

$$\rho = 1$$
$$p = 1$$
$$B_x = 100$$
$$\lambda = 200$$
$$v_y = -\delta_v \cos(kx)$$
$$v_z = \delta_v \sin(kx)$$
$$B_y = \delta_B \cos(kx)$$
$$B_z = -\delta_B \sin(kx)$$

$$\frac{\delta_v}{\delta_B} = \frac{|B_x|}{v_{ph}\rho}$$
$$v_{ph} = \frac{v_w}{2} + \sqrt{v_A^2 + v_w^2/4} = 169.345$$
$$\text{Period} = \lambda/v_{ph} = 1.181022$$

Benchmark: Whistler Wave

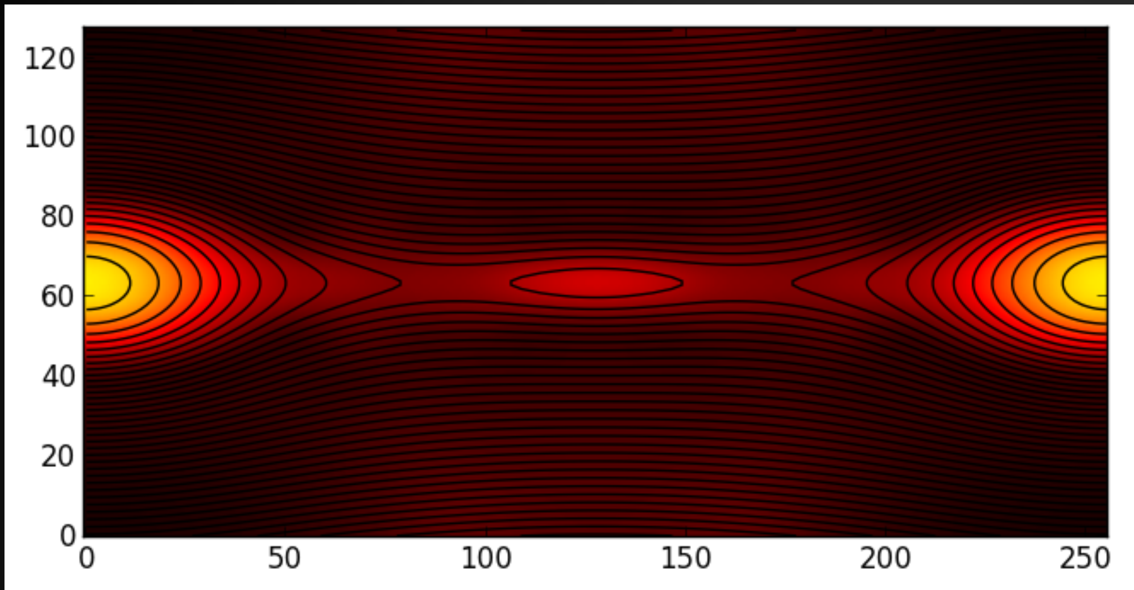
Followed procedure of Toth+ (2008)

Relative errors after one period:

$$E_n = \frac{\sum_{i=1}^{n} |v_{z,i}(t_{\max}) - v_{z,i}(0)|}{\sum_{i=1}^{n} |v_{z,i}(0)|}$$

N	1D	1D Ratio	2D	2D Ratio
16	0.26732	–	0.244582	–
32	0.06657	4.01	0.06156	3.97
64	0.01556	4.2	0.01442	4.27
128	0.00372	4.2	0.003449	4.2
256	0.00091	4.1	0.00084	4.1

