BLUE WATERS SUSTAINED PETASCALE COMPUTING

Rethinking Computer Architecture for Energy Limited Computing

Wen-mei Hwu ECE, CS, PCI, NCSA University of Illinois at Urbana-Champaign





GREAT LAKES CONSORTIUM







Agenda

- Blue Waters and recent progress in petascale GPU computing
- Important lessons and new innovations in energylimited exascale computing
 - System Architecture
 - Programming Systems
- Conclusion and discussions





A major shift of paradigm

- In the 20th Century, we were able to understand, design, and manufacture what we can *measure*
 - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, understand natural processes, control artificial processes...
- In the 21st Century, we are able to understand, design, create what we can *compute*
 - Computational models are allowing us to see even farther, going back and forth in time, relate better, test hypothesis that cannot be verified any other way, create safe artificial processes





CRAY

Examples of Paradigm Shift

20th Century

- Small mask patterns and short light waves
- Electronic microscope and Crystallography with computational image processing
- Anatomic imaging with computational image processing
- Teleconference

21st Century

- Computational optical proximity correction
- Computational microscope with initial conditions from Crystallography
- Metabolic imaging sees disease before visible anatomic change
- Tele-emersion



Blue Waters Science Breakthrough Example

- Determination of the structure of the HIV capsid at atomic-level
- Collaborative effort of experimental groups, at the U. of Pittsburgh and Vanderbilt U., and the Schulten's computational team at the U. of Illinois.
- 64-million-atom HIV capsid simulation of the process through which the capsid disassembles, releasing its genetic material, is a critical step in HIV infection and a potential target for antiviral drugs.





Blue Waters - Operational at Illinois since 3/2013



13.2 PF 1.6 PB DRAM \$250M

120+ Gb/sec



WAN



LAKES CONSORTIUM





Cray XK7 Nodes



Blue Waters contains 4,224 Cray XK7 compute nodes.

- Dual-socket Node
 - One AMD Interlagos chip
 - 8 core modules, 32 threads
 - 156.5 GFs peak performance
 - 32 GBs memory
 - 51 GB/s bandwidth
 - One NVIDIA Kepler chip
 - 1.3 TFs peak performance
 - 6 GBs GDDR5 memory
 - 250 GB/sec bandwidth
 - Gemini Interconnect
 - Same as XE6 nodes





Cray XE6 Nodes



Blue Waters contains 22,640 Cray XE6 compute nodes.

- Dual-socket Node
 - Two AMD Interlagos chips
 - 16 core modules, 64 threads
 - 313 GFs peak performance
 - 64 GBs memory
 - 102 GB/sec memory bandwidth
 - Gemini Interconnect
 - Router chip & network interface
 - Injection Bandwidth (peak)
 - 9.6 GB/sec per direction







LESSON #1

Current Generation of GPUs are making narrow, deep impact on applications



Science Area	Number of Teams	Codes	Struct Grids	Unstruct Grids	Dense Matrix	Sparse Matrix	N- Body	Monte Carlo	FFT	PIC	Sig I/O
Climate and Weather	3	CESM, GCRM, CM1/WRF, HOMME	Х	Х		Х		X			X
Plasmas/ Magnetosphere	2	H3D(M),VPIC, OSIRIS, Magtail/UPIC	Х				х		Х		x
Stellar Atmospheres and Supernovae	5	PPM, MAESTRO, CASTRO, SEDONA, ChaNGa, MS-FLUKSS	X			x	X	X		Х	Х
Cosmology	2	Enzo, pGADGET	Х			Х	Х				
Combustion/ Turbulence	2	PSDNS, DISTUF	х						Х		
General Relativity	2	Cactus, Harm3D, LazEV	х			Х					
Molecular Dynamics	4	AMBER, Gromacs, NAMD, LAMMPS				x	X		Х		
Quantum Chemistry	2	SIAL, GAMESS , NWChem			Х	x	X	X			X
Material Science	3	NEMOS, OMEN, GW, QMCPACK			Х	Х	Х	Х			
Earthquakes/ Seismology	2	AWP-ODC, HERCULES, PLSQR, SPECFEM3D	X	Х			X				x
Quantum Chromo Dynamics	1	Chroma, MILC, USQCD	х		Х	Х					
Social Networks	1	EPISIMDEMICS									
Evolution	1	Eve									
Engineering/System of Systems	1	GRIPS,Revisit						х			
Computer Science	1			X	х	х			Х		Х





