GPU-acceleration of an Established Solar MHD code using OpenACC

Ronald M. Caplan, Jon A. Linker, Zoran Mikić, Cooper Downs, and Tibor Török



Slides available at:
predsci.com/~caplanr



ASTRONUM 2

NUMERICAL MODELING OF SPACE PLASMA F IUNE 25 – 29, 2018, PANAMA CITY BEACH, FI



Outline

Accelerated Computing **V**OpenACC The MAS Code OpenACC Implementation
 Results • Outlook





Accelerated Computing

An accelerator is a discrete piece of hardware designed for massively parallel computations

Many brands/types of accelerators, here we focus on NVIDIA GPUs.

Why use accelerators?

1) Performance (FLOP/s and Memory Bandwidth)

2) Compact Performance







8xGPU

- 3) Saves Energy
- 4) Saves Money







Accelerated Computing

Who uses accelerators?

GPU Developers 10X in 5 Yrs



Total GPU FLOPS of Top 50 Systems 15X in 5 Yrs



Top stories





Move Over, China: U.S. Is Again Home to World's Speediest Supercom... This computer can do more calculations per second than the world...

... consists of 4,608 compute servers, each containing two 22-core IBM Power9 processors and six NVIDIA Tesla V100 GPU accelerators ..."

ASTRONUM 2018

- 9:40 AM M. Zingale Tues.
- Tues. 3:30 PM M. Zhang
- Wed. 8:25 AM N. Pogorelov MS-FLUKSS
- Thurs. 1:55 PM P. Woodward PPMStar

IBM and the Department of Energy show off world's fastest superc...



Accelerated Computing

NVIDIA

CUDA

Why not use accelerators?

- Not all algorithms suitable
- Hard to program
 Originally, only option was language extension APIs
 - V CUDA (NVIDIA-specific)
 - OpenCL (more general)
 - This involves rewriting large sections of code and maintaining at least two code bases.

Portability and longevity risk
 What if GPUs go away?

