SAND2015-3191C

On the performance of fully-coupled algebraic multigrid preconditioning at large-scale: application to FEM CFD/MHD

PANACM 2015

Buenos Aires, Argentina

April 27, 2015

Paul Lin, John N. Shadid, Jonathan J. Hu, Eric C. Cyr, Roger P. Pawlowski, Andrey Prokopenko

Sandia National Labs

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Big picture overview

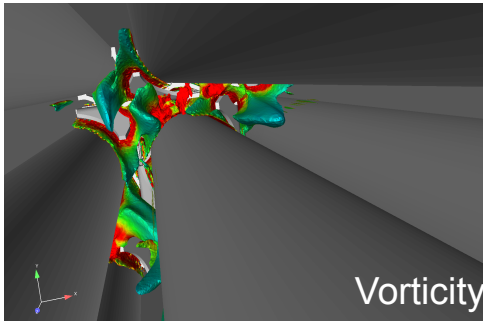
- SNL need for high-fidelity solutions with complex physics
- SNL has many huge production application codes
- Challenge: get apps to run efficiently on future platforms
- Includes FEM apps on unstructured meshes
  - Many rely on multigrid for performance and scalability
  - Algebraic multigrid (AMG) on unstructured meshes has additional challenges compared with geometric multigrid
- Many apps at SNL and outside rely on “classic” Trilinos (Epetra): 32-bit limit, no path forward for future arch
- Next-generation “modern” Trilinos (Tpetra): templated data types, potential path forward for future architectures

Focus of this talk

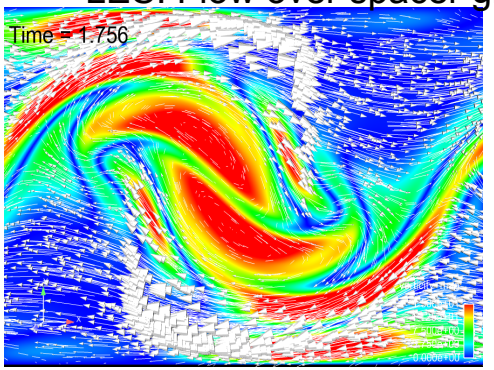
- Talk will focus on one SNL CFD/MHD FEM unstructured mesh application code (Drekar) that relies on the Trilinos solver stack and AMG for performance and scaling (Newton-Krylov solver)
  - Work-in-progress talk (more questions than answers)
  - Issues apply to other app codes that use Trilinos with AMG
- App migrated from classic to modern Trilinos
  - Perform simulations with > 2 billion unknowns
  - App is still flat MPI-only
  - Still many challenges
    - Evaluate and improve algorithmic scaling
    - Evaluate and improve multigrid setup scaling
- Possible path forward to handle future architectures
  - brief discussion: matrix assembly, linear solve (work from other SNL colleagues)
- Future work (lots of it)

Drekar

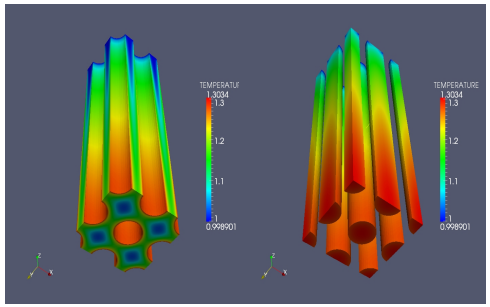
(J. Shadid, R. Pawlowski, E. Cyr, T. Smith, T. Wildey)



LES: Flow over spacer grid



MHD: Hydromagnetic Kelvin-Helmholtz



Conjugate Heat Transfer

- Scalable parallel implicit/IMEX FE code
- Includes: Navier-Stokes, MHD, LES, RANS
- Architecture admits new coupled physics
- Support for advanced discretizations
  - mixed, physics compatible and high-order basis functions
  - multi-physics capable (conjugate heat transfer)
- Advanced UQ tools/techniques
  - Adjoint based sensitivities and error-estimates
- Advanced solution methods
  - Parallel solvers from SNL's Trilinos framework
  - Physics-based preconditioning
  - Fully-coupled multigrid for monolithic systems

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories

Brief classic Trilinos overview

Object-oriented software framework for the solution of large-scale, complex multi-physics engineering and scientific problems