Initial Production Use Results

- NAMD
 - 100 million atom benchmark with Langevin dynamics and PME once every 4 steps, from launch to finish, all I/O included
 - 768 nodes, Kepler+Interlagos is 3.9X faster over Interlagos-only
 - 768 nodes, XK7 is 1.8X XE6
- Chroma
 - Lattice QCD parameters: grid size of 48³ x 512 running at the physical values of the quark masses
 - 768 nodes, Kepler+Interlagos is 4.9X faster over Interlagos-only
 - 768 nodes, XK7 is 2.4X XE6
- QMCPACK
 - Full run Graphite 4x4x1 (256 electrons), QMC followed by VMC
 - 700 nodes, Kepler+Interlagos is 4.9X faster over Interlagos-only
 - 700 nodes, XK7 is 2.7X XE6







LESSON #2

Production GPU code often involve complex tradeoffs.





Tridiagonal Solver

- Implicit finite difference methods, cubic spline interpolation, pre-conditioners
- An algorithm to find a solution of Ax = d, where A is an nby-n tridiagonal matrix and d is an n-element vector









Efficient Storage Format for A

• Stored as 3 1D arrays

- No need for column indices
- No need for row pointers



GPU Tridiagonal System Solver Building Blocks

• Thomas (sequential)



• Cyclic Reduction (1 step)

- Hybrid Methods
 - PCR-Thomas (Kim 2011, Davidson 2011)
 - PCR-CR (CUSPARSE 2012)
 - Etc.



CUSPARSE is supported by NVIDIA







Runtime of solving an 8M-row matrix







NCSA



Relative Backward Error

Matrix type	SPIKE-diag_pivoting	SPIKE-Thoma	CUSPARSE	MKL	ntel SPIKE	Matlab
1	1.82E-14	1.97E-14	7.14E-12	1.88E-14	1.39E-15	1.96E-14
2	1.27E-16	1.27E-16	1.69E-16	1.03E-16	1.02E-16	1.03E-16
3	1.55E-16	1.52E-16	2.57E-16	1.35E-16	1.29E-16	1.35E-16
4	1.37E-14	1.22E-14	1.39E-12	3.10E-15	1.69E-15	2.78E-15
5	1.07E-14	1.13E-14	1.82E-14	1.56E-14	4.62E-15	2.93E-14
6	1.05E-16	1.06E-16	1.57E-16	9.34E-17	9.51E-17	9.34E-17
7	2.42E-16	2.46E-16	5.13E-16	2.52E-16	2.55E-16	2.27E-16
8	2.14E-04	2.14E-04	1.50E+10	3.76E-04	2.32E-16	2.14E-04
9	2.32E-05	3.90E-04	1.93E+08	3.15E-05	9.07E-16	1.19E-05
10	4.27E-05	4.83E-05	2.74E+05	3.21E-05	4.72E-16	3.21E-05
11	7.52E-04	6.59E-02	4.54E+11	2.99E-04	2.20E-15	2.28E-04
12	5.58E-05	7.95E-05	5.55E-04	2.24E-05	5.52E-05	2.24E-05
13	5.51E-01	5.45E-01	1.12E+16	3.34E-01	3.92E-15	3.08E-01
14	2.86E+49	4.49E+49	2.92E+51	1.77E+48	3.86E+54	1.77E+48
15	2.09E+60	Nan	Nan	1.47E+59	Fail	3.69E+58
16	Inf	Nan	Nan	Inf	Fail	4.68E+171





ĮΑ

, F_i

A

Partitioning

B

I F

C_i

 SPIKE (Polizzi et al), diagonal pivoting for numerical stability



SX = Y(5) DY = F(6) AiYi = Fi(7)







Fast Transposition on GPU



Chang, et al, SC 2012





Dynamic Tiling

Runtime







I

NESA

GREAT

LAKES CONSORTIUM

CB.