glo	ba]	L	V	oi	d	Sā	axp	by ((ir	۱t
									f]	loi
									f]	loi
int	i	=	b1	00	k1	[d>	<.>	(*	blc	bcl
if	(i	<	N)	у	[i	.]	=	a'	*x[i
}										
• • •										
const	: ir	nt	BL	00	K_	_S1	ZE	=	204	18
float	*0	א_t	* <mark>ب</mark> *	d_	<u>у;</u>					
dim3	din	nB1	.00	k (BL	.00	CK_	_S:	IZE	=)
dim3	din	nGr	id	((ir	nt)	Ce	2i	1(((N-
cuda№	1al]	loc	:((v	oi	.d	**	<)	&c	1_ 2
cuda№	1al]	loc	:((V	oi	.d	**	<)	&c	<u>; </u> ;
cuda№	lemo	сру	'(d	_x	و	ر X	. N	ر ا	сι	ld
cuda№	1emc	сру	'(d	_y	و	у,	. N	ر ا	cι	ıda
saxpy	/<<<	di	.mG	iri	d,	di	mE	310	ock	(>]
cuda№	1emc	сру	'(y	'	d_	<u>ر</u> y _	N	, ا	Cι	ıda
cudaF	ree	e(d	_x	:);						
cudaF	ree	e(d	[_у	');						
	glo int if } const float dim3 dim3 cudaM cudaM cudaM cudaM cudaM cudaM cudaM cudaM cudaM	global int i if (i } const ir float *d dim3 din dim3 din cudaMal1 cudaMal1 cudaMemd cudaMemd cudaMemd cudaMemd cudaFree cudaFree	<pre>global int i = if (i < } const int float *d_x dim3 dimB1 dim3 dimB1 dim3 dimGr cudaMalloc cudaMemcpy cudaMemcpy cudaFree(d cudaFree(d)</pre>	<pre>globalv int i = bl if (i < N) } const int BL float *d_x,* dim3 dimBloc dim3 dimGrid cudaMalloc(cudaMalloc(cudaMemcpy(d cudaMemcpy(d cudaFree(d_x cudaFree(d_x cudaFree(d_y)</pre>	<pre>globalvoi int i = bloc if (i < N) y } const int BLOC float *d_x,*d_ dim3 dimBlock(dim3 dimGrid((cudaMalloc((v cudaMalloc((v cudaMemcpy(d_x cudaMemcpy(d_y saxpy<<<dimgri cudaFree(d_x); cudaFree(d_y);</dimgri </pre>	<pre>globalvoid int i = blockI if (i < N) y[i } const int BLOCK_ float *d_x,*d_y; dim3 dimBlock(BL dim3 dimBlock(BL dim3 dimGrid((ir cudaMalloc((voi cudaMalloc((voi cudaMemcpy(d_x, cudaMemcpy(d_y, saxpy<<<dimgrid, saxpy<<<dimgrid, cudaFree(d_x); cudaFree(d_y);</dimgrid, </dimgrid, </pre>	<pre>globalvoid sa int i = blockIdx if (i < N) y[i] } const int BLOCK_SI float *d_x,*d_y; dim3 dimBlock(BLOC dim3 dimGrid((int)) cudaMalloc((void cudaMalloc((void cudaMemcpy(d_x, x, cudaMemcpy(d_y, y, saxpy<<<dimgrid,di cudaMemcpy(y, d_y, cudaFree(d_x); cudaFree(d_y);</dimgrid,di </pre>	<pre>globalvoid saxp int i = blockIdx.> if (i < N) y[i] = } const int BLOCK_SIZE float *d_x,*d_y; dim3 dimBlock(BLOCK_ dim3 dimGrid((int)ce cudaMalloc((void ** cudaMalloc((void ** cudaMemcpy(d_x, x, N cudaMemcpy(d_y, y, N saxpy<<<dimgrid,dime cudaMemcpy(y, d_y, N cudaFree(d_x); cudaFree(d_y);</dimgrid,dime </pre>	<pre>globalvoid saxpy int i = blockIdx.x*N if (i < N) y[i] = a } const int BLOCK_SIZE= float *d_x,*d_y; dim3 dimBlock(BLOCK_SI dim3 dimGrid((int)cei dim3 dimGrid((int)cei cudaMalloc((void **) cudaMalloc((void **) cudaMemcpy(d_x, x, N, cudaMemcpy(d_y, y, N, saxpy<<<dimgrid,dimbloc cudaMemcpy(y, d_y, N, cudaFree(d_x); cudaFree(d_y);</dimgrid,dimbloc </pre>	<pre>globalvoid saxpy(ir f] f] int i = blockIdx.x*blo if (i < N) y[i] = a*x[} const int BLOCK_SIZE=204 float *d_x,*d_y; dim3 dimBlock(BLOCK_SIZE dim3 dimBlock(BLOCK_SIZE dim3 dimGrid((int)ceil((cudaMalloc((void **) &c cudaMalloc((void **) &c cudaMalloc((void **) &c cudaMemcpy(d_x, x, N, cu cudaMemcpy(d_y, y, N, cu saxpy<<<dimgrid,dimblock cudaMemcpy(y, d_y, N, cu cudaFree(d_x); cudaFree(d_y);</dimgrid,dimblock </pre>



)

; +0.0)/dimBlock.x));

x, sizeof(float)*N);
y, sizeof(float)*N);
aMemcpyHostToDevice);
aMemcpyHostToDevice);

>>(N, a, d_x, d_y);

aMemcpyDeviceToHost);

OpenACC

More Science, Less Programming
 Directive-based API, began as off-shoot of OpenMP

- Uniform source code (no branches!)
- Low-risk (can compile to CPU as before)

UpenACC

- Vendor-independent
 (PGI, CRAY, GNU, OMNI, SunWay)
- Multiple Target Architectures
 (GPU, Multicore x86, FPGA, SunWay)
- Designed for rapid development, especially for pre-existing codes
- Used by >90% of GPU Industry codes run on Titan at ORNL

BOOSTING INDUSTRY WITH OPENACC

BY JONATHAN HINES • 2 WEEKS AGO • INDUSTRY

FORTRAN:

INDUSTRIAL USERS BENEFIT FROM CODES ACCELERATED BY DIRECTIVE-BASED PROGRAMMING STANDARD

ne of the biggest hurdles for users who want to take adv accelerated computing is the time required to write and software. That's especially true for industrial users, who carefully evaluate the projected returns on such an investment.