- nonlinear/linear solvers, discretization, optimization, load balance, I/O, etc. (Heroux, Bartlett, Bochev, Boman, Cyr, Devine, Gaidamour, Gee, Hoemman, Howle, Hu, Kolda, Long, Pawlowski, Peterson, Phipps, Prokopenko, Rajamanickam, Ridzal, Sala, Thornquist, Tuminaro, et al.)
- Includes developers and users around the world
- Epetra-based
  - Limited by 32-bit integer global objects
    - Severe limitation on high fidelity simulations required for challenging problems at SNL
- Most packages employ flat MPI-only; future architectures?
- PETSc is another well-known solvers library (ANL; Smith, Knepley, Brown, et al.)

Trilinos ML Library: algebraic multigrid preconditioners

(R. Tuminaro, J. Hu, C. Siefert, M. Sala, M. Gee, C. Tong, etc.)

- Aggregates to produce a coarser operator
  - Create graph where vertices are block nonzeros in matrix A_k
  - Edge between vertices i and j added if block $B_k(i,j)$ contains nonzeros
  - Uncoupled aggregation
- Restriction/prolongation operator
- $A_{k-1} = R_k A_k P_k$
- Repartition coarser level matrix (ML+Zoltan)
  - Coarser level matrices on a subset of MPI processes to reduce communication
- Coarsest level: serial direct solve (KLU; T. Davis) on 1 MPI process

Another well-known AMG library: LLNL Hypre (R. Falgout, A. Baker, E. Chow, T. Kolev, C. Tong, U. Meier Yang, et al.)

Resistive MHD model

(J. Shadid, R. Pawlowski, E. Cyr, L. Chacon)

Navier-Stokes + Magnetic Induction

$$\rho\frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u}\cdot\nabla\mathbf{u}) - \nabla\cdot(\mathbf{T} + \boxed{\mathbf{T_M}}) - \rho\mathbf{g} = 0$$

$$\mathbf{T} = -(P + \frac{2}{3}\mu(\nabla\cdot\mathbf{u}))\mathbf{I} + \mu[\nabla\mathbf{u} + \nabla\mathbf{u}^T]$$

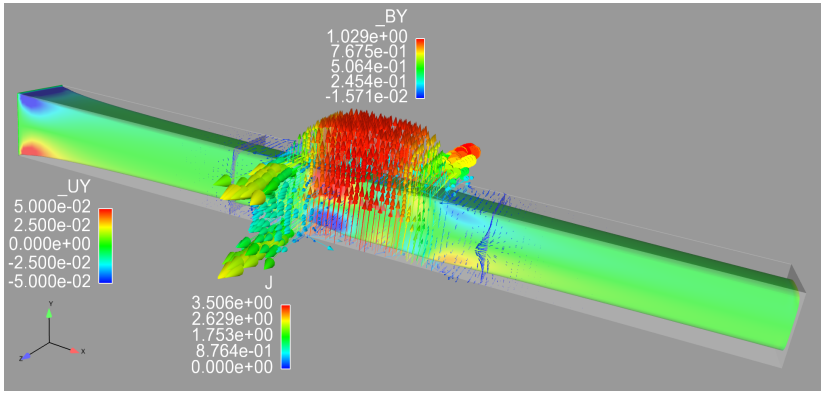
$$\boxed{\mathbf{T_M} = \frac{1}{\mu_0}\mathbf{B}\otimes\mathbf{B} - \frac{1}{2\mu_0}\|\mathbf{B}\|^2\mathbf{I}}$$

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

$$\rho C_p\left[\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T\right] + \nabla\cdot\mathbf{q} - \boxed{\eta\|\mathbf{J}\|^2} = 0$$

