

#### **Towards Petascale Computing for Science**

#### Horst Simon Lenny Oliker, David Skinner, and Erich Strohmaier Lawrence Berkeley National Laboratory

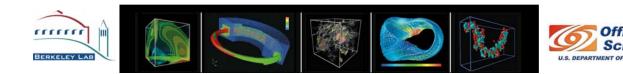
The Salishan Conference on High-Speed Computing

April 19, 2005



# Outline

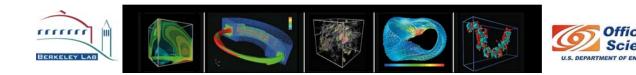
- Science Driven Architecture
- Performance on today's (2004 2005) platforms
- Challenges with scaling to the Petaflop/s level
- Two tools that can help: IPM and APEX/MAP



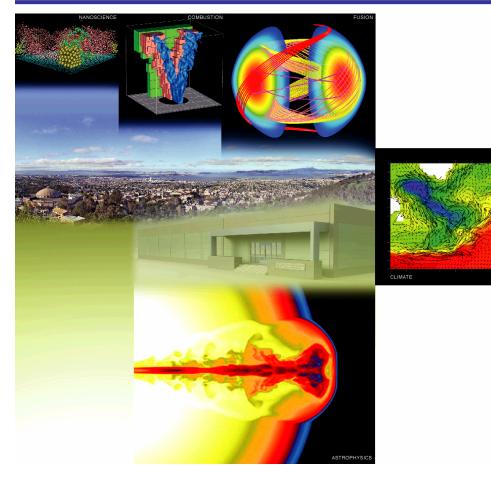
## Scientific Applications and Underlying Algorithms Drive Architectural Design

- 50 Tflop/s 100 Tflop/s **sustained** performance on applications of national importance
- Process:
  - identify applications
  - identify computational methods used in these applications
  - identify architectural features most important for performance of these computational methods

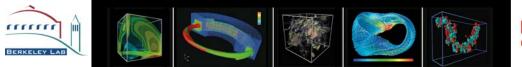
**Reference:** Creating Science-Driven Computer Architecture: A New Path to Scientific Leadership, (Horst D. Simon, C. William McCurdy, William T.C. Kramer, Rick Stevens, Mike McCoy, Mark Seager, Thomas Zacharia, Jeff Nichols, Ray Bair, Scott Studham, William Camp, Robert Leland, John Morrison, Bill Feiereisen), Report LBNL-52713, May 2003. (see www.nersc.gov/news/reports/HECRTF-V4-2003.pdf)



# Capability Computing Applications in DOE/SC



- Accelerator modeling
- Astrophysics
- Biology
- Chemistry
- Climate and Earth Science
- Combustion
- Materials and Nanoscience
- Plasma Science/Fusion
- QCD
- Subsurface Transport



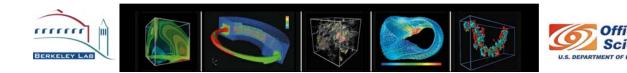


# Capability Computing Applications in DOE/SC (cont.)

These applications and their computing needs have been well-studied in the past years:

• "A Science-Based Case for Large-scale Simulation", David Keyes, Sept. 2004 (http://www.pnl.gov/scales).

• "Validating DOE's Office of Science "Capability" Computing Needs", E. Barsis, P. Mattern, W. Camp, R. Leland, SAND2004-3244, July 2004.



#### Science Breakthroughs Enabled by Leadership Computing Capability

Science Areas	Goals	Computational Methods	Breakthrough Target (50-100 Tflop/s)
Nanoscience	Simulate the synthesis and predict the properties of multi-component nanosystems	Quantum molecular dynamics Quantum Monte Carlo Iterative eigensolvers Dense linear algebra Parallel 3D FFTs	Simulate nanostructures with hundreds to thousands of atoms as well as transport and optical properties and other parameters
Combustion	Predict combustion processes to provide efficient, clean and sustainable energy	Explicit finite difference Implicit finite difference Zero-dimensional physics Adaptive mesh refinement Lagrangian particle methods	Simulate laboratory scale flames with high fidelity representations of governing physical processes
Fusion	Understand high-energy density plasmas and develop an integrated simulation of a fusion reactor	Multi-physics, multi-scale Particle methods Regular and irregular access Nonlinear solvers Adaptive mesh refinement	Simulate the ITER reactor
Climate	Accurately detect and attribute climate change, predict future climate and engineer mitigation strategies	Finite difference methods FFTs Regular and irregular access Simulation ensembles	Perform a full ocean/ atmosphere climate model with 0.125 degree spacing, with an ensemble of 8-10 runs
Astrophysics	Determine through simulations and analysis of observational data the origin, evolution and fate of the universe, the nature of matter and energy, galaxy and stellar evolutions	Multi-physics, multi-scale Dense linear algebra Parallel 3D FFTs Spherical transforms Particle methods	Simulate the explosion of a supernova with a full 3D model
			S Office of

U.S. DEPARTMENT OF ENERG



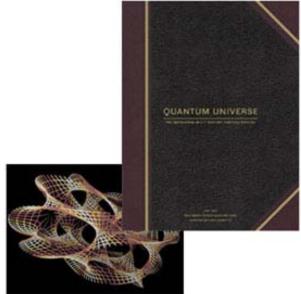
# **Opinion Slide**

One reason why we have failed so far to make a good case for increased funding in supercomputing is that we have not yet made a compelling science case.