<u>OpenACC</u>, a directive-based programming standard for accelerator systems, offers a potential alternative to labor-intensive code rebuil programmers to adapt specific sections of an application for GPU ac while leaving other sections unchanged.

As home to one of the most powerful GPU-accelerated supercompu world, the <u>Oak Ridge Leadership Computing Facility</u> (OLCF), a <u>US D</u> <u>Energy</u> (DOE) <u>Office of Science</u> User Facility located at DOE's <u>Oak F</u>

#pragma acc !\$acc

OVER 100 APPS* USING OpenACC

ANSYS Fluent Gaussian VASP LSDalton MPAS GAMERA

GTC XGC ACME FLASH COSMO Numeca

* Applications in production and development

OpenACC





- Established MHD code with over 15 years of development used Ψ extensively in solar physics research
- Written in FORTRAN 90 (~50,000 lines), parallelized with MPI Ψ
- Available for use at the Community Coordinated Modeling Center (CCMC) Ψ



Predicted Corona of the August 21st, 2017 Total Solar Eclipse







Community MODELLAR Centres

Simulation of the Feb. 13th, 2009 CME

MAS: `Full' MHD Model Equations





CORONAL HEATING

 $H = H^* + rac{
ho}{4\,\lambda_+} \left[|z_-| \, z_+^2 + |z_+| \, z_-^2
ight]$ $\lambda_{\perp} = \lambda_0 \sqrt{\frac{B_w}{|\mathbf{B}|}} |z_{\pm}(r = R_{\odot})| = \mathbf{z}_0$

 $S = (\Delta t^2 \,\tilde{k}^2)^{-1} \, \left(C_w^2 / (1 - C_f)^2 - 1 \right)$ $C_f = \Delta t \, \tilde{k} \cdot \mathbf{v}$ $C_w^2 = 0.25 \,\Delta t^2 \,\tilde{k}^2 \,(v_c^2 + |\mathbf{v}_A|^2)$ $\tilde{k}^2 = 4 \left(\Delta r^{-2} + (r \Delta \theta)^{-2} + (r \Delta \phi \sin \theta)^{-2} \right)$

MAS: MHD Model Equations ("Zero-Beta")

- In the low corona outside of active regions, the plasma beta is very small Ψ (i.e. dynamics dominated by magnetic field)
- This allows a simplified "zero-beta" model to be useful in many cases Ψ (e.g. modeling the initial configuration and onset dynamics of a CME eruption)



Since the core algorithms are the same as the full model, this makes an ideal target Ψ for our initial OpenACC implementation (stepping stone)

MAS: Algorithm Summary

- Finite difference on non-uniform spherical grid
- Parallelized with MPI
- Explicit and implicit time-stepping Ψ algorithms
- Implicit time-step (backward-Euler) solved with Preconditioned Conjugate Gradient
- Two communication-free Ψ preconditioners: **PC1** and **PC2**
- For `hard' solves, PC2 faster than PC1 for 'easy' solves, **PC1** faster than **PC2**



PC1

Point-Jacobi



Block-Jacobi with ILU0

Production Test Run

Zero-Beta Unstable Flux Rope Eruption

Run information

Physical code time duration: **198 seconds** Number of time-steps: **695**

 $160 \times 267 \times 246 \sim 10.5$ million points

Spherical Domain with $\,r_{
m max}=10\,R_\odot$

Detailed run information

	Ν	Δ_{\min}	Δ_{\max}	$\max \left \frac{\Delta_{i+1} - \Delta_i}{\Delta_{i+1}} \right $
r	160	$800\mathrm{km}$	$530000\mathrm{km}$	9%
θ	267	0.066°	9.45°	11%
ϕ	246	0.067°	14.61°	10%
t	695	0.001 sec	$0.17 \sec$	11%

PCG Solver Iterations per Time Step (mean)

	SI Predictor	SI Corrector	Viscosity
PC1	72	75	554
PC2	$30 \rightarrow 34$	$31 \rightarrow 35$	$104 \rightarrow 197$