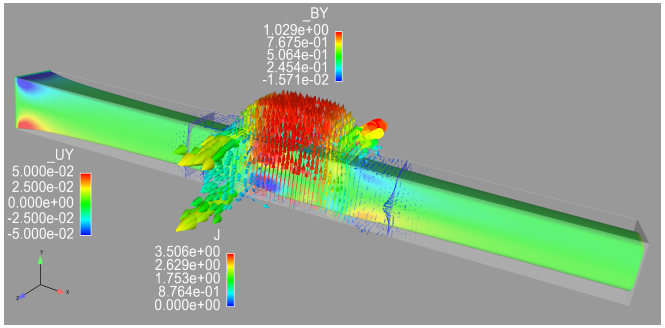
$$\boxed{\frac{\partial \mathbf{B}}{\partial t} - \nabla\times(\mathbf{u}\times\mathbf{B}) + \nabla\times(\frac{\eta}{\mu_0}\nabla\times\mathbf{B}) = \mathbf{0}}$$

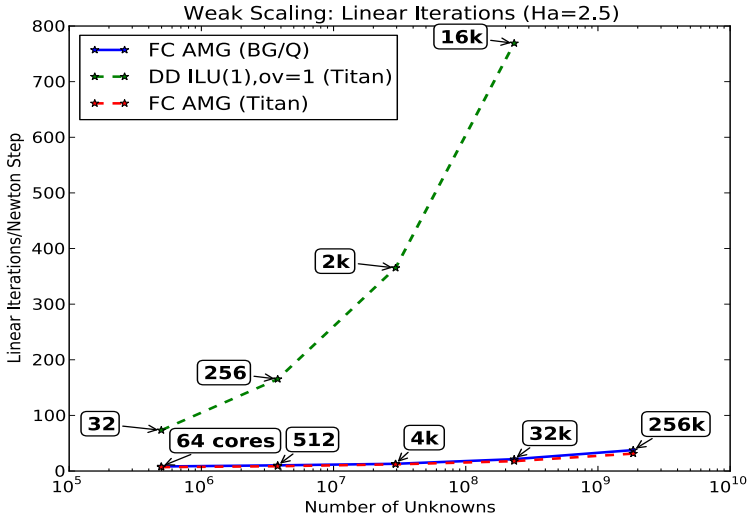
- Steady-state MHD generator (stabilized FE; fully-coupled multigrid preconditioned Newton-Krylov solve)
- 3D flow with external cross-stream B field
- 8 DOFs/mesh node



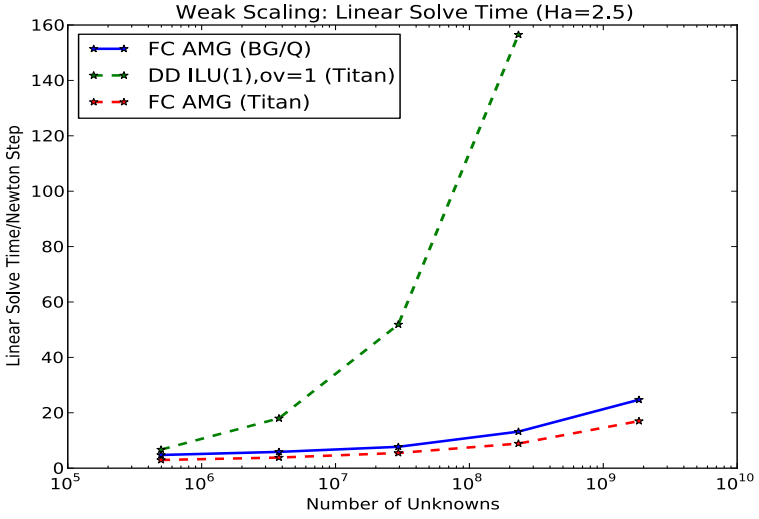
Weak scaling study: MHD generator (classic Trilinos)

- Re = 500, Re_m = 1, Ha = 2.5
- Cray XK7
- IBM Blue Gene/Q (BG/Q)
- Largest problem: 1.8b DOFs (32-bit limit for Epetra)





GMRES Iterations/Newton step



Linear solve time/Newton step

Additive Schwarz domain decomposition does not scale
Multigrid critical for performance and scaling

Brief next-generation "modern" Trilinos overview

- Classic Trilinos (Epetra-based):
  - Limited by 32-bit integer global objects
  - Most packages employ flat MPI-only; future architectures?
- Modern Trilinos solver stack (Tpetra-based):
  - No 32-bit limitation on global objects (employs C++ templated data types)
  - Path forward for future architectures
    - Trilinos Kokkos (Edwards, Trott, Sunderland; not part of this talk)
  - Want Trilinos to impact production engineering applications
  - Significant effort to mature modern Trilinos over past 2.5 years
  - Several SNL applications using modern Trilinos