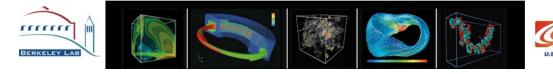
# A better example: "The Quantum Universe"

"It describes a revolution in particle physics and a quantum leap in our understanding of the mystery

and beauty of the universe."



http://interactions.org/quantumuniverse/



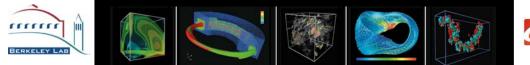
# **How Science Drives Architecture**

State-of-the-art computational science requires increasingly diverse and complex algorithms

Science Areas	Multi- Physics and Multi- Scale	Dense Linear Algebra	FFTs	Particle Methods	AMR	Data Parallelism	Irregular Control Flow
Nanoscience	X	Х	Х	Х		Х	Х
Combustion	Х			Х	Х	Х	Х
Fusion	X	Х		Х	Х	Х	Х
Climate	X		Х		Х	Х	Х
Astrophysics	Х	Х	Х	Х	Х	Х	X

Only balanced systems that can perform well on a variety of problems will meet future scientists' needs!

Data-parallel and scalar performance are both important



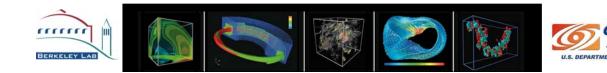


# Phil Colella's "Seven Dwarfs"

Algorithms that consume the bulk of the cycles of current high-end systems in DOE:

- Structured Grids
- Unstructured Grids
- Fast Fourier Transform
- Dense Linear Algebra
- Sparse Linear Algebra
- Particles
- Monte Carlo

(Should also include optimization / solution of nonlinear systems, which at the high end is something one uses mainly in conjunction with the other seven)

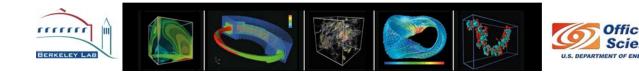


# "Evaluation of Leading Superscalar and Vector Architectures for Scientific Computations"

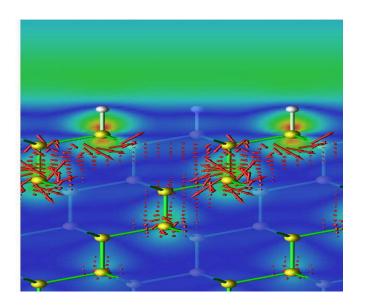
#### Leonid Oliker, Andrew Canning, Jonathan Carter LBNL

#### Stephane Ethier PPPL

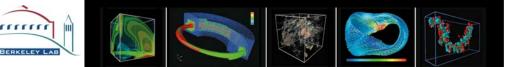
(see SC04 paper at <a href="http://crd.lbl.gov/~oliker/">http://crd.lbl.gov/~oliker/</a> )



# **Material Science: PARATEC**



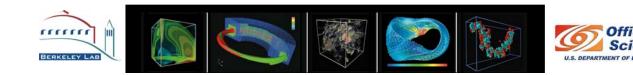
- PARATEC performs first-principles quantum mechanical total energy calculation using pseudopotentials & plane wave basis set
- Density Functional Theory to calc structure & electronic properties of new materials
- DFT calc are one of the largest consumers of supercomputer cycles in the world
- PARATEC uses all-band CG approach to obtain wavefunction of electrons
- Part of calc. in real space other in Fourier space using specialized 3D FFT to transform wavefunction
- Generally obtains high percentage of peak on different platforms
- Developed with Louie and Cohen's groups (UCB, LBNL), Raczkowski



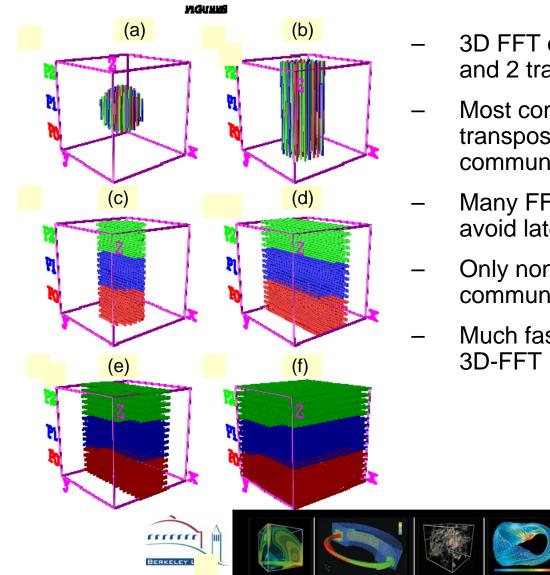


# **PARATEC: Code Details**

- Code written in F90 and MPI (~50,000 lines)
- 33% 3D FFT, 33% BLAS3, 33% Hand coded F90
- Global Communications in 3D FFT (Transpose)
- 3D FFT handwritten, minimize comms. reduce latency (written on top of vendor supplied 1D complex FFT)
- Code has setup phase then performs many (~50) CG steps to converge the charge density of the system (data on speed is for 5CG steps, does not include setup)



# **PARATEC: 3D FFT**

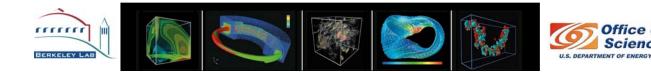


- 3D FFT done via 3 sets of 1D FFTs and 2 transposes
- Most communication in global transpose (b) to (c) little communication (d) to (e)
- Many FFTs done at the same time to avoid latency issues
- Only non-zero elements communicated/calculated
- Much faster than vendor supplied 3D-FFT



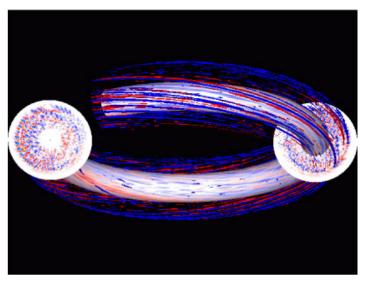
# **PARATEC: Performance**

Data		Powe	er 3	Pow	er4	Alt	ix	ES	5	<b>X</b> 1	
Data Size	P	Gflops/ P	% peak	Gflops/ P	% peak	Gflops/ P	% peak	Gflops/ P	%pea k	Gflops/ P	% peak
	32	0.95	63%	2.0	39%	3.7	62%	4.7	60%	3.0	24%
122	64	0.85	57%	1.7	33%	3.2	54%	4.7	5 <b>9</b> %	2.6	20%
432 Atom	128	0.74	49%	1.5	29%			4.7	5 <b>9</b> %	1.9	15%
Alom	256	0.57	38%	1.1	21%			4.2	52%		
	512	0.41	28%					3.4	42%		
686	128							4.9	62%	3.0	24%
Atom	256							4.6	57%	1.3	10%