Matrix type SF	IKE-diag_pivoti	ា៖ SPIKE-Thomas	CUSPARSE	MKL	ntel SPIKE	Matlab
1	1.82E-14	1.97E-14	7.14E-12	1.88E-14	1.39E-15	1.96E-14
2	1.27E-16	1.27E-16	1.69E-16	1.03E-16	1.02E-16	1.03E-16
3	1.55E-16	1.52E-16	2.57E-16	1.35E-16	1.29E-16	1.35E-16
4	1.37E-14	1.22E-14	1.39E-12	3.10E-15	1.69E-15	2.78E-15
5	1.07E-14	1.13E-14	1.82E-14	1.56E-14	4.62E-15	2.93E-14
6	1.05E-16	1.06E-16	1.57E-16	9.34E-17	9.51E-17	9.34E-17
7	2.42E-16	2.46E-16	5.13E-16	2.52E-16	2.55E-16	2.27E-16
8	2.14E-04	2.14E-04	1.50E+10	3.76E-04	2.32E-16	2.14E-04
9	2.32E-05	3.90E-04	1.93E+08	3.15E-05	9.07E-16	1.19E-05
10	4.27E-05	4.83E-05	2.74E+05	3.21E-05	4.72E-16	3.21E-05
11	7.52E-04	6.59E-02	4.54E+11	2.99E-04	2.20E-15	2.28E-04
12	5.58E-05	7.95E-05	5.55E-04	2.24E-05	5.52E-05	2.24E-05
13	5.51E-01	5.45E-01	1.12E+16	3.34E-01	3.92E-15	3.08E-01
14	2.86E+49	4.49E+49	2.92E+51	1.77E+48	3.86E+54	1.77E+48
15	2.09E+60	Nan	Nan	1.47E+59	Fail	3.69E+58
16	Inf	Nan	Nan	Inf	Fail	4.68E+171

Chang, et al, SC 2012





GPU Performance Advantage

Runtime of solving an 8M matrix



Chang, et al, SC 2012







LESSON #3

Accelerators as I/O devices limit their effectiveness

BLUE WATERS



GREAT LAKES CONSORTIUM

NESA





NESA

ngv

Led by mobile designs for improved power-performance.







LESSON #4

Current parallel programming interfaces are too low-level for mainstream development



















































GREAT LAKES CONSORTIUM

requires too many versions of code!









Programming in Triolet Python Nonuniform FT (real part)

$$y_i = \sum_{j=0}^{n-1} x_j \cos(r_i k_j)$$

for all $0 \le i < m$





Programming in Triolet Python Nonuniform FT (real part) Inner loop $y_i = \sum_{j=1}^{n-1} x_j \cos(r_i k_j)$ for all $0 \le i < m$

Inner loop
sum(x * cos(r*k) for (x, k) in zip(xs, ks))

j=0





$$y_i = \sum_{i=0}^{n-1} x_j \cos($$

 $(r_i k_j)$ for all $0 \le i < m$











- "map and reduce" style programming—no new paradigm to learn
- Parallel details are implicit—easy to use
- Automated data partitioning, MPI rank generation, MPI messaging, OpenMP, GPU kernel generation, data layout transformation, etc.





Library-Driven Optimization

- The basic idea:
 - Compilers are smart (inlining, unboxing, ...)
 - Software abstractions have no overhead when the compiler can "evaluate" them statically
 - Just write components as library code!
 - Embody optimizations in the library code as alternative function implementations, use auto-tuning to select a final arrangement
- Triolet uses this approach [Rodrigues, et al '14]
 - Prior work loop fusion, shared-memory parallelism, vectorization [Coutts et al. '07, Keller et al. '10, Mainland et al. '13]





sum(map(square, range(k)))







Triolet

- ys = [sum(x * cos(r*k) for (x, k) in zip(xs, ks))]for r in par(rs)]
- Library functions factor out data decomposition, lacksquareparallelism, and communication

128-way Speedup	Triolet	C with MPI+OpenMP
(10 cores × 8 noues)	99	115



!/01'2)%%'+"345/01'26//'7689:,!;%&"'#&()*+<. !/01'2)%%'(=#>5/01'26//'7689:,!;%&"'"-<. !*)#+\$!"#\$!(=#>?!@!%&"'"-!@@!?.