Modern Trilinos

Functionality	Classic solver stack	Modern solver stack
Distributed linear alg	Epetra	Tpetra (Hoemmen,Trott, etc)
Iterative linear solve	Aztec	Belos (Thornquist,Hoemmen,etc.)
Incomplete factor	Aztec, Ifpack	Ifpack2 (Hoemmen,Hu,Siefert, etc.)
Algebraic multigrid	ML	MueLu (Hu,Prokopenko,Tsuji,Siefert,Tuminaro,etc.)
Partition & load bal	Zoltan	Zoltan2 (Devine,Boman,Rajamanickam,Wolf,etc.)
Direct solve interface	Amesos	Amesos2 (Rajamanickam,etc.)

MueLu library: algebraic multigrid preconditioners

(J. Hu, A. Prokopenko, J. Gaidamour, P. Tsuji, C. Siefert, R. Tuminaro)

- Smoothed aggregation; aggregates to produce a coarser operator
  - Create graph where vertices are block nonzeros in matrix A_k
  - Edge between vertices i and j added if block $B_k(i,j)$ contains nonzeros
  - Uncoupled aggregation
- Restriction/prolongation operator; $A_{k-1} = R_k A_k P_k$
- Repartition coarser level matrix (MueLu+Zoltan2)
- Research/experimental capabilities: EMIN, Petrov-Galerkin, limited geometric multigrid (for specific MHD problems only)

Preliminary Weak Scaling BG/Q: CFD Jet (Re = 10^6, CFL ~0.25)

Drekar/Epetra/Aztec/ML/Ifpack SGS (classic Trilinos)

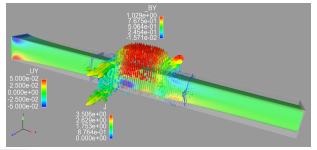
cores	DOF	Newt /dt	Iter/ Newt	Iter/ dt	Total time	Time/Newt (sec)		
						Prec	Solve	Jac
32	901056	3.40	8.41	28.6	1329	1.79	7.76	27.72
256	6931504	3.50	10.8	37.8	1435	1.94	9.56	27.85
2048	54,723,004	3.60	15.72	56.6	1657	2.95	13.6	27.92
16,384	434,886,004	3.60	22.42	80.7	1855	3.02	19.3	27.94
131,072	3,467,532,004	Cannot run problems with DOF > 2.1b						

Drekar/Tpetra/Belos/MueLu/Ifpack2 SGS (modern Trilinos)

cores	DOF	Newt /dt	Iter/ Newt	Iter/ dt	Total time	Time/Newt (sec)		
						Prec	Solve	Jac
32	901,056	3.40	8.44	28.7	1533	2.13	7.29	32.76
256	6,931,504	3.60	11.92	42.9	1729	2.28	10.05	32.86
2048	54,723,004	3.70	16.32	60.4	1932	2.53	13.99	32.89
16,384	434,886,004	3.90	24.67	96.2	2328	3.34	20.67	32.9
131,072	3,467,532,004	4.00	36.23	144.9	3421	19.02	30.55	32.9

- Tpetra slower, iterations higher with Tpetra
- Tpetra/MueLu enables simulations > 2.1b (32-bit)
- Solved 40 billion DOF for 10 time steps

Steady MHD generator weak scaling: linear solve

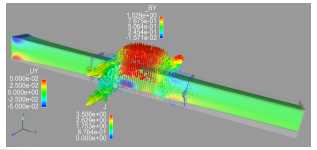


MPI tasks	unknowns	iter/ Newt	Setup(smoother) /Newt(s)	Solve time/ Newt(s)
128	845,000	17.4	9.45 (7.2)	4.11
1024	6,473,096	21.6	10.8 (8.7)	5.31
8192	50,658,056	31	12.2 (9.4)	7.89
65,536	400,799,240	53.3	25.5 (10.4)	14.2
524,288	3,188,616,200	104.8	261 (13.9)	29.8