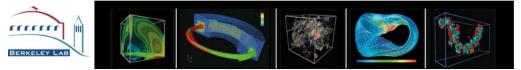


**Office of** Science

# **Magnetic Fusion: GTC**



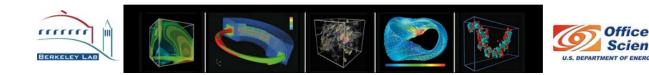
- Gyrokinetic Toroidal Code: transport of thermal energy (plasma microturbulence)
- Goal magnetic fusion is burning plasma power plant producing cleaner energy
- GTC solves gyroaveraged gyrokinetic system w/ particle-in-cell approach (PIC)
- PIC scales N instead of N<sup>2</sup> particles interact w/ electromag field on grid
- Allows solving equation of particle motion with ODEs (instead of nonlinear PDEs)
- Main computational tasks:
  - <u>Scatter</u>: deposit particle charge to nearest grid points
  - Solve the Poisson eqn to get potential at each grid point
  - Gather: Calc force on each particle based on neighbors potential
  - Move particles by solving eqn of motion along the characteristics
  - Find particles moved outside local domain and update
- Developed at Princeton Plasma Physics Laboratory, vectorized by Stephane Ethier



### **GTC: Performance**

Numbe		Powe	er 3	Pow	er4	Alt	ix	E	S	X	1
r Particl es	Ρ	Gflops /P	%pea k	Gflops/ P	%pea k	Gflops/ P	/%pea k	Gflops /P	%pea k	Gflops/ P	%pea k
10/cell	32	0.13	9%	0.29	5%	0.29	5%	1.15	14%	1.00	8%
20M	64	0.13	9%	0.32	5%	0.26	4%	1.00	13%	0.80	6%
	32	0.13	9%	0.29	5%	0.33	6%	1.62	20%	1.50	12%
100/cell	64	0.13	9%	0.29	5%	0.31	5%	1.56	20%	1.36	11%
200M	102 4	0.06	4%								

GTC is now scaling to 2048 processors on the ES for a total of 3.7 TFlops/s



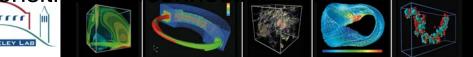
# **Issues in Applications Scaling**

#### **Applications Status in 2005**

- A few Teraflop/s sustained performance
- Scaled to 512 1024 processors

#### Applications on Petascale Systems need to deal with

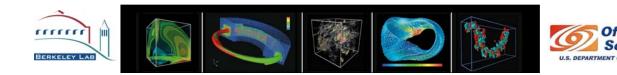
- 100,00 processors (assume nominal Petaflop/s system with 100,000 processors of 10 Gflop/s each)
- Multi-core processors
- Topology sensitive interconnection network





# Integrated Performance Monitoring (IPM)

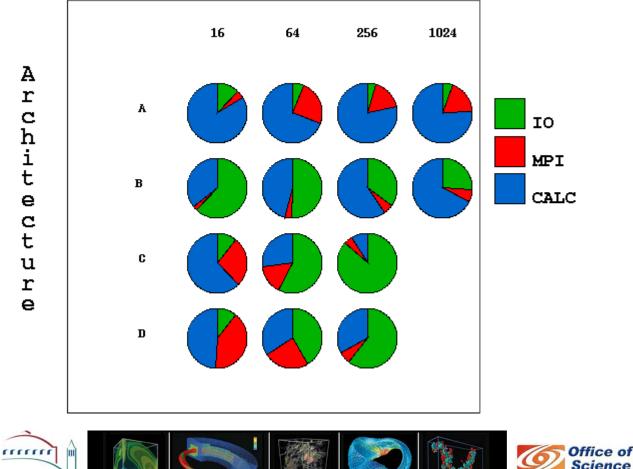
- brings together multiple sources of performance metrics into a single profile that characterizes the overall performance and resource usage of the application
- maintains low overhead by using a unique hashing approach which allows a fixed memory footprint and minimal CPU usage
- open source, relies on portable software technologies and is scalable to thousands of tasks
- developed by David Skinner at NERSC (see <a href="http://www.nersc.gov/projects/ipm/">http://www.nersc.gov/projects/ipm/</a> )



# Scaling Portability: Profoundly Interesting

A high level description of the

performance of a well known cosmology code on four well known architectures.