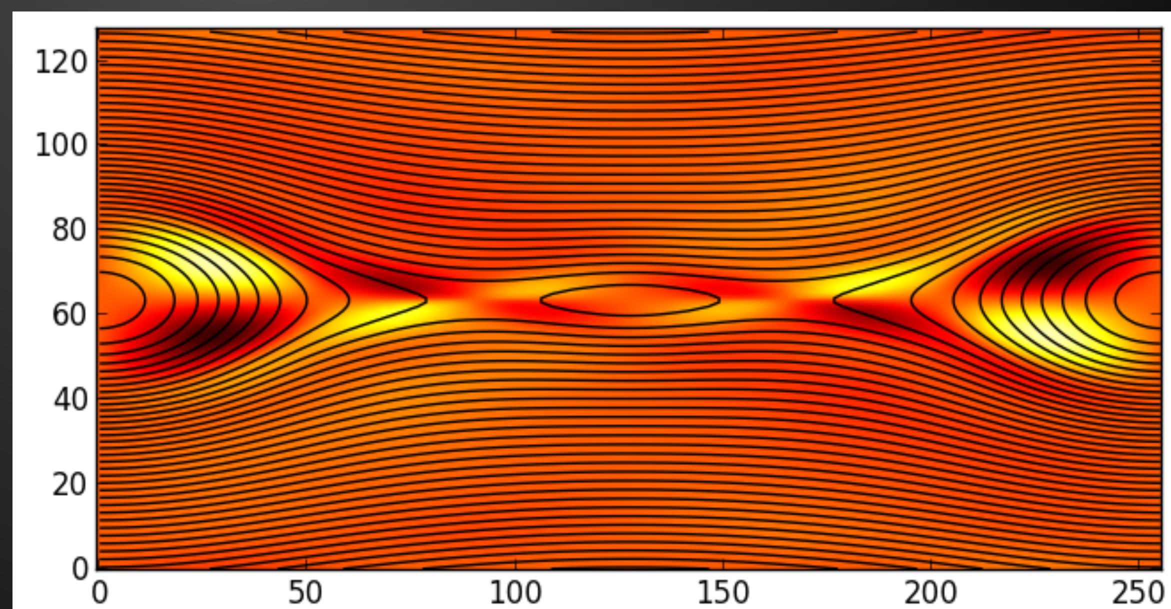
Benchmark: GEM



Based on Birn+ (2001)