Time = 0.0000 n = 1



OpenACC Implementation: Preliminaries

Profile code

PCG over 90% of run-time

Viscosity is hardest solve

Analyze algorithms for GPU-compatibility

 Most PCG steps and explicit time-stepping
 "vector-friendly"

Preconditioners

PC1: directives only (**portable**) PC2: cuSparse (not portable)

Test performance through "Proof-of-concepts" **W** DIFFUSE: Explicit finite-difference POT3D: PCG+PC1/PC2

Based on results of POT3D, we only accelerate PC1 in MAS



Run using 16 nodes of 24-core Haswell CPUs (PC2)



OpenACC Implementation: Examples

$CPU \leftrightarrow GPU$ Data transfers

allocate and initialize "y" !\$acc enter data copyin (y) use "y" in OpenACC compute regions !\$acc update self (y) CPU version of "y" updated for I/O, etc. ... !\$acc exit data delete(y)

Basic Loop

!\$acc parallel default(present) !\$acc loop **do** i=1,n $y(i) = a^*x(i) + y(i)$ enddo !\$acc end parallel

Reductions

!\$acc kernels loop present(y) !\$acc& reduction(+:sum) do j=1,m sum = sum + y(j)enddo

FORTRAN Array-syntax

!\$acc kernels default(present) $y(:) = a^*x(:) + y(:)$ **!**\$acc end kernels

OpenACC Implementation: Multi-GPU

Multiple GPUs with MPI



(assumes linear affinity)



(code shown assumes **#GPUs/node = #ranks/node)** call MPI Comm rank (MPI COMM WORLD, iprocw, ierr) ngpus_per_node = 4 igpu = MODULO(iprocw, ngpus per node) !\$acc set device_num(igpu)

call MPI_Comm_split_type (MPI_COMM_WORLD, MPI_COMM_TYPE_SHARED, & 0,MPI INFO NULL, comm shared, ierr) call MPI Comm size (comm shared, nprocsh, ierr) call MPI Comm rank (comm shared, iprocsh, ierr) igpu = MODULO(iprocsh, nprocsh) !\$acc set device_num(igpu)

Use GPU data directly with MPI calls ("CUDA-aware MPI")

!\$acc ho	st_data	use_d	evice(y)	if_pres	sent			
call M	PI_Allr	educe	(MPI_IN_	PLACE, y,	,n,MPI_	_DOUBLE,MPI_	_SUM,MPI_	_C
!\$acc en	d host_	data						



OMM WORLD, ierr)

OpenACC Implementation: Effort Summary

OpenACC comment lines added

<2%

Total modified lines of code

<5%



Total lines in ori Total lines in ac Total !\$acc/!\$ac Total modified

Factors to consider:

- Optional CPU code simplifications
- Some CPU changes are temporary compiler bug work-arounds, or waiting for future OpenACC features
- (zero-beta only!)



iginal code	52,600
celerated code	55,460
c& lines added	776 (1.5%)
ines	2451 (4.7%)



OpenACC Implementation: Difficulties

Difficulties...

!\$acc cache(a%y(i-1:i+1))

Compiler Issues
 Documentation lag
 Implementation lag
 Bugs



>I'm sorry, I'm afraid I can't do that... yet

System issues
 Compiler licenses/updates
 Library versions and setup
 Hardware setups



TEPN709XV-R-20401x9V-2g0X7AA*,q1:00g-0_0_0000000 FAS_1vbt9+6-08-4.0omb01*090*970.0000



ac die C. 1200000 (C. 1990) MC (EDpary 0000000000) Sale (D. 1990) C. 1990) C. 1990 C. 19

10601b/c10010111

-IvxEoxC: Uhi() =1

#20411.

- 15-300%ap (1658 01p5*-y., 0La0.045.2.5 5.-056) 102580 (0550%AE55) - 155--- 070000 (*X'00-\$NE9* 8216X*01-p5*05*08.6-0

Timing Procedures

- "Time-to-solution" Includes I/O, comm, setup, etc. (Queue times excluded, but important!)
- We use best available compiler, compiler version, instruction sets, library versions, and *algorithm* for each hardware

Why is this fair?

We're not benchmarking hardware

Want to test the maximum "effective" performance on each system for solving our problem, using our code









Hardware and Environments

	NASA NAS Pleiades & Electra				Local Workstation	Local Desktop		
Compiler		Int	GNU 5.4.0					
MPI Library		SG	il MPT 2.15r20)		OpenMPI 1.10.2		
Family	Sandy Bridge	Ivy Bridge	Haswell Broadwell Skylake		Haswell	Broadwell		
Instruction Set	AV	AVX		AVX2 AVX		AVX2		
Model	E5-2670	E5-2680v2	E5-2680v3	E5-2680v4	Gold 6148	E5-2680v3	E5-1650v4	
Clock Rate	2.6 GHz	2.8 GHz	2.5 GHz	2.4 GHz	2.4 GHz	2.5 GHz	3.6 GHz	
#Sockets x #Cores	2 x8	2 ×10	2 x12	2 x14	2 x20	2 x12	1 x6	
Total Mem Bandwidth	51.2 GB/s	59.7 GB/s	68 GB/s	76.8 GB/s	128 GB/s	68 GB/s	76.8 GB/s	

	NVIDIA PSG	SDSC Comet	Local Desktop	Compiler Flags:
Compiler	PGI 18.3	PGI 18.4		Intel (CPU): -0
MPI Library	OpenMPI 1.10.7	OpenM	PI 2.1.2	-f
CUDA Library		CUDA 9.1		- X
Driver Version	396.26	367.48	396.26	
# GPUs x Model	4x V100	4x P100	1x TitanXP	GNU (CPU): -03
Clock Rate	1.38 GHz	1.33 GHz	1.58 GHz	
# CUDA DP Cores/GPU	2560	1792	120	PGI (GPU): -03
Mem Bandwidth/GPU	900 GB/s	732 GB/s	547.6 GB/s	-ta=



Hast of the second seco

3 -heap-arrays p-model precise CORE_AVX#

-mtune=native

tesla:cuda9.1,cc##

Timing Results



Timing Results "In-house" Single Server



Timing Results "In-house" Single Desktop



(est) ~1 hour <1.5x Cost >10x Speed



~\$9000

Wall Clock:

Alternative Algorithms: Super Time-Stepping

- Want vectorizable PC as good as PC2 in reducing Ψ iterations
- Geometric/algebraic multigrid attractive choice but Ψ requires massive code changes
- At ASTRONUM 2016 we tested RKL2 Super Time-Stepping (STS) (Meyers et al 2014) in MAS as an alternative to PCG for viscosity
- Performance of the STS method was great, but had accuracy issues
- Since the STS algorithm is highly vectorizable, its worth testing an OpenACC implementation for the current problem (where viscosity is most timeconsuming)





Alternative Algorithms: Super Time-Stepping





Alternative Algorithms: Super Time-Stepping





CPU **STS** exhibits better scaling, but similar run times to PC2

GPU **STS** ~ twice as fast as **PC1**, but similar scaling

Performance Summary of Equivalent Wall Time





24 cores





24 cores



24 cores

Summary and Outlook

- For this run (representative of many similar cases), we can move from HPC cluster to "in-house"
- Future improvements
 - Vectorizable Preconditioners
 - PC2 with single-precision
 - Make STS method accuracyrobust
- Next steps
 - Heliospheric runs(PC1 faster than PC2 on CPU!)
 - Thermodynamic (coronal) runs (on GPU-cluster like Summit)





Heliospheric CME Simulation TI



Thermodynamic CME Simulation

Questions?

penACC User Group Directives for Accelerators —> More Science. Less programming

> Twitter @OpenACCorg Facebook @OpenACCorg LinkedIn OpenACC Developers

This work was supported by - NSF's Frontiers in Earth System Dynamics - NASA's Living with a Star program - Air Force Office of Scientific Research Computer allocations provided by - NASA NAS (Pleiades/Electra) - NVIDIA Cooperation (PSG) - XSEDE/SDSC (Comet)







Extreme Science and Engineering Discovery Environment



Predictive Science Inc.

Contact: caplanr@predsci.com

Slides available at: predsci.com/~caplanr