U.S. DEPARTMENT OF ENERG

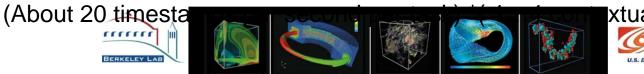
BERKELEY LA

Concurrency

# 16 Way for 4 seconds

36.2 s 36.202 s 36.204 s 36.206 s 36.208 s 36.21 s 36.212 s 36.214 s 36.216 s

Process O <mark>300 User_Code</mark>	User_Code		User_Code 300 MP1_Allreduce	
Process 1 300 User_Cod	e 300 User_Code	User_Code l	User_Code User_Code 300 MPI_Allreduce	
Process 2 User_Code User_Co	de User_Code	300 300 Uşer_Ca	ode User_Code 300. MPI_Allreduce	
Process 3 300 User_Co	de 300 300 300	300 300 300 300 300 300	300 User_Code 300 M21_Allred	uce
Process 4 Deta Conte User_C	odeUser_Cod	e 300 vs	ser_Code 7 300 User_Code 300 MP1_Allre	duce
Process 5	300 300 300	Wser_Code User_Code 300	User_Code 300 User_Code 300 MPL	Allreduce
Process 6 800 User one		800 300 <b>300 - 200</b>	User_Code 300 300 User_Code 300, MPI.	Alfreduce
Process 7 Dear Code		n code		er_Code 300
Process 8	cone aser core	Turne cone	User_Gade 1000 1000 1000 100 Allreduce	
Process 9 300		and there case were cas	e 300 300 300 300 1300 MPL_Allreduce	
Process 10 User_Code	A Stare IN	and the second	User Sole She 200 300 MP1_Allred	uce the state state of the
Process 11 User_Code	ALL HA	A BAR AND	The second secon	Ireduce
Process 12 <mark>User_Code</mark>			Code 2 Serie 2 Serie 300 MP1_Allreduce	
Process 13 <mark>User_Code</mark>			Soor Sda 1300 - 300 300 MPL Allreduc	e en la constant de la constant
Process 14 <mark>User_Code</mark>			1070 042 360 200 300 MPL_Allr	aduce
Process 15 <mark>User_Code</mark>		Der fou	User_code_3t	MPI_Allreduce
Process 16 User_Code	300	300 USE ADDE USE	Cher 300 80. 1 200 300 MPL_Allreduce	



xtual variables)