- BG/Q: 1 MPI task/core

- Preliminary results: development underway to improve scaling
- Algorithmic scaling challenging for nonsymmetric matrices
  - 4096x increase in size: 6.0x increase in iterations, 7.3x increase in time
  - Petrov-Galerkin or energy minimization approaches promising
  - Need better aggregation, better smoothers, etc.
  - Multiple sweeps of ILU significantly improves scaling; demonstrated by ML/ifpack, need to implement in modern Trilinos
  - AMG involves a lot of communication and lots of small messages
  - Ideas for communication reducing multigrid have been proposed (Brown, Adams, Knepley, Falgout, Vassilevski, Nakajima, etc.)
- Memory issues for large number of MPI tasks; further investigation needed (MPI fault? app fault?)

Steady MHD generator weak scaling: multigrid setup



MPI tasks	unknowns	Setup(smoother)/Newt(s)
128	845,000	9.45 (7.2)
1024	6,473,096	10.8 (8.7)
8192	50,658,056	12.2 (9.4)
65,536	400,799,240	25.5 (10.4)
524,288	3,188,616,200	261 (13.9)

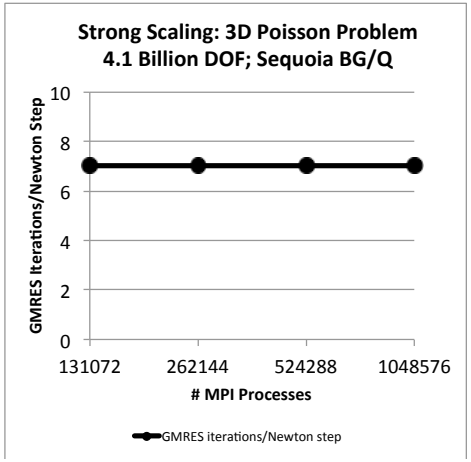
- BG/Q: 1 MPI task/core
- Multigrid setup time/Newton step
- Smoother is ILU(0) with overlap=1

- Setup does not scale (bad things happen 64k -> 512k)
  - Challenge: Tpetra sparse matrix-matrix multiply (A_c=R*A*P) ILU with overlap for large number of tasks; further investigation needed
- Reuse of construction of hierarchy and smoothers
  - Critical for transient simulations (10^4 or 10^5 time steps)
    - Sierra low Mach code: PPE constant, construct once and save
  - In ML/ifpack (use this when run Drekar with Epetra)
  - Some reuse recently added to MueLu
  - No reuse for results above
  - But if mesh connectivity changes, cannot reuse (e.g. adaptive mesh)

How many MPI tasks needed for multigrid?

- Our experience (app dependent): really bad multigrid setup scaling not until > couple 10,000s of MPI tasks (definitely ~100k)
- Do we need to worry about multigrid with one million MPI tasks?
  - Flat MPI-only clearly not the way to go
- Multigrid with O(10^5) MPI tasks relevant for MPI+X approaches ?
  - MPI+X: number of MPI tasks the same as number of compute nodes
  - Number is clearly app dependent
  - May need minimum number per node that is > 1 to utilize NIC bandwidth
  - Future DOE machines (2016-2019) O(10^3) to O(10^5) nodes
  - Exascale machines will be O(10^4) to O(10^6) nodes
- Good multigrid performance on O(10^5) tasks appears to be still relevant
  - Need to continue to work on algorithmic scaling
- Potential approach for "X" in MPI+X -> SNL Trilinos Kokkos

  - Drekar 3D Poisson problem
  - Simple cube geometry
  - Optimal iteration count to 1 million cores
  - Preconditioner setup does not scale

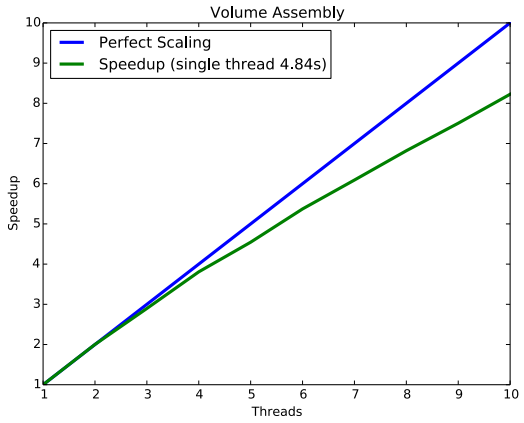


U.S. DEPARTMENT OF ENERGY | National Nuclear Security Administration

Sandia National Laboratories

Future looking slide: Kokkos for FEM matrix assembly

- Matrix assembly for Maxwell's eqns (2nd order form); edge-based
  - Results courtesy of Cyr, Bettencourt, Demeshko, Pawlowski
  - Initial results for Kokkos implementation of the Trilinos Phalanx library: directed acyclic graph (DAG) based assembly abstraction
  - Template-based embedded C++ data types used within DAG to assemble Jacobian