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

!/01'A*=+\$5;+"34'B,!C,!/01'1DE,!?,!/01'26//'7689:<. !/01'A*=+\$5;+"34'>,!C,!/01'1DE,!?,!/01'26//'7689:<

!*)#+\$!"#\$!*FG#>'+"34'B!@!*4"H-"I5+"34'B,!%&"'#&()*+<. !*)#+\$!"#\$!&=--4-'+"34'B!@!*FG#>'+"34'B!J!%&"'#&()*+.

!KH)=\$!J>+,!JB+. !"K!5(=#>?<!L !!!>+!@!"#&G\$'>+. !!!B+!@!"#&G\$'B+.

14H+41L !!!>+!@!%=HH)*5+"34'>!J!+"34)K5KH)=\$<<. !!!B+!@!%=HH)*5+"34'>!J!+"34)K5KH)=\$<<.</pre>

!KH)=\$!J(+'*FG#>!@!%=HH)*5*FG#>'+"34'B!J!+"34)K5KH)=\$<<. !KH)=\$!JN+'*FG#>!@!%=HH)*5*FG#>'+"34'B!J!+"34)K5KH)=\$<<.

!"K!5(=#>?<!L

!!!"#\$!#0)(>4(+!@!%&"'#&()*+PC. !!!/01'84QG4+\$!J(4Q+!@!%=HH)*5#O)(>4(+!J!R!J!+"34)K5/01'84QG4+\$<<. 111"#\$10. !!!K)(!50!@!?.!0!S!#0)(>4(+.!OTT<!L !!!!!"#\$!0)(>4('"-!@!OTC. !!!!!/01'1+4#-5>+,!+"34'>,!/01'U96VE,!0)(>4('"-, !!!!!!!!!!!!!!?,!/01'26//'7689:,!;(4Q+W0X<.

!!!!!/01'1+4#-5B+, !+"34'>, !/01'U96VE, !0)(>4('"-, !!!!!!!!!!!!!!!?, !/01'26//'7689:, !; (40+W#0)(>4(+TOX<.</pre> !!!!!/01'1+4#-5(+!T!0)(>4(''-)*F6#>'+"34'B,!*F6#>'+"34'B,!*O1'U96VE,!0)(>4(''-, !!!!!!!!!!!!!!!!!!?,!/01'26//'7689:,!;(40+WYI#0)(>4(+TOX<.</pre>

!!!%4%*&N5(+'*FG#>,!(+,!*FG#>'+"34'B!J!+"34)K5KH)=\$<<.

!!!/01'7="\$=HH5#0)(>4(+JR,!(4Q+,!/01'ZEVE[Z\Z'1]D68\<. !!!K(445(4Q+<.

1 I M

!4H+4!L !!!/01'84*I5>+,!+"34'>,!/01'U96VE,!?, !!!!!!!!!!?,!/01'26//'7689:,!/01'ZEVE[Z'1]D68\<. !!!/01'84*I5B+,!+"34'>,!/01'U96VE,!?, !!!!!!!!!!?,!/01'26//'7689:,!/01'ZEVE[Z'1]D68\<. !!!/01'84*I5(+'*FG#>,!*FG#>'+"34'B,!/01'U96VE,!?, !!!!!!!!!!?,!/01'26//'7689:,!/01'ZEVE[Z'1]D68\<.</pre>