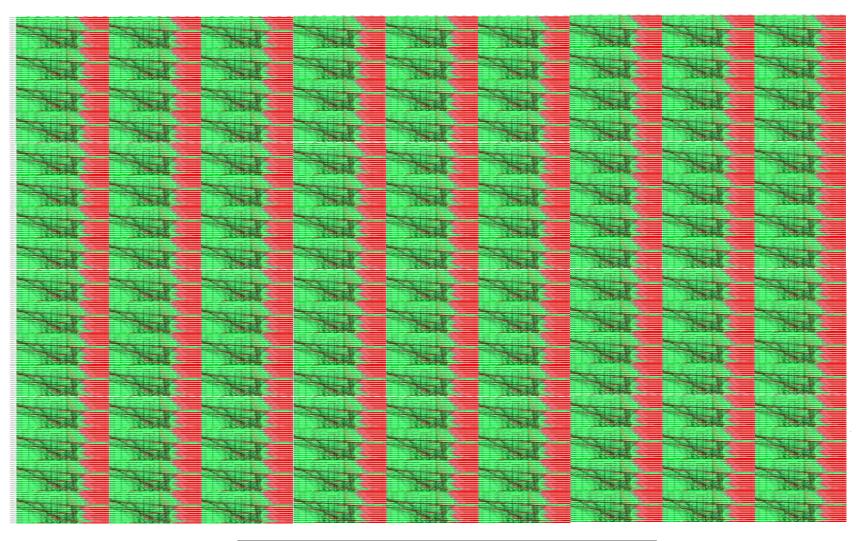
# 64 way for 12 seconds

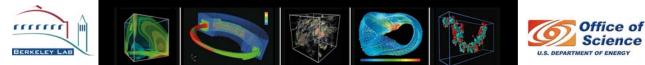
m mor, Cada mer, Cada	The Price Cade Price Cade	Party Calls The Target Street and Street	ann mar Cada Marris Cada	Barr, Cale Barr, Cale State
		International Contraction of the International Contractional Contracti		
an here the ball having the ball and a second the second second second second second second second second second	and and a set of the lost of the lost of the lost	and and an	allow and the part has been been been	when her a little war, take Bank wet, allestate
and party and the contract of the second	alla for a state of the state o	ant anti-take anti-take anti-take anti-taketer	an particular particular and antiparticular and a second s	Designer, Cafe
The second se	AND ADDRESS OF TAXABLE AND ADDRESS OF TAXABLE AND ADDRESS OF TAXABLE A	an and the second s		and provide and interiors and residence
· Property and a statement of the statem	NY ADVANCE			
	mar carls and and calls 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	A CANADA		
an contrate Walton Man an and make the state	and the second s	A CONTRACTOR OF THE REAL PROPERTY AND ADDRESS OF	Bern continues to Barliner M. House	A DESCRIPTION OF THE OWNER OWNER OF THE OWNER
				Con an and a second sec
	Anna and a second a	A CONTRACTOR OF A CONTRACTOR O		And State of the second s
		A Real Property and a second s		and the second sec
	Anna Anna Anna Anna Anna Anna Anna Anna	LE CANA DE CAMPA DE ANNO DE LA CANADA	and the second sec	
		N P S S S S S S S S S S S S S S S S S S	Annual State	Contraction of the second seco
	allering and a second sec	and the second sec		
				Constant of the second se
n par, Cale Part, Cale	and the Difference of the	Provide the second seco	and the line in the second	Martin Cale and Anticipation
n June, Cada June June, Cada June, Tana June , Cada June, Anda June , Cada		The second line set	m Dise Cada millione Cada Biar Cada	mar case have cade have strategy
er case pase case provide per provide and pase and pase case and and and and and	Plan Cold Diar, Cold Diar, Call	Into Date Directories and a linear control Date Arts Antonio a	tur, Cale Diar, Cale Diar, Cale Diar	Prise_chile
an han (en por bin boo parties (an parties (an ) ) and an interaction (and )	made a log and the log and the log matter	an anter an anter an anter and	and part the part and party and	and the second s
and any training any training and any training and any training and any training any trainin	allisters and a state of the local data	and the second sec	and the second s	
The second	MELADOREA MARKET	an anna Californi, Ipan Californi, anna Inna Californi, anna Anna Anna Anna Anna Anna Anna Ann	The second	AND THE DESCRIPTION OF THE OWNER
	and the set of the set		m leven 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	and the second se
	Bellecter Marine Marine Marine	an and a state of the state of	and and the second s	The state of the second second
		And the second sec		and an
	and a second	Property allocations		A CONTRACTOR OF A CONTRACTOR O
ar cada	Libedes and the case	A Real Property of the second se	tan, cada	and the second sec
	Anna Anna Canta	Name of the state	tor con	And a set of the set o
	effect	IN PARTY AND AND AND A PROPERTY AND AND A	Rest, Cada	The states
				A State of the second sec
				Contraction of the second se
n joser, Cade Press, Cade	me pare, Cade Bran, Cade	Bruer, Cade and Catherine	me pare, Cade pres, Cade	Burr, Cade Day, and Allena
a plase, Cada pare plase, Cada plase, Cada and Cada and Cada and Cada and Cada	me pros cade pros cade prost, cade	mart, Cale	ant place, Cado and place, Cado place, Cado	man, Caller
er, cade aner, cade peor, cade peor pier, cade energies and peor, cade	a por tade	and and plan_claim	ener, Code Door, Code Door, Code Door, Code Door, Code	Dist. Chief and the state and
en hansten perjen perjen perjen per i marten ber	anologi (anologi (ano	ant parties and a second se	200 000 000 000 000 000 000 000 000 000	and and a second and and and and and a second and a
and and the set and an	alledges and a second s	and an provide and and the and art distant	and the second s	Sector States and states
The second	Mart allowers and an and and and and and and	te une contrast proc.com per la con com any pressione		WARE DESCRIPTION OF THE OWNER
a land, but we want to a substitute the set of the state of the set of the se	Mart Allerting	and the state of the set of the s	The second	And a second sec
	mar. Colo	The color	an bereiter ihr anderen Barren	The Colo
an and a second s			and real and the second and the	
		No. 1		
				The first second s
	adapt and a second a			
	Allesters Here cate	The second se	and take	The second se
	and the statement of th		and the second sec	
	and the little and the little and the		and the loss of the second second	
mar, Cade mer, Cade mer, Cade ang Cat, Sineshot	me pror. Cade Pror. Cade	Breer, Cade ang Sert Streets C	ant pase, Cade Print, Cade	Peer, Cade and Serie Classifier
a pres, cada and pres, cada pres, cada mer, cada and cada and cada and and cada	ann prost, Cade aner, Cade aver, Cade	mar_Cade Disar_Cade Data_Sati_satis_tata	ant prost, Cade 200 prost, Cade prost, Cade	Base, Caller Baser, Caller State, Solid State
r.Cade Deser.Cade Des Des Dies Dies Deser.Cade Des Deser.Cade	Bier, Con Dier, Con	and and pier claim	Fint Cade Dient Cade Dient Cade and Dient	anter and anter and anter all and anter all and
ne aver, ces pol per pos juniter des polites aus alle ver, cete aut ave	andre a state and a state and a state and a state and	200 200 200 200 200 200 City Dog 011_Allocate	200 0107_CARE 200 200 200_200 200 2	and and a second and an and a second and a second a secon
and purpose for loss an purpose call an purpose call	COLUMN COLUMN COLUMN	ant and an and a set a	and a second and a	ann anne Cade anne Cade ann Art allerters
		Contraction and the second sec		aligned (some case) and (some case) and allowing a
				and the configuration of the Code
	and the second s	An Can	and an entry of the second second second	
	Instant Control of Con			And a second sec
		A Design of the second se		
	and and a second s		the case	
	adara ana taka		tary Cade	
	Allendary Cade	A CONTRACTOR OF A CONTRACTOR O	ter tak	Provide and the second second
	a long wet, edited as a second s			