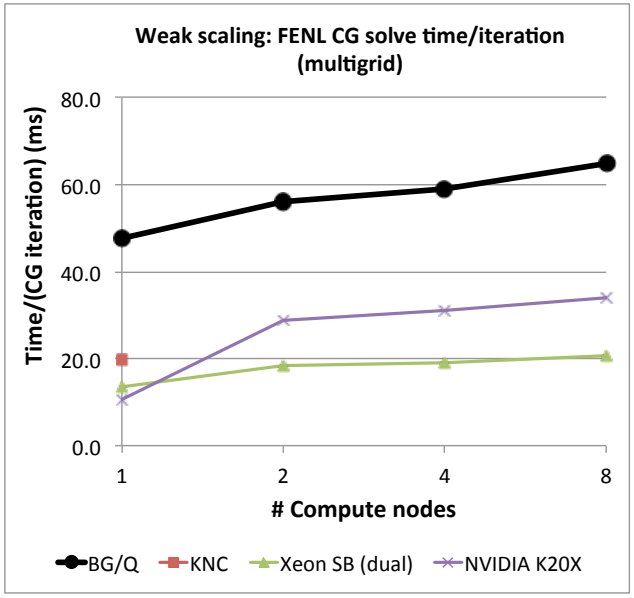


- Strong scaling 40^3 cube of hex elements
- OpenMP threading on dual Intel Xeon (Westmere)

- Drekar employs Phalanx for assembly; will leverage this work

Future looking slide: Kokko for linear solve

- Finite Element Nonlinear (FENL) “miniapp” (C. Edwards): 3D nonlinear heat equation using Kokkos
- MueLu setup and Tpetra mat-mat multiply not yet implemented with Kokkos (setup on CPU)
- MueLu apply using Kokkos (Tpetra converted to Kokkos)
- Results courtesy of E. Phipps



- Preliminary results
- Solve times (setup not included)
- BG/Q:1 MPI rank/node, 64 threads/rank
- KNC: Intel Xeon Phi 224 threads (multi-node results need further investigation)
- Xeon SB (dual): dual-socket Sandy Bridge 1MPI rank/node, 16 threads/rank
- NVIDIA K20X GPU 1 MPI rank/node

FENL with Kokkos runs on CPU, GPU, and Xeon Phi (FENL with MueLu apply)

- Drekar will leverage this work

Concluding Remarks and Future Work

- 40 billion unknown FEM multigrid-preconditioned linear solve (MPI-only) prototyped with Drekar app with modern Trilinos
- Many challenges for multigrid-preconditioned linear solve
  - Scaling; $O(10^5)$ tasks still relevant (app dependent)
  - Multigrid preconditioner setup (sparse mat-mat; ILU overlap)
- Tpetra with Kokkos implementation is ongoing work
- Other Trilinos packages need to be implemented with Kokkos: MueLu setup, additional Ifpack2 smoothers, etc.
- Drekar progress depends on above Kokkos for solver packages
  - Assembly done (I. Demeshko)
- Don't forget app code physics and discretization issues, e.g.
  - Strong convection effects, hyperbolic systems
  - non-uniform FE aspect ratios

Thanks For Your Attention!

Paul Lin (ptlin@sandia.gov)

Acknowledgments:

- Ray Tuminaro, Chris Siefert
- Eric Phipps, Mark Hoemmen
- LLNL LC BG/Q team (especially John Gyllenhaal and Scott Futral)

Funding Acknowledgment:
The authors gratefully acknowledge funding from the DOE NNSA Advanced Simulation & Computing (ASC) program and ASCR Applied Math Program

Extra Slides