Lx = 25.6 d_i
Lz = 12.8 d_i

Density



Out of plane B

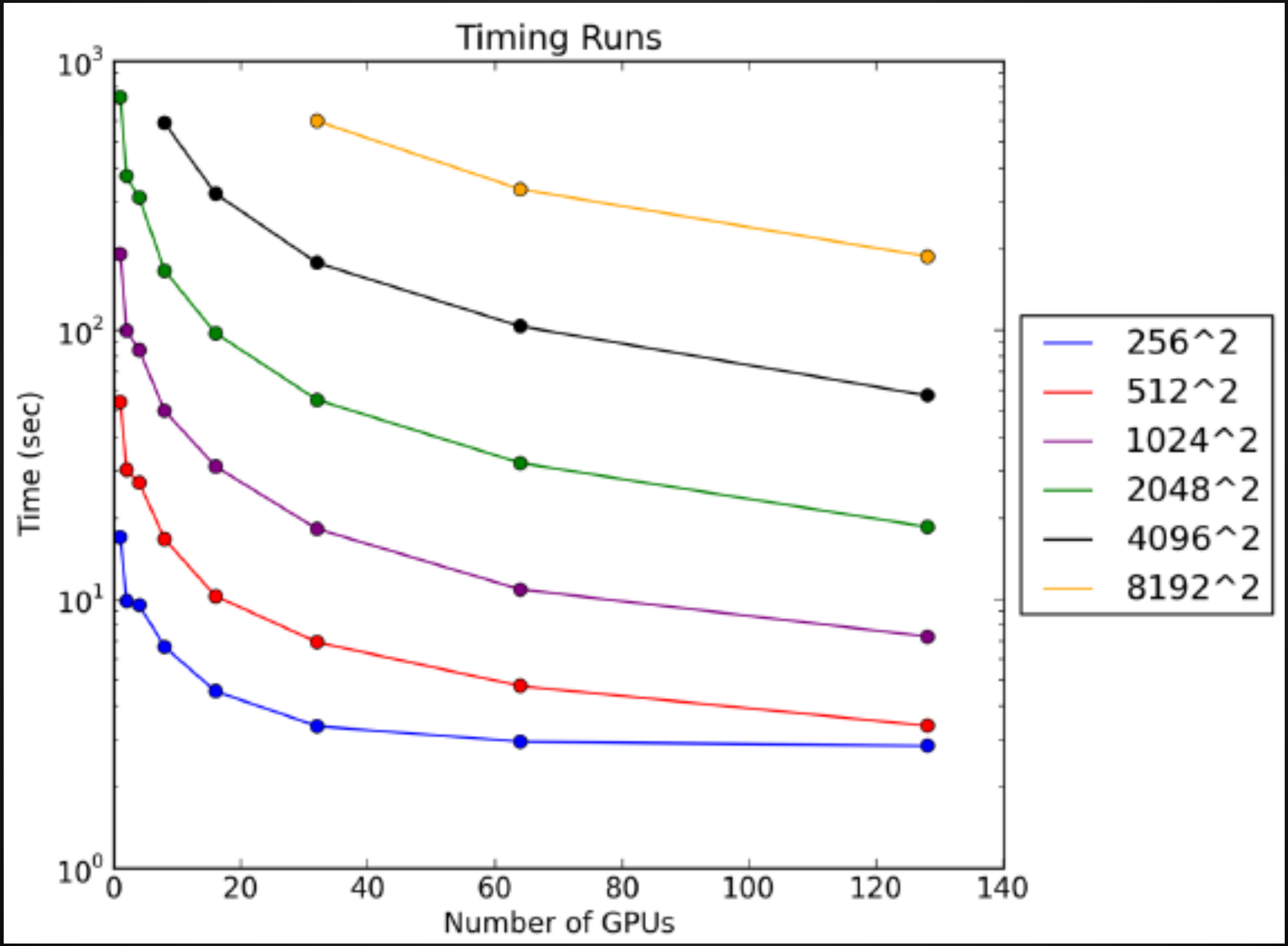
Large GEM Movie

$L_x = 204.8\ d_i$

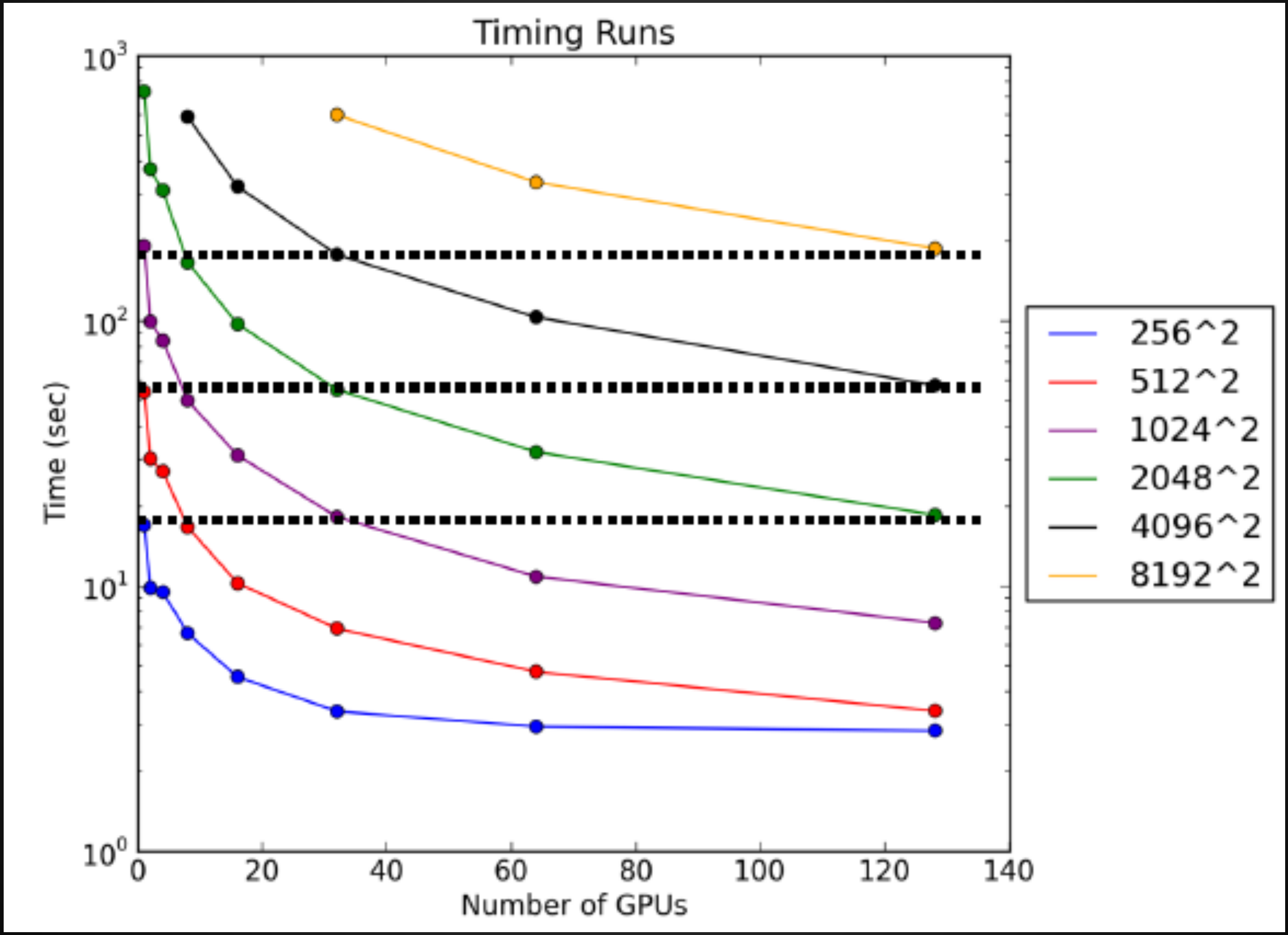
$L_z = 102.4\ d_i$

All other parameters same as Birn+ (2001)

Timing Results - 2D



Weak Scaling - 2D



Future Work

- Continue scaling tests
  - Currently running on up to 128 GPUs (512^3 grid)
  - Distant Future goal: 2048^3

- Timing results
  - Compare to multi-CPU + MPI versions

- Investigate phenomena
  - Compare Hall MHD with kinetic PIC
  - Magnetospheres (Planetary/Massive Star)