**Office of** ience

# 256 Way for 36 Seconds

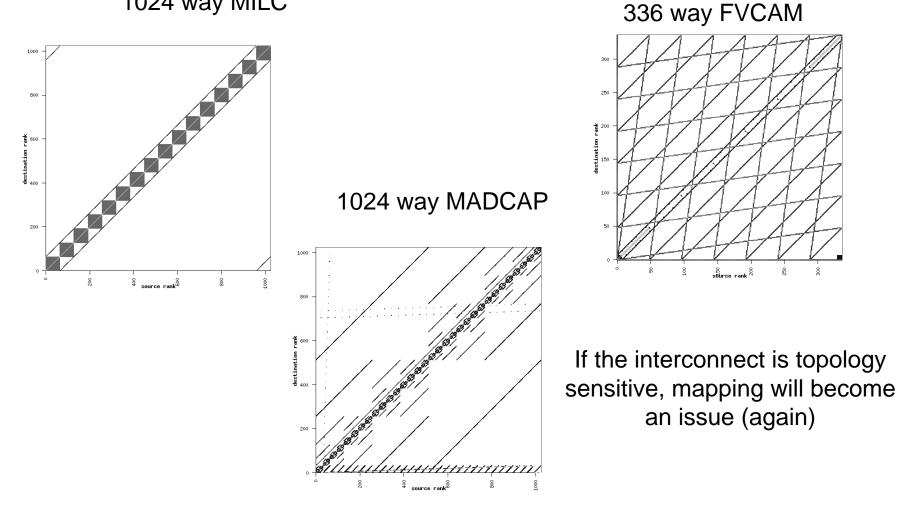


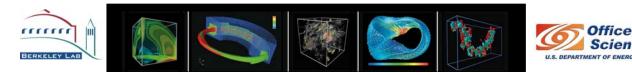


**Office of** 

# **Application Topology**

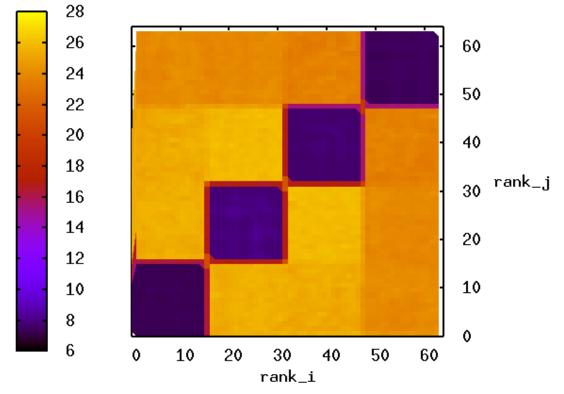
1024 way MILC





Office of

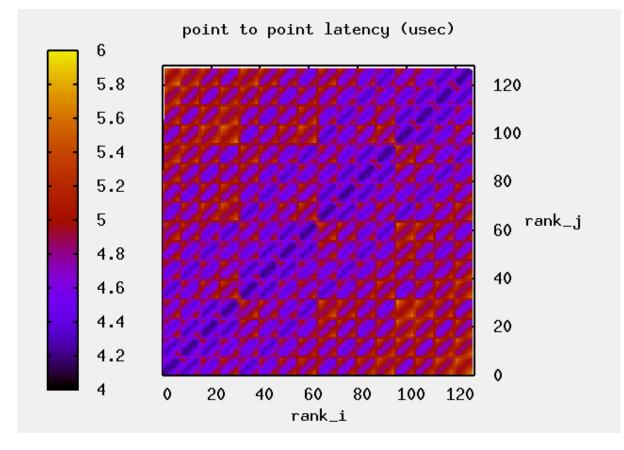
# **Interconnect Topology**

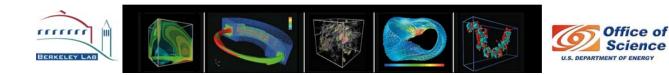


point to point latency (usec)

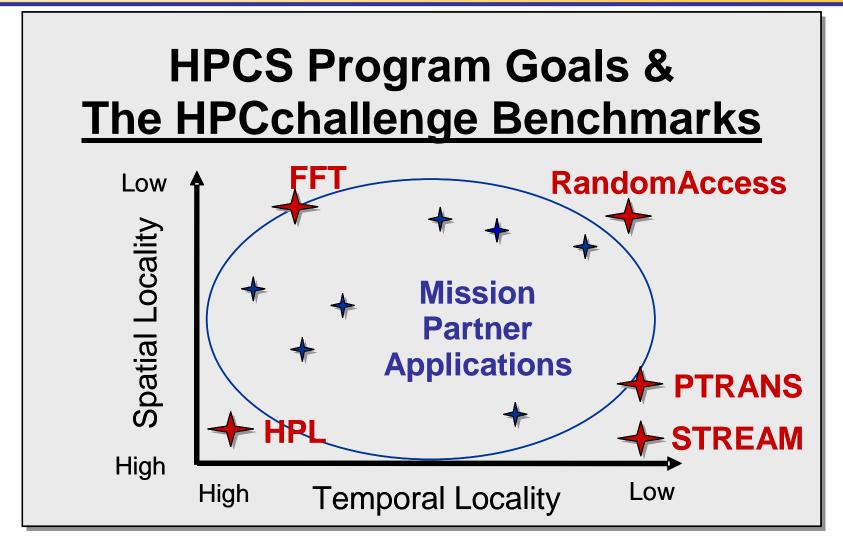


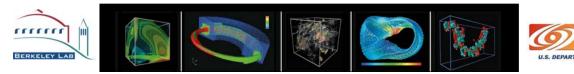
# **Interconnect Topology**





## DARPA HPCS will characterize applications

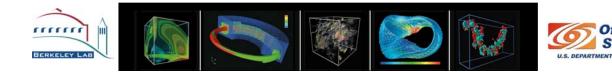




# APEX-Map: A Synthetic Benchmark to Explore the Space of Application Performances

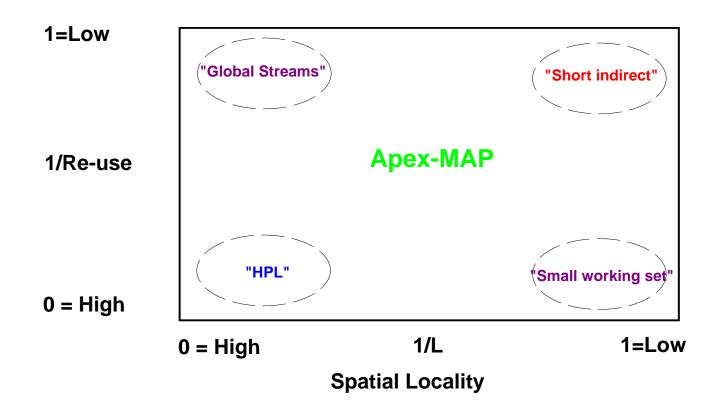
Erich Strohmaier, Hongzhang Shan Future Technology Group, LBNL EStrohmaier@lbl.gov

Co-sponsored by DOE/SC and NSA

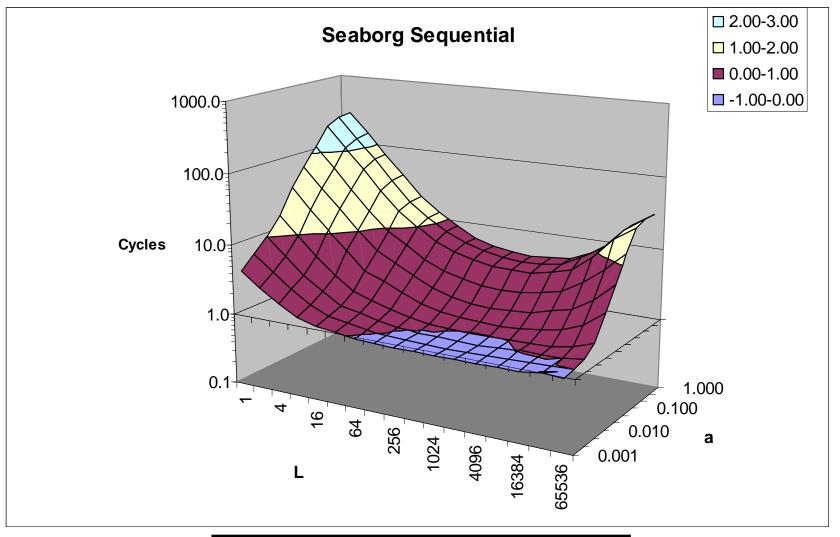


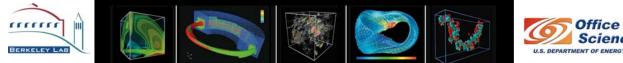
# Apex-MAP characterizes architectures through a synthetic benchmark

#### **Temporal Locality**



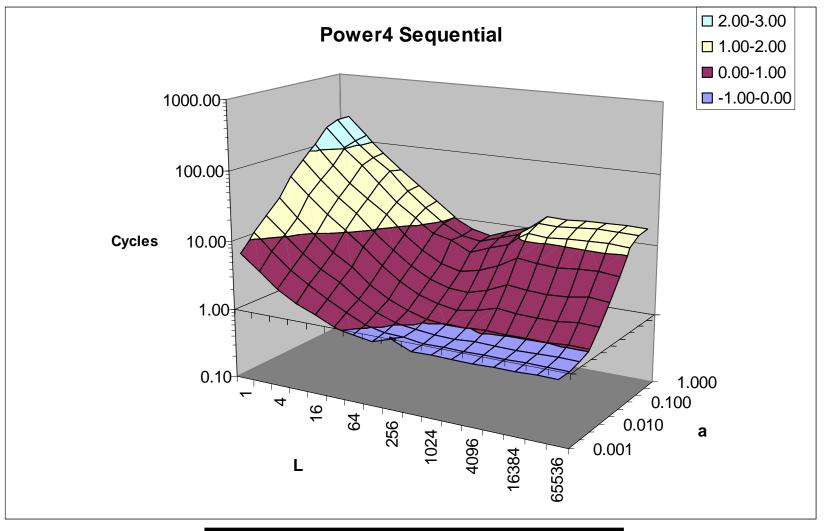


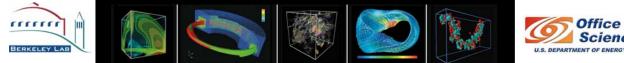




Office of

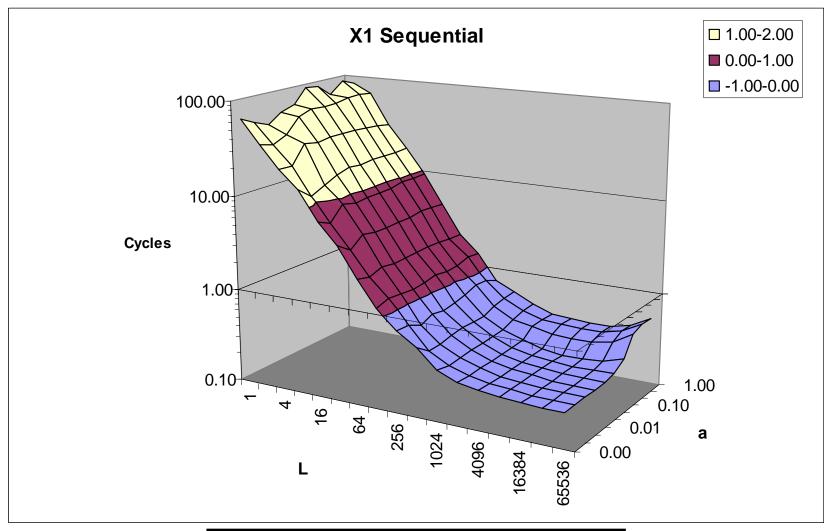
Science





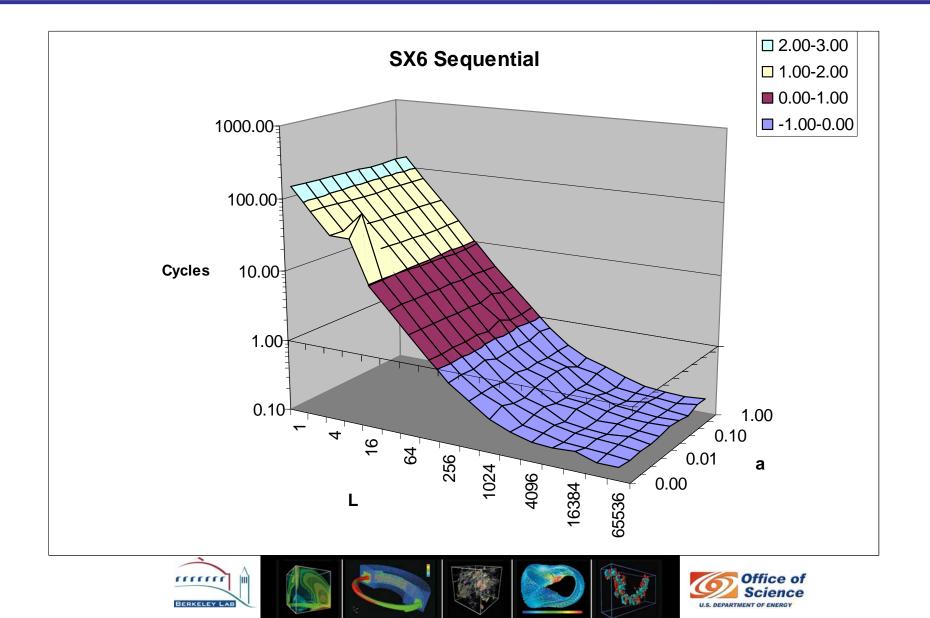
Office of

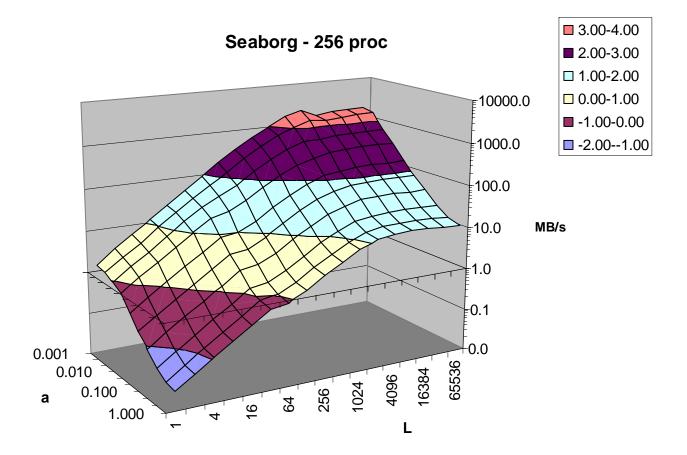
Science

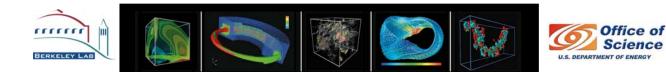


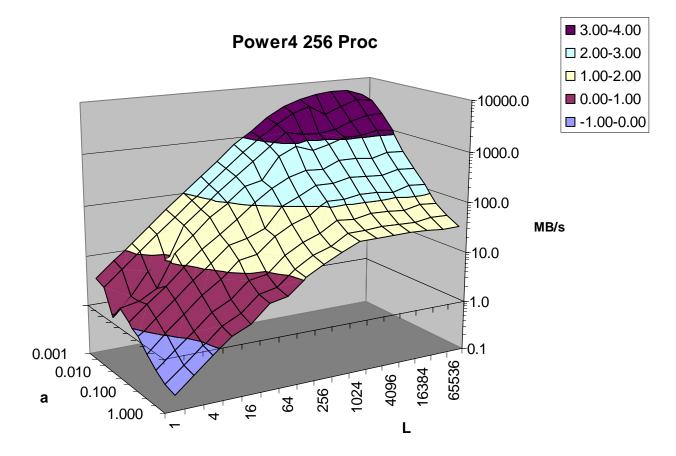


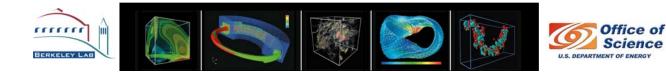
Office of Science

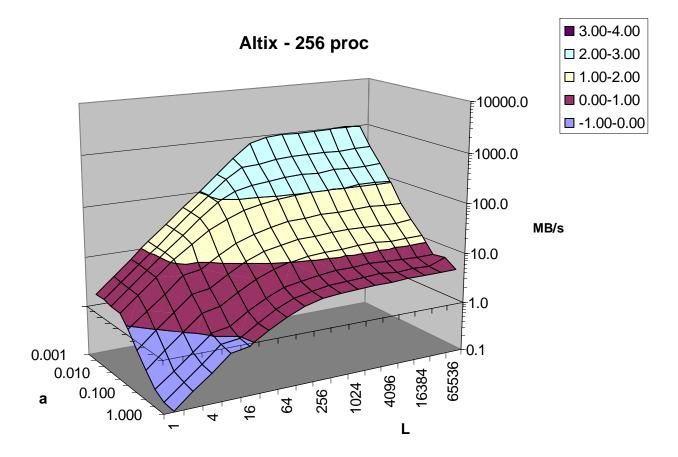


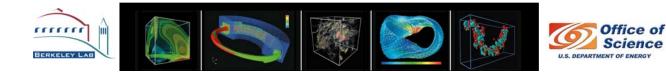


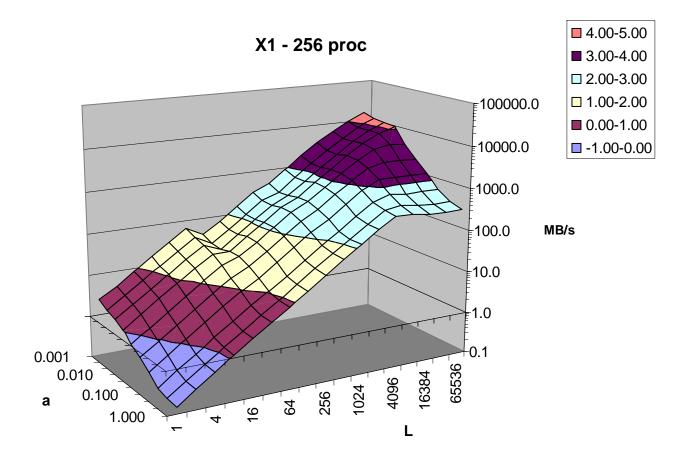


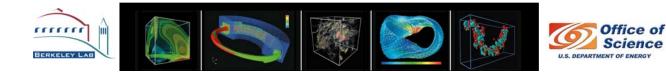


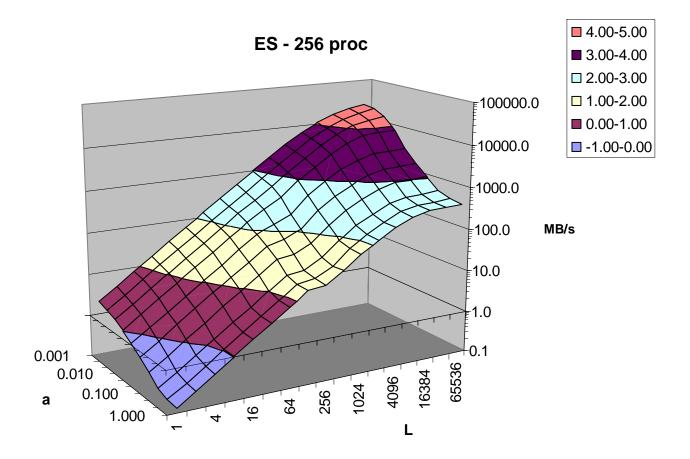


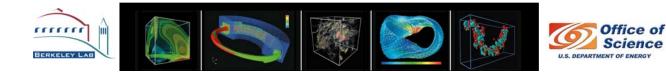












# Summary

- Three sets of tools (applications benchmarks, performance monitoring, quantitative architecture characterization) have been shown to provide critical insight into applications performance
- Need better quantitative data and measurements (like the ones discussed here) to help applications to scale to the next generation of platforms