111 1111"#\$!". &(=`%=!)%&!&=(=HH4H!K)(!+*F4-GH45+\$=\$"*< !!!!K)(!5"!@!?.!"!S!*FG#>'+"34'B.!"TT<!L !!!!!!KH)=\$!+!@!?. !!!!!!"#\$!a. !!!!!!K)(!5a!@!?.!a!S!+"34'>.!aTT< !!!!!!!+!T@!B+WaX!J!*)+K5(+'*FG#>W"X!J!>+WaX<. !!!!!!N+'*FG#>W"X!@!+. LITIM

111111111111, 1/01'26//'7689:<. !K(445(+'*FG#><. !K(445N+'*FG#><. !"K!5(=#>?<!L !!!K(445(+<. . H+4!I !!!K(445B+< !!K(445>+<.

/01']=\$F4(5N+'*FG#>,!*FG#>'+"34'B,!/01'U96VE,!N+,!*FG#>'+"34'B,!/01'U96VE,

Cleanu

10x10 Project Progress



• Composite evaluation: 5x5

Compiler

GPM

Compilation: Mapreduce, micro-engine, and robust vectorization and locality management

 Global Power Management: power & performance models for core and link frequency changes + opportunity assessment





- Processor cores, accelerators, NICs, and storage controllers are the new data path components in the SOC era.
- Impactful microarchitecture research will likely need to involve OS and compiler innovations.
 - Java Bytecode is arguably the most important ISA
 - OS and VM implementations are long overdue for revamping.
- Performance/efficiency of many applications can be improved by ~100x by squeezing out inefficiency from the implementation stack.
 - Will likely take a decade or more to materialize in a scaling manner
 - Much of the progress will need microarchitecture support
 - Be radical!





- Blue Waters and related HPC systems have demonstrated that GPU computing can have narrow but deep impact on science
- New architectural innovations enable much easier work migration between CPUs and accelerators
- New programming systems innovations enable much faster application development and lowered maintenance cost





- D. August (Princeton), S. Baghsorkhi (Illinois), N. Bell (NVIDIA), D. Callahan (Microsoft), A. Chien (U. Chicago), J. Cohen (NVIDIA), B. Dally (Stanford), J. Demmel (Berkeley), P. Dubey (Intel), M. Frank (Intel), M. Garland (NVIDIA), Isaac Gelado (BSC), M. Gschwind (IBM), R. Hank (Google), J. Hennessy (Stanford), P. Hanrahan (Stanford), M. Houston (AMD), T. Huang (Illinois), D. Kaeli (NEU), K. Keutzer (Berkeley), W. Kramer (NCSA), I. Gelado (NVIDIA), B. Gropp (Illinois), D. Kirk (NVIDIA), D. Kuck (Intel), S. Mahlke (Michigan), T. Mattson (Intel), N. Navarro (UPC), J. Owens (Davis), D. Padua (Illinois), S. Patel (Illinois), Y. Patt (Texas), D. Patterson (Berkeley), C. Rodrigues (Illinois), P. Rogers (AMD), S. Ryoo (ZeroSoft), B. Sander (AMD), K. Schulten (Illinois), B. Smith (Microsoft), M. Snir (Illinois), I. Sung (Illinois), P. Stenstrom (Chalmers), J. Stone (Illinois), S. Stone (Harvard) J. Stratton (Illinois), H. Takizawa (Tohoku), M. Valero (UPC)
- And many others!





Two Petascale Computing Systems

I

System Attribute	NCSA Blue Waters	ORNL Titan
Vendors	Cray/AMD/NVIDIA	Cray/AMD/NVIDIA
Processors	Interlagos/Kepler	Interlagos/Kepler
Total Peak Performance (PF)	13.2	27.1
Total Peak Performance (CPU/GPU)	7.7/5.5	2.6/24.5
Number of CPU Chips	49,504	18,688
Number of GPU Chips	4,224	18,688
Amount of CPU Memory (TB)	1600	584
Interconnect	3D Torus	3D Torus
Amount of On-line Disk Storage (PB)	26	13.6
Sustained Disk Transfer (TB/sec)	>1	0.4-0.7
Amount of Archival Storage	300	15-30
Sustained Tape Transfer (GB/sec)	100